

ECONOMICS

PRINCIPLES AND APPLICATIONS



ROBERT E. HALL

MARC LIEBERMAN

UPDATED

SECOND EDITION

WHAT IS ECONOMICS?

E*conomics*. The word conjures up all sorts of images: manic stock traders on Wall Street, an economic summit meeting in a European capital, a somber television news anchor announcing good or bad news about the economy. . . . You probably hear about economics several times each day. What exactly is economics?

First, economics is a *social science*, so it seeks to explain something about *society*. In this sense, it has something in common with psychology, sociology, and political science. But economics is different from these other social sciences, because of *what* economists study and *how* they study it. Economists ask fundamentally different questions, and they answer them using tools that other social scientists find rather exotic.

ECONOMICS, SCARCITY, AND CHOICE

A good definition of economics, which stresses the difference between economics and other social sciences, is the following:

Economics is the study of choice under conditions of scarcity.

This definition may appear strange to you. Where are the familiar words we ordinarily associate with economics: “money,” “stocks and bonds,” “prices,” “budgets,” . . . ? As you will soon see, economics deals with all of these things and more. But first, let’s take a closer look at two important ideas in this definition: scarcity and choice.

SCARCITY AND INDIVIDUAL CHOICE

Think for a moment about your own life—your daily activities, the possessions you enjoy, the surroundings in which you live. Is there anything you don’t have right now that you’d *like* to have? Anything that you already have but that you would like *more* of? If your answer is “no,” congratulations! Either you are well advanced on the path of Zen self-denial, or else you are a close relative of Bill Gates. The rest of us, however, feel the pinch of limits to our material standard of living. This simple truth is at the very core of economics. It can be restated this way: We all face the problem of **scarcity**.

CHAPTER OUTLINE

Economics, Scarcity, and Choice

- Scarcity and Individual Choice
- Scarcity and Social Choice
- Scarcity and Economics

The World of Economics

- Microeconomics and Macroeconomics
- Positive and Normative Economics

Why Study Economics?

- To Understand the World Better
- To Gain Self-Confidence
- To Achieve Social Change
- To Help Prepare for Other Careers
- To Become an Economist

The Methods of Economics

- The Art of Building Economic Models
- Assumptions and Conclusions
- The Four-Step Process

Math, Jargon, and Other Concerns . . .

How to Study Economics

Economics The study of choice under conditions of scarcity.

Scarcity A situation in which the amount of something available is insufficient to satisfy the desire for it.



To make good use of the Internet, you will need the Adobe Acrobat Reader. It can be downloaded from <http://www.adobe.com/prodindex/acrobat/readstep.html>. An economic question is: Why does Adobe give the Reader away free?

At first glance, it may seem that you suffer from an infinite variety of scarcities. There are so many things you might like to have right now—a larger room or apartment, a new car, more clothes . . . the list is endless. But a little reflection suggests that your limited ability to satisfy these desires is based on two other, more basic limitations: scarce *time* and scarce *spending power*.

As individuals, we face a scarcity of time and spending power. Given more of either, we could each have more of the goods and services that we desire.

The scarcity of spending power is no doubt familiar to you. We've all wished for higher incomes so that we could afford to buy more of the things we want. But the scarcity of time is equally important. So many of the activities we enjoy—seeing a movie, taking a vacation, making a phone call—require time as well as money. Just as we have limited spending power, we also have a limited number of hours in each day to satisfy our desires.

Because of the scarcities of time and spending power, each of us is forced to make *choices*. We must allocate our scarce *time* to different activities: work, play, education, sleep, shopping, and more. We must allocate our scarce *spending power* among different goods and services: housing, food, furniture, travel, and many others. And each time we choose to buy something or do something, we are also choosing *not* to buy or do something else.

Economists study the choices we make as individuals and how those choices shape our economy. For example, over the next decade, we may each—as individuals—decide to make more of our purchases over the Internet. Collectively, this decision will determine which firms and industries will expand and hire new workers (such as Internet consulting firms and manufacturers of Internet technology) and which firms will contract and lay off workers (such as traditional “brick and mortar” retailers).

Economists also study the more subtle and indirect effects of individual choice on our society. Will most Americans continue to live in houses, or—like Europeans—will most of us end up in apartments? Will we have an educated and well-informed citizenry? Will traffic congestion in our cities continue to worsen, or is there relief in sight? Will the Internet create faster economic growth and more rapidly rising living standards for years to come or just a short burst of economic activity that will soon subside? These questions hinge, in large part, on the separate decisions of millions of people. To answer them requires an understanding of how individuals make choices under conditions of scarcity.

SCARCITY AND SOCIAL CHOICE

Now let's think about scarcity and choice from *society's* point of view. What are the goals of our society? We want a high standard of living for our citizens, clean air, safe streets, good schools, and more. What is holding us back from accomplishing all of these goals in a way that would satisfy everyone? You already know the answer: scarcity.

In society's case, the problem is a scarcity of **resources**—the things we use to make goods and services that help us achieve our goals. Economists classify resources into three categories:

1. **Labor** is the time human beings spend producing goods and services.
2. **Capital** consists of the long-lasting tools people use to produce goods and services. This includes *physical capital*, such as buildings, machinery, and equipment, as well as *human capital*—the *skills and training* that workers possess.

Resources The land, labor, and capital that are used to produce goods and services.

Labor The time human beings spend producing goods and services.

Capital Long-lasting tools used in producing goods and services.

Human capital The skills and training of the labor force.

3. **Land** is the physical space on which production takes place, as well as the natural resources found under it or on it, such as oil, iron, coal, and lumber.

Land The physical space on which production occurs, and the natural resources that come with it.

Anything *produced* in the economy comes, ultimately, from some combination of these resources. Think about the last lecture you attended at your college. You were consuming a service—a college lecture. What went into producing that service? Your instructor was supplying labor. Many types of capital were used as well. The physical capital included desks, chairs, a chalkboard or transparency projector, and the classroom building itself. It also included the computer your instructor may have used to compose lecture notes. In addition, there was human capital—your instructor’s specialized knowledge and lecturing skills. Finally, there was land—the property on which your classroom building sits.

Besides the three resources, other things were used to produce your college lecture. Chalk, for example, is a tool used by your instructor, so you might think it should be considered capital, but it is not. Why not? Because it is not *long lasting*. Typically, economists consider a tool to be capital only if it lasts for a few years or longer. Chalk is used up as the lecture is produced, so it is considered a *raw material* rather than capital.

But a little reflection should convince you that a piece of chalk is itself produced from some combination of the three resources (labor, capital, and land). In fact, all of the raw materials needed to produce the lecture—the energy used to heat or cool your building, the computer paper used for your instructor’s lecture notes, and so on—come, ultimately, from society’s three resources. And the scarcity of these resources, in turn, causes the scarcity of all goods and services produced from them.

As a society, our resources—land, labor, and capital—are insufficient to produce all the goods and services we might desire. In other words, society faces a scarcity of resources.

This stark fact about the world helps us understand the choices a society must make. Do we want a more educated citizenry? Of course. But that will require more labor—construction workers to build more classrooms and teachers to teach in them. It will require more natural resources—land for classrooms and lumber to build them. And it will require more capital—cement mixers, trucks, and more. These very same resources, however, could instead be used to produce *other* things that we find desirable—things such as new homes, hospitals, automobiles, or feature films. As a result, every society must have some method of *allocating* its scarce resources—choosing which of our many competing desires will be fulfilled and which will not be.

Many of the big questions of our time center on the different ways in which resources can be allocated. The cataclysmic changes that rocked Eastern Europe and the former Soviet Union during the early 1990s arose from a very simple fact: The method these countries used for decades to allocate resources was not working. Closer to home, the never-ending debates between Democrats and Republicans in the United States reflect subtle but important differences of opinion about how to allocate resources. Often, these are disputes about whether the private sector can handle the allocation of resources on its own or whether the government should be involved.

SCARCITY AND ECONOMICS

The scarcity of resources—and the choices it forces us to make—is the source of all of the problems you will study in economics. Households have limited incomes for satisfying their desires, so they must choose carefully how they allocate their spending

among different goods and services. Business firms want to make the highest possible profit, but they must pay for their resources, so they carefully choose *what* to produce, *how much* to produce, and *how* to produce it. Federal, state, and local government agencies work with limited budgets, so they must carefully choose which goals to pursue. Economists study these decisions made by households, firms, and governments to explain how our economic system operates, to forecast the future of our economy, and to suggest ways to make that future even better.

THE WORLD OF ECONOMICS

The field of economics is surprisingly broad. It extends from the mundane—why does a pound of steak cost more than a pound of chicken?—to the personal and profound—how do couples decide how many children to have? With a field this broad, it is useful to have some way of classifying the different types of problems economists study and the different methods they use to analyze them.

MICROECONOMICS AND MACROECONOMICS

The field of economics is divided into two major parts: microeconomics and macroeconomics. **Microeconomics** comes from the Greek word *mikros*, meaning “small.” It takes a close-up view of the economy, as if looking through a microscope. Microeconomics is concerned with the behavior of *individual* actors on the economic scene—households, business firms, and governments. It looks at the choices they make, and how they interact with each other when they come together to trade *specific* goods and services. What will happen to the cost of movie tickets over the next five years? How many jobs will open up in the fast-food industry? How would U.S. phone companies be affected by a tax on imported cell phones? These are all microeconomic questions because they analyze individual *parts* of an economy, rather than the *whole*.

Macroeconomics—from the Greek word *makros*, meaning “large”—takes an *overall* view of the economy. Instead of focusing on the production of carrots or computers, macroeconomics lumps all goods and services together and looks at the economy’s *total output*. Instead of focusing on employment in the fast-food industry or the manufacturing sector, it considers *total employment* in the economy. Instead of asking why credit card loans carry higher interest rates than home mortgage loans, it asks what makes interest rates *in general* rise or fall. In all of these cases, macroeconomics focuses on the big picture and ignores the fine details.

POSITIVE AND NORMATIVE ECONOMICS

The micro versus macro distinction is based on the level of detail we want to consider. Another useful distinction has to do with the *purpose* in analyzing a problem. **Positive economics** deals with what *is*—with *how* the economy works, plain and simple. If we lower income tax rates in the United States next year, will the economy grow faster? If so, by how much? And what effect will this have on total employment? These are all positive economic questions. We may disagree about the answers, but we can all agree that the correct answers to these questions do *exist*—we just have to find them.

Normative economics concerns itself with what *should be*. It is used to make judgments about the economy, identify problems, and prescribe solutions. While positive economics is concerned with just the facts, normative economics requires

Microeconomics The study of the behavior of individual households, firms, and governments; the choices they make; and their interaction in specific markets.

Macroeconomics The study of the economy as a whole.

Positive economics The study of what *is*, of how the economy works.

Normative economics The study of what *should be*; it is used to make value judgments, identify problems, and prescribe solutions.

us to make value judgments. When an economist advises that we cut government spending—an action that will benefit some citizens and harm others—the economist is engaging in normative analysis.

Positive and normative economics are intimately related in practice. For one thing, we cannot properly argue about what we should or should not do unless we know certain facts about the world. Every normative analysis is therefore based on an underlying positive analysis. But while a positive analysis can, at least in principle, be conducted without value judgments, a normative analysis is always based, at least in part, on the values of the person conducting it.

Why Economists Disagree. The distinction between positive and normative economics can help us understand why economists sometimes disagree. Suppose you are watching a television interview in which two economists are asked whether the United States should eliminate all government-imposed barriers to trading with the rest of the world. The first economist says, “Yes, absolutely,” but the other says, “No, definitely not.” Why the sharp disagreement?

The difference of opinion may be *positive* in nature: The two economists may have different views about what would actually happen if trade barriers were eliminated. Differences like this sometimes arise because our knowledge of the economy is imperfect, or because certain facts are in dispute.

More likely, however, the disagreement will be *normative*. Economists, like everyone else, have different values. In this case, both economists might agree that opening up international trade would benefit *most* Americans, but harm *some* of them. Yet they may still disagree about the policy move because they have different values. The first economist might put more emphasis on benefits to the overall economy, while the second might put more emphasis on preventing harm to a particular group. Here, the two economists have come to the same *positive* conclusion, but their *different values* lead them to different *normative* conclusions.

In the media, economists are rarely given enough time to express the basis for their opinions, so the public hears only the disagreement. People may then conclude—wrongly—that economists cannot agree about how the economy works when the *real* disagreement is over which goals are most important for our society.

WHY STUDY ECONOMICS?

Students take economics courses for all kinds of reasons.

TO UNDERSTAND THE WORLD BETTER

Applying the tools of economics can help you understand global and cataclysmic events such as wars, famines, epidemics, and depressions. But it can also help you understand much of what happens to you locally and personally—the worsening traffic conditions in your city, the raise you can expect at your job this year, or the long line of people waiting to buy tickets for a popular concert. Economics has the power to help us understand these phenomena because they result, in large part, from the choices we make under conditions of scarcity.

Economics has its limitations, of course. But it is hard to find any aspect of life about which economics does not have *something* important to say. Economics cannot explain why so many Americans like to watch television, but it *can* explain how TV networks decide which programs to offer. Economics cannot protect you from a



The Federal Reserve Bank of Minneapolis asked some Nobel Prize winners how they became interested in economics. Their stories can be found at <http://woodrow.mpls.frb.fed.us/pubs/rjion/98-12/quotes.html>.

robbery, but it *can* explain why some people choose to become thieves and why no society has chosen to eradicate crime completely. Economics will not improve your love life, resolve unconscious conflicts from your childhood, or help you overcome a fear of flying, but it *can* tell us how many skilled therapists, ministers, and counselors are available to help us solve these problems.

TO GAIN SELF-CONFIDENCE

Those who have never studied economics often feel that mysterious, inexplicable forces are shaping their lives, buffeting them like the bumpers in a pinball machine, determining whether or not they'll be able to find a job, what their salary will be, whether they'll be able to afford a home, and in what kind of neighborhood. If you've been one of those people, all that is about to change. After you learn economics, you may be surprised to find that you no longer toss out the business page of your local newspaper because it appears to be written in a foreign language. You may no longer lunge for the remote and change the channel the instant you hear "And now for news about the economy. . . ." You may find yourself listening to economic reports with a critical ear, catching mistakes in logic, misleading statements, or out-and-out lies. When you master economics, you gain a sense of mastery over the world, and thus over your own life as well.

TO ACHIEVE SOCIAL CHANGE

If you are interested in making the world a better place, economics is indispensable. There is no shortage of serious social problems worthy of our attention—unemployment, hunger, poverty, disease, child abuse, drug addiction, violent crime. Economics can help us understand the origins of these problems, explain why previous efforts to solve them have failed, and enable us to design new, more effective solutions.

TO HELP PREPARE FOR OTHER CAREERS

Economics has long been the most popular college major for individuals intending to work in business. But in the last two decades it has also become popular among those planning careers in politics, international relations, law, medicine, engineering, psychology, and other professions. This is for good reason: Practitioners in each of these fields often find themselves confronting economic issues. For example, lawyers increasingly face judicial rulings based on the principles of economic efficiency. Doctors will need to understand how new laser technologies or changes in the structure of HMOs will affect their practices. Industrial psychologists need to understand the economic implications of workplace changes they may advocate, such as flexible scheduling or on-site child care.

TO BECOME AN ECONOMIST

Only a tiny minority of this book's readers will decide to become economists. This is welcome news to the authors, and after you have studied labor markets in your *microeconomics* course, you will understand why. But if you do decide to become an economist—obtaining a master's degree or even a Ph.D.—you will find many possibilities for employment. Of 16,780 members of the American Economic Association who responded to a recent survey,¹ 65 percent were employed at colleges or universities. The rest were engaged in a variety of activities in both the private sector (21 percent) and government (14 percent). Economists are hired by banks to as-

¹ *American Economic Review*, December 1993, p. 635.

sess the risk of investing abroad; by manufacturing companies, to help them determine new methods of producing, marketing, and pricing their products; by government agencies, to help design policies to fight crime, disease, poverty, and pollution; by international organizations, to help create aid programs for less developed countries; by the media to help the public interpret global, national, and local events; and even by nonprofit organizations, to provide advice on controlling costs and raising funds more effectively.

THE METHODS OF ECONOMICS

One of the first things you will notice as you begin to study economics is the heavy reliance on *models*. Indeed, the discipline goes beyond any other social science in its insistence that every theory be represented by an explicit, carefully constructed *model*.

You've no doubt encountered many models in your life. As a child, you played with model trains, model planes, or model people—dolls. In a high school science course, you probably saw a model of an atom—one of those plastic and wire contraptions with red, blue, and green balls representing protons, neutrons, and electrons. You may have also seen architects' cardboard models of buildings. These are physical models, three-dimensional replicas that you can pick up and hold. Economic models, on the other hand, are built not with cardboard, plastic, or metal but with words, diagrams, and mathematical statements.

What, exactly, is a model?

A model is an abstract representation of reality.

Model An abstract representation of reality.

The two key words in this definition are *abstract* and *representation*. A model is not supposed to be exactly like reality. Rather, it *represents* the real world by *abstracting*, or *taking from* the real world that which will help us understand it. In any model, many real-world details are left out.

THE ART OF BUILDING ECONOMIC MODELS

When you build a model, how do you know which details to include and which to leave out? There is no simple answer to this question. The right amount of detail depends on your purpose in building the model in the first place. There is, however, one guiding principle:

A model should be as simple as possible to accomplish its purpose.

This means that a model should contain only the *necessary* details.

To understand this a little better, think about a map. A map is a model—it represents a part of the earth's surface. But it leaves out many details of the real world. First, maps are two-dimensional, so they leave out the third dimension—height—of the real world. Second, maps always ignore small details, such as trees and houses and potholes. Third, a map is much smaller than the area it represents. But when you buy a map, how much detail do you want it to have?

Let's say you are in Boston, and you need a map (your *purpose*) to find the best way to drive from Logan Airport to the downtown convention center. In this case, you would want a very detailed city map, with every street, park, and plaza in Boston clearly illustrated and labeled. A highway map, which ignores these details, wouldn't do at all.



These maps are *models*. But each would be used for a different purpose.

Simplifying assumption Any assumption that makes a model simpler without affecting any of its important conclusions.

Critical assumption Any assumption that affects the conclusions of a model in an important way.

But now suppose your purpose is different: to select the best driving route from Boston to Cincinnati. Now you want a highway map. A map that shows every street between Boston and Cincinnati would have *too much* detail. All of that extraneous information would only obscure what you really need to see.

Although economic models are more abstract than road maps, the same principle applies in building them: The level of detail that would be just right for one purpose will usually be too much or too little for another. When you feel yourself objecting to a model in this text because something has been left out, keep in mind the purpose for which the model is built. In introductory economics, the purpose is entirely educational. The models are designed to help you understand some simple, but powerful, principles about how the economy operates. Keeping the models simple makes it easier to see these principles at work and remember them later.

Of course, economic models have other purposes besides education. They can help businesses make decisions about pricing and production, help households decide how and where to invest their savings, and help governments and international agencies formulate policies. Models built for these purposes will be much more detailed than the ones in this text, and you will learn about them if you take more advanced courses in economics. But even complex models are built around a very simple framework—the same framework you will be learning here.

ASSUMPTIONS AND CONCLUSIONS

Every economic model begins with *assumptions* about the world. There are two types of assumptions in a model: simplifying assumptions and critical assumptions.

A **simplifying assumption** is just what it sounds like—a way of making a model simpler without affecting any of its important conclusions. The purpose of a simplifying assumption is to rid a model of extraneous detail so its essential features can stand out more clearly. A road map, for example, makes the simplifying assumption, “There are no trees,” because trees on a map would only get in the way. Similarly, in an economic model, we might assume that there are only two goods that households can choose from or that there are only two nations in the world. We make such assumptions *not* because they are true, but because they make a model easier to follow and do not change any of the important insights we can get from it.

A **critical assumption**, by contrast, is an assumption that affects the conclusions of a model in important ways. When you use a road map, you make the critical assumption, “All of these roads are open.” If that assumption is wrong, your conclusion—the best route to take—might be wrong as well.

In an economic model, there are always one or more critical assumptions. You don’t have to look very hard to find them, because economists like to make these assumptions explicit right from the outset. For example, when we study the behavior of business firms, our model will assume that firms try to earn the highest possible profit for their owners. By stating this assumption up front, we can see immediately where the model’s conclusions spring from.

THE FOUR-STEP PROCESS

As you read this textbook, you will learn how economists use economic models to address a wide range of problems. In Chapter 2, for example, you will see how a simple economic model can give us important insights about society’s production choices. And subsequent chapters will present still different models that help us understand the U.S. economy and the global economic environment in which it operates. As you read, it may seem to you that there are a lot of models to learn and remember . . . and, indeed, there are.

But there is an important insight about economics that—once mastered—will make your job easier than you might think. The insight is this: There is a remarkable similarity in the types of models that economists build, the assumptions that underlie those models, and what economists actually *do* with them. In fact, you will see that economists follow the same *four-step procedure* to analyze almost any economic problem. The first two Key Steps explain how economists *build* an economic model, and the second two Key Steps explain how they *use* the model.

What are these four steps that underlie the economic approach to almost any problem? Sorry for the suspense, but you'll have to wait a bit—until the end of Chapter 3—for the answer. By that time, you'll have learned a little more about economics, and the four-step procedure will make more sense to you.

MATH, JARGON, AND OTHER CONCERNS . . .

Economists often express their ideas using mathematical concepts and a special vocabulary. Why? Because these tools enable economists to express themselves more precisely than with ordinary language. For example, someone who has never studied economics might say, “When used textbooks are available, students won't buy new textbooks.” That statement might not bother you right now. But once you've finished your first economics course, you'll be saying it something like this: “When the price of used textbooks falls, the demand curve for new textbooks shifts leftward.”

Does the second statement sound strange to you? It should. First, it uses a special term—a *demand curve*—that you haven't yet learned. Second, it uses a mathematical concept—a *shifting curve*—with which you might not be familiar. But while the first statement might mean a number of different things, the second statement—as you will see in Chapter 3—can mean only *one* thing. By being precise, we can steer clear of unnecessary confusion. If you are worried about the special vocabulary of economics, you can relax. All of the new terms will be defined and carefully explained as you encounter them. Indeed, this textbook does not assume you have any special knowledge of economics. It is truly meant for a “first course” in the field.

But what about the math? Here, too, you can relax. While professional economists often use sophisticated mathematics to solve problems, only a little math is needed to understand basic economic *principles*. And virtually all of this math comes from high school algebra and geometry.

Still, you may have forgotten some of your high school math. If so, a little brushing up might be in order. This is why we have included an appendix at the end of this chapter. It covers some of the most basic concepts—such as the equation for a straight line, the concept of a slope, and the calculation of percentage changes—that you will need in this course. You may want to glance at this appendix now, just so you'll know what's there. Then, from time to time, you'll be reminded about it when you're most likely to need it.

HOW TO STUDY ECONOMICS

As you read this book or listen to your instructor, you may find yourself nodding along and thinking that everything makes perfect sense. Economics may even seem easy. Indeed, it *is* rather easy to follow economics, since it's based so heavily on simple logic. But *following* and *learning* are two different things. You will eventually discover (preferably *before* your first exam) that economics must be studied actively, not passively.



An on-line introduction to the use of graphs can be found at <http://syllabus.syr.edu/cid/graph/book.html>.

If you are reading these words lying back on a comfortable couch, a phone in one hand and a remote control in the other, you are going about it in the wrong way. Active studying means reading with a pencil in your hand and a blank sheet of paper in front of you. It means closing the book periodically and *reproducing* what you have learned. It means listing the steps in each logical argument, retracing the cause-and-effect steps in each model, and drawing the graphs that represent the model. It means *thinking* about the basic principles of economics and how they relate to what you are learning. It is hard work, but the payoff is a good understanding of economics and a better understanding of your own life and the world around you.

S U M M A R Y

Economics is the study of choice under conditions of scarcity. As individuals, and as a society, we have unlimited desires for goods and services. Unfortunately, the *resources*—land, labor, and capital—needed to produce those goods and services are scarce. Therefore, we must choose which desires to satisfy and how to satisfy them. Economics provides the tools that explain those choices.

The field of economics is divided into two major areas. *Microeconomics* studies the behavior of individual households, firms, and governments as they interact in specific markets. *Macroeconomics*, by contrast, concerns itself with the

behavior of the entire economy. It considers variables such as total output, total employment, and the overall price level.

Economics makes heavy use of *models*—abstract representations of reality. These models are built with words, diagrams, and mathematical statements that help us understand how the economy operates. All models are simplifications, but a good model will have *just enough detail for the purpose at hand*.

When analyzing almost any problem, economists follow a four-step procedure in building and using economic models. This four-step procedure will be introduced at the end of Chapter 3.

K E Y T E R M S

economics
scarcity
resources
labor

capital
human capital
land
microeconomics

macroeconomics
positive economics
normative economics
model

simplifying assumption
critical assumption

R E V I E W Q U E S T I O N S

- Discuss (separately) how scarcity arises for households, businesses, and governments.
- Would each of the following be classified as microeconomics or macroeconomics? Why?
 - Research into why the growth rate of total production increased during the 1990s.
 - A theory of how consumers decide what to buy.
 - An analysis of Dell Computer's share of the personal computer market.
 - Research on why interest rates were unusually high in the late 1970s and early 1980s.
- Discuss whether each statement is an example of positive economics or normative economics or if it contains elements of both:
 - An increase in the personal income tax will slow the growth rate of the economy.
 - The goal of any country's economic policy should be to increase the well-being of its poorest, most vulnerable citizens.
 - Excess regulation of small business is stifling the economy. Small business has been responsible for most of the growth in employment over the last 10 years, but regulations are putting a severe damper on the ability of small businesses to survive and prosper.
 - The 1990s were a disastrous decade for the U.S. economy. Income inequality increased to its highest level since before World War II.
- What determines the level of detail that an economist builds into a model?
- What is the difference between a simplifying assumption and a critical assumption?

P R O B L E M

1. Come up with a list of critical assumptions that could lie behind each of the following statements. Discuss whether each assumption would be classified as normative or positive.
 - a. The United States is a democratic society.
 - b. European movies are better than American movies.
 - c. The bigger the city, the higher the quality of the newspaper.

E X P E R I E N T I A L E X E R C I S E

1. Go to the Bank of Sweden's Web page on the Nobel Prize in economic science at <http://www.ee.nobel.se/prize/memorial.html>. Review the descriptions of some recent awards and try to determine whether each of those awards was primarily for work in microeconomics or macroeconomics.



APPENDIX

GRAPHS AND OTHER USEFUL TOOLS

TABLES AND GRAPHS

A brief glance at this text will tell you that graphs are important in economics. Graphs provide a convenient way to display data. Take the example of Len & Harry's, an up-and-coming manufacturer of high-end ice cream products, located in Texas. Suppose that you've just been hired to head Len & Harry's advertising department, and you want to learn as much as you can about how advertising can help the company's sales.

Table A.1 records the company's total advertising outlay per month in the left-hand column, and the company's ice cream sales during that same month are shown in the right-hand column. Notice that the data are organized so that advertising outlay increases as we move down the first column. Often, just looking at such a table can reveal useful patterns. In this case, it seems that higher advertising outlays are associated with higher monthly sales. This suggests that there may be some *causal relationship* between advertising and sales.

To explore this relationship further, we might decide to plot the data and draw a graph (see Figure A.1). First, we need to choose units for our two variables. We'll measure both advertising and sales in thousands of dollars. Different values of one variable are then measured along the horizontal axis, increasing as we move rightward from the origin. The corresponding values of the other variable are measured along the ver-

tical axis, increasing as we move upward, away from the origin.

Using the data in the table, let X stand for advertising outlay per month, and let Y stand for sales per month. Notice that each row of the table gives us a pair of numbers: The first is always the value of the variable we are calling X , and the second is the value of the variable we are calling Y . We often write such pairs in the form (X, Y) . For example, we would write the first three rows of the table as $(2, 46)$, $(3, 49)$, and $(6, 58)$, respectively.

To plot the pair (X, Y) on a graph, begin at the origin, where the axes meet. Count rightward X units along the horizontal axis, then count upward Y units parallel to the vertical axis, and then mark the spot. For example, to plot the pair $(2, 46)$, we go rightward 2 units along the horizontal axis and then upward 46 units along the vertical axis, arriving at the point marked A in Figure A.1. To plot the next pair, $(3, 49)$, we go rightward from the origin 3 units and then upward 49 units, arriving at the point marked B . Carrying on in just this way, we can plot all remaining pairs in Table A.1 as the points C, D, E , and F .

If we connect points A through F , we see that they all lie along the same straight line. Now we are getting somewhere. The relationship we've discovered appears from the graph to be very regular, indeed.

Study the graph closely. You will notice that each time advertising increases (moves rightward) by \$1,000,

TABLE A.1

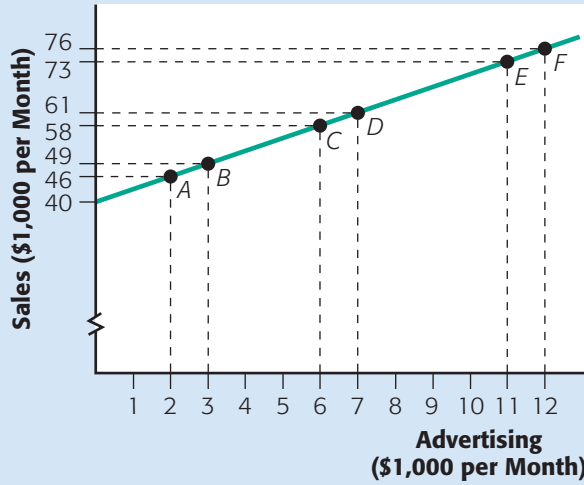
ADVERTISING AND SALES
AT LEN & HARRY'S

Advertising
(\$1,000s per Month)

Sales
(\$1,000s per Month)

2	46
3	49
6	58
7	61
11	73
12	76

FIGURE A.1



Y moves upward by \$3,000. For example, when advertising rises from \$2,000 to \$3,000, sales rise from \$46,000 to \$49,000. By checking between any other two points on the graph, you will see that every time X increases horizontally by one unit (here, a unit is \$1,000), Y increases vertically by three units (here, by \$3,000). Thus, we conclude that the *rate of change* in Y is three units of Y for every one-unit increase in X.

The *slope* of a graph tells us the rate at which the Y-variable changes for every one-unit change in the X-variable. The slope of a straight line between any two points (X_1, Y_1) and (X_2, Y_2) is defined as the change in Y—the vertical “rise”—divided by the change in X—the horizontal “run.” This is why the slope is often described as “rise over run.” Supposing we start at (X_1, Y_1) and end at (X_2, Y_2) ; then the change in the X-variable is $(X_2 - X_1)$. The corresponding change in the Y-variable is $(Y_2 - Y_1)$. We therefore compute the slope as follows:

$$\begin{aligned} \text{Slope of the line} \\ \text{from } (X_1, Y_1) \text{ to } (X_2, Y_2) &= \frac{\text{Rise along vertical axis}}{\text{Run along horizontal axis}} \\ &= \frac{Y_2 - Y_1}{X_2 - X_1} \end{aligned}$$

We sometimes use the capital Greek letter, Δ (“delta”), to denote a change in a variable. Here we would write $\Delta X = X_2 - X_1$ to denote the change in X, and $\Delta Y = Y_2 - Y_1$ to denote the corresponding change

in Y. We then could write that same formula for the slope more compactly as

$$\text{Slope of the line from } (X_1, Y_1) \text{ to } (X_2, Y_2) = \frac{\Delta Y}{\Delta X}.$$

NONLINEAR GRAPHS

Although many of the relationships we encounter in economics have straight-line graphs, many do not. Still, graphs can help us understand the underlying relationships, and the concept of slope remains very useful.

As an example, look at the data in Table A.2, which records the price of a share of Len and Harry’s stock at different points in time since the stock first appeared on the market. To understand how the price of this stock has behaved over time, we might again start by plotting a graph of the data in the table. It seems natural to measure time—in “weeks since launch”—on the X-axis and stock price—in “dollars per share”—on the Y-axis. As you can see in Figure A.2, Len and Harry’s has had a rocky ride since it came on the market. In its first 10 weeks, the stock’s price rose, so the slope of the underlying relationship was positive during that time. Over the next 10 weeks, the story changed: The stock’s price decreased, so the slope of the relationship was negative then. Between weeks 20 and 30, things leveled off: There was no change in the stock’s price, so the slope of the

TABLE A.2

PRICE OF LEN & HARRY'S STOCK SINCE LAUNCH

Weeks Since Launch	Stock Price
3	\$20
10	50
18	35
20	20
25	20
30	20
40	75

graph was zero during that time. However, between weeks 30 and 40 things picked up, and once again the slope turned positive, since the price of the stock increased.

From this example, we can see the following:

- The slope is positive whenever an increase in X is associated with an increase in Y .
- The slope is negative whenever an increase in X is associated with a decrease in Y .
- The slope is equal to zero whenever an increase in X is associated with no change in Y .

LINEAR EQUATIONS

Let's go back to the relationship between advertising and sales, as shown in Table A.1. What if you need to know how much sales the firm could expect if it spent \$5,000 on advertising next month? What if it spent \$8,000, or \$9,000? Wouldn't it be nice to be able to an-

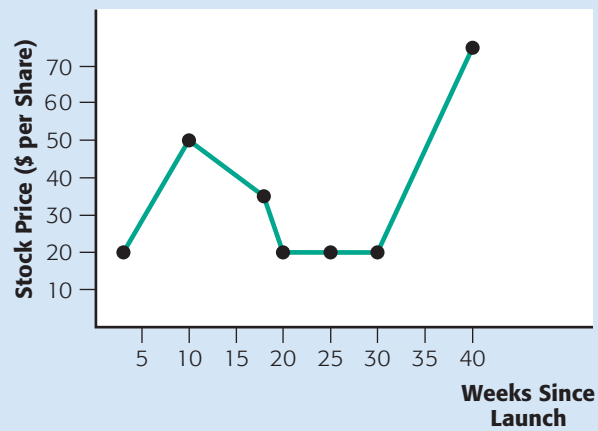
swer questions like this without having to pull out tables and graphs to do it? As it turns out, anytime the relationship you are studying has a straight-line graph, it is easy to figure out the equation for the entire relationship. You then can use the equation to answer any such question that might be put to you.

All straight lines have the same general form. If Y stands for the variable on the vertical axis and X for the variable on the horizontal axis, every straight line has an equation of the form

$$Y = a + bX,$$

where a stands for some number and b for another number. The number a is called the vertical *intercept*, because it marks the point where the graph of this equation hits (intercepts) the vertical axis; this occurs when X takes the value zero. (If you plug $X = 0$ into the equation, you will see that, indeed, $Y = a$.) The number b is the slope of the line, telling us how much Y will

FIGURE A.2



change every time X changes by one unit. To confirm this, note that as X increases from 0 to 1, Y goes from a to $a + b$. The number b is therefore the change in Y corresponding to a one-unit change in X —exactly what the slope of the graph should tell us.

More generally, if X changes from some value X_1 to some other value X_2 , Y will change from

$$Y_1 = a + bX_1$$

to

$$Y_2 = a + bX_2.$$

If we subtract Y_1 from Y_2 to compute how much Y has changed (ΔY), we find that

$$\begin{aligned} \Delta Y = Y_2 - Y_1 &= (a + bX_2) - (a + bX_1) \\ &= a + bX_2 - a - bX_1 \\ &= b(X_2 - X_1) \\ &= b\Delta X. \end{aligned}$$

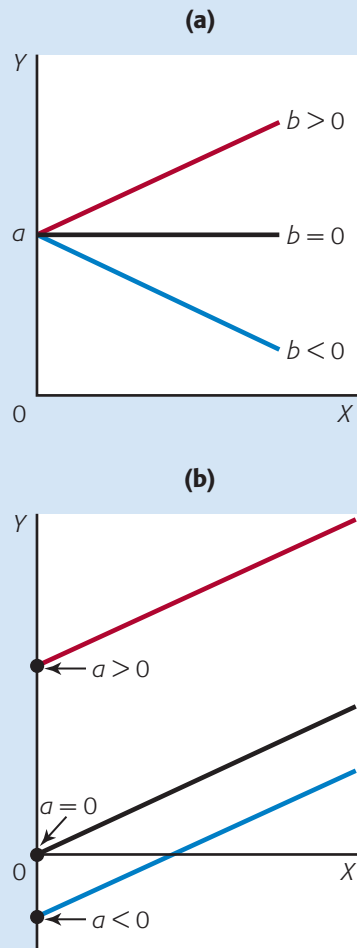
Dividing both sides of the equation $\Delta Y = b\Delta X$ by ΔX , we get

$$\frac{\Delta Y}{\Delta X} = b,$$

confirming that b really does measure the slope.

If b is a positive number, a one-unit increase in X causes Y to increase by b units, so the graph of our line would slope upward, as illustrated by the red line in panel (a) of Figure A.3. If b is a negative number, then a one-unit increase in X will cause Y to decrease by b units, so

FIGURE A.3



the graph would slope downward, as the blue line does in panel (a). Of course, b could equal zero. If it does, a one-unit increase in X causes no change in Y , so the graph of the line is flat, like the black line in panel (a).

The value of a has no effect on the slope of the graph. Instead, different values of a determine the graph's position. When a is a positive number, the graph will intercept the vertical Y -axis above the origin, as the red line does in panel (b) of Figure A.3. When a is negative, however, the graph will intercept the Y -axis *below* the origin, like the blue line in panel (b). When a is zero, the graph intercepts the Y -axis right at the origin, as the black line does in panel (b).

Let's see if we can figure out the equation for the relationship depicted in Figure A.1. There, X denotes advertising and Y denotes sales. On the graph, it is easy to see that when advertising expenditure is zero, sales are \$40,000. Therefore, our equation will have a *vertical* intercept of $a = 40$. Earlier, we calculated the slope of this graph to be 3. Therefore, the equation will have $b = 3$. Putting these two observations together, we find that the equation for the line in Figure A.1 is

$$Y = 40 + 3X.$$

Now if you need to know how much in sales to expect from a particular expenditure on advertising, you'd be able to come up with an answer: You'd simply multiply the amount spent on advertising by 3, add \$40,000, and that would be your sales. To confirm this, plug in for X in this equation any amount of advertising from the left-hand column of Table A.1. You'll see that you get the corresponding amount of sales in the right-hand column.

HOW LINES AND CURVES SHIFT

So far, we've focused on relationships where some variable Y depends on a single other variable, X . But in many of our theories, we recognize that some variable of interest to us is actually affected by more than just one other variable. When Y is affected by both X and some third variable, changes in that third variable will usually cause a *shift* in the graph of the relationship between X and Y . This is because whenever we draw the graph between X and Y , we are holding fixed every other variable that might possibly affect Y .

A graph between two variables X and Y is only a picture of their relationship when all other variables affecting Y are constant. Changes in any one or more of those other variables will shift the graph of X and Y .

Think back to the relationship between advertising and sales. Earlier, we supposed sales depend only on advertising. But suppose we make an important discovery: Ice cream sales are *also* affected by how hot the weather is. What's more, all of the data in Table A.1 on which we previously based our analysis turns out to have been from the month of June, when the average temperature in Texas is 80 degrees. What's going to happen in July, when the average temperature rises to 100 degrees?

In Figure A.4 we've redrawn the graph from Figure A.1, this time labeling the line "June." Often, a

FIGURE A.4

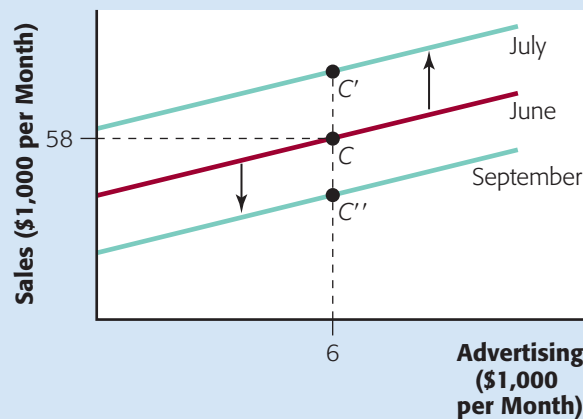
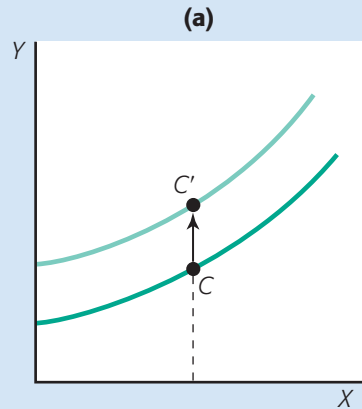
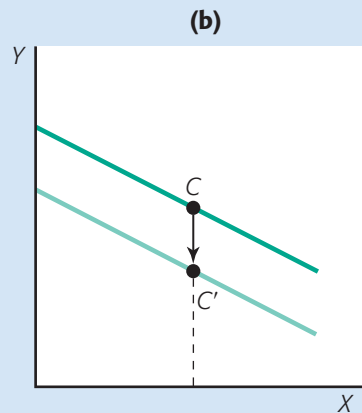


FIGURE A.5



An increase in Z causes an increase in Y at any value of X



An increase in Z causes a decrease in Y at any value of X

good way to determine how a graph will shift is to perform a simple experiment like this: Put your pencil tip anywhere on the graph labeled June—let's say at point C . Now ask the following question: If I hold advertising constant at \$6,000, do I expect to sell more or less ice cream as temperature rises in July? If you expect to sell more, then the amount of sales corresponding to \$6,000 of advertising will be *above* point C , at a point such as C' . From this, we can tell that the graph will shift upward as temperature rises. In September, however, when temperatures fall, the amount of sales corresponding to \$6,000 in advertising would be less than it is at point C . It would be

shown by a point such as C'' . In that case, the graph would shift downward.

The same procedure works well whether the original graph slopes upward or downward and whether it is a straight line or a curved one. Figure A.5 sketches two examples. In panel (a), an increase in some third variable, Z , increases the value of Y for each value of X , so the graph of the relationship between X and Y shifts upward as Z increases. We often phrase it this way: "An increase in Z causes an increase in Y , *at any value of X .*" In panel (b), an increase in Z *decreases* the value of Y , at any value of X , so the graph of the relationship between X and Y shifts *downward* as Z increases.

SOLVING EQUATIONS

When we first derived the equation for the relationship between advertising and sales, we wanted to know what level of sales to expect from different amounts of advertising. But what if we're asked a slightly different question? Suppose, this time, you are told that the sales committee has set an ambitious goal of \$64,000 for next month's sales. The treasurer needs to know how much to budget for advertising, and you have to come up with the answer.

Since we know how advertising and sales are related, we ought to be able to answer this question. One way is just to look at the graph in Figure A.1. There, we could first locate sales of \$64,000 on the vertical axis. Then, if we read over to the line and then down, we find the amount of advertising that would be necessary to generate that level of sales. Yet even with that carefully drawn diagram, it is not always easy to see just exactly how much advertising would be required. If we need to be precise, we'd better use the equation for the graph instead.

According to the equation, sales (Y) and advertising (X) are related as follows:

$$Y = 40 + 3X.$$

In the problem before us, we know the value for sales, and we need to solve for the corresponding amount of advertising. Substituting the sales target of \$64,000 for Y , we need to find that value of X for which

$$64 = 40 + 3X.$$

Here, X is the unknown value for which we want to solve.

Whenever we solve one equation for one unknown, say, X , we need to *isolate* X on one side of the equals sign and everything else on the other side of the equals sign. We do this by performing identical operations on both sides of the equals sign. Here, we can first subtract 40 from both sides, getting

$$24 = 3X.$$

We can then divide both sides by 3 and get

$$8 = X.$$

This is our answer. If we want to achieve sales of \$64,000, we'll need to spend \$8,000 on advertising.

By looking back over what we just did, we can come up with a useful formula that will help to solve similar equations. Starting with an equation of the form

$$Y = a + bX,$$

we first subtracted a from both sides to get

$$Y - a = bX.$$

We then divided both sides by b to get our answer:

$$\frac{(Y - a)}{b} = X.$$

This is a formula you can use to solve for X whenever X and Y are linearly related and whenever b is not equal to zero. Of course, not all relationships are linear, so this formula will not work in every situation. But no matter what the underlying relationship, the idea remains the same:

To solve for X in any equation, rearrange the equation, following the rules of algebra, so that X appears on one side of the equals sign and everything else in the equation appears on the other side.

PERCENTAGE CHANGES

It is often convenient to express changes in percentage terms, rather than absolute terms. While we are all quite used to thinking in percentages, a quick review of how to calculate them may be helpful. If some variable X starts at one value and ends at another, the percentage change in X , denoted, $\% \Delta X$, is computed as follows:

$$\% \Delta X = \frac{\text{ending value of } X - \text{starting value of } X}{\text{starting value of } X} \times 100$$

Look at this formula for a moment. It says that, to calculate the *percentage* change in X , first compute the *change* in X by subtracting the ending value from the starting value, and then divide by the "base," or starting value, of X . The resulting fraction is then multiplied by 100. The formula shows us that:

*Whenever a variable decreases, the percentage change in its value will be negative.
Whenever a variable increases, the percentage change in its value will be positive.*

TABLE A.3

Variable	Beginning Value	Ending Value	Calculated Percentage Change
B	100	103	3%
C	20	21	5%
$B \times C$	2,000	2,163	8.15%
B/C	5	4.905	1.9%

RULES OF THUMB FOR PERCENTAGE CHANGES

Sometimes, we are interested in computing the percentage change in a product or a ratio. There are some useful rules of thumb that can simplify those computations. Specifically, we have:

Product Rule: If $A = B \times C$,
then $\% \Delta A = \% \Delta B + \% \Delta C$.

Quotient Rule: If $A = \frac{B}{C}$,
then $\% \Delta A = \% \Delta B - \% \Delta C$.

The product rule says that when A is the product of B and C , to find the percentage change in A , we simply *add* the percentage change in B to the percentage change in C . The quotient rule says that when A is the quotient B/C , to find the percentage change in A , simply *subtract* the percentage change in C from the percentage change in B .

Strictly speaking, these rules are *approximations*. They are most accurate when the percentage changes in B and C are extremely small. Yet as long as those percentage changes remain “relatively small,” the rules will provide “reasonably good” approximations. A few examples will help to convince you.

Suppose B rises from 100 to 103, while C rises from 20 to 21. To keep things straight, we’ve recorded the relevant data in Table A.3. The first two rows of the table record the beginning and ending values of B and C , and the percentage change in each variable. The last two rows show the beginning and ending values for the product $B \times C$ and the quotient B/C , respectively, and the percentage change in each of these, calculated exactly.

Now look at what we have. Moving across the third row, we see that $B \times C$ rises from 2,000 to 2,163, a percentage increase of 8.15% when computed exactly. Notice that this is very close to what we would get if, instead, we just applied our product rule, adding the 3% change in B to the 5% change in C to get an estimate of 8% for the change in the product $B \times C$. Thus, our approximation is very close. Similarly, moving across the fourth row, we find that the quotient B/C declines from 5 to 4.905, a percentage decrease of exactly 1.9%. Had we applied our quotient rule instead, we would have taken the 3% increase in B and subtracted the 5% increase in C to get $3\% - 5\% = -2\%$ —again, very close to the exact result of 1.9%.

CHAPTER

2

SCARCITY, CHOICE, AND ECONOMIC SYSTEMS

CHAPTER OUTLINE

The Concept of Opportunity Cost

Opportunity Cost for Individuals
Opportunity Cost and Society
Production Possibilities Frontiers
The Search for a Free Lunch

Economic Systems

Specialization and Exchange
Resource Allocation
Resource Ownership
Types of Economic Systems

Using the Theory: Are We Saving Lives Efficiently?

Opportunity cost The value of the best alternative sacrificed when taking an action.



<http://>

Is college worth the opportunity cost for you? Find out by trying Professor Jane Leuthold's COLLEGE CHOICE program at <http://www.cba.uiuc.edu/college/econ/choice/choice.html>.

What does it cost you to go to the movies? If you answered eight or nine dollars, because that is the price of a movie ticket, then you are leaving out a lot. Most of us are used to thinking of “cost” as the money we must pay for something. A Big Mac costs \$2.50, a new Toyota Corolla costs \$15,000, and the baby-sitter costs \$8.00 an hour. Certainly, the money we pay for a good or service is a *part* of its cost. But economics takes a broader view of costs, recognizing monetary as well as nonmonetary components.

THE CONCEPT OF OPPORTUNITY COST

The total cost of any choice we make—buying a car, producing a computer, or even reading a book—is everything we must *give up* when we take that action. This cost is called the *opportunity cost* of the action, because we give up the opportunity to have other desirable things.

The opportunity cost of any choice is all that we forego when we make that choice.

Opportunity cost is the most accurate and complete concept of cost—the one we should use when making our own decisions or analyzing the decisions of others.

OPPORTUNITY COST FOR INDIVIDUALS

Virtually every action we take as individuals uses up scarce money, scarce time, or both. Hence, every action we choose requires us to sacrifice other enjoyable goods and activities for which we could have used our money and time. For example, it took a substantial amount of the authors' time to write this textbook. Suppose that the time devoted to writing the book could instead have been used by one of the authors to either (1) go to law school, (2) write a novel, or (3) start a profitable business.

Do all three of these alternatives combined make up the opportunity cost of writing this book? Not really. Choosing not to write the book would have released some time but not enough time to pursue all three activities. To measure opportunity cost, we look only at the alternatives that *would* have been chosen—the ones

that are actually given up. Suppose that for one of the authors the next best alternative to writing this book was to start a profitable business. Then the opportunity cost of co-authoring this book was the foregone opportunity to start the business. Since the other, less valuable alternatives would not have been chosen anyway, they are not part of the cost of writing the book.

To explore this notion of opportunity cost further, let's go back to the earlier question: What does it cost to see a movie? That depends on *who* is seeing the movie. Suppose some friends ask Jessica, a college student, to go with them to a movie located 10 minutes from campus. To see the movie, Jessica will use up scarce *funds* to buy the movie ticket and scarce *time* traveling to and from the movie and sitting through it. Suppose the *money* she uses for the movie ticket would otherwise have been spent on a long-distance phone call to a friend in Italy—Jessica's next best use of the money—and the *time* would otherwise have been devoted to studying for her economics exam—her next best use of time. For Jessica, then, the opportunity cost of the movie consists of two things given up: (1) a phone call to her friend *and* (2) a higher score on her economics exam. Seeing the movie will require Jessica to sacrifice *both* of these valuable alternatives, since the movie will cost Jessica both money and time.

Now consider Samantha, a highly paid consultant who lives in New York City a few miles from the movie theater, and who has a backlog of projects to work on. As in Jessica's case, seeing the movie will use scarce funds and scarce time. But for Samantha, both costs will be greater. First, the direct money costs: There is not only the price of the movie ticket, but also the round-trip cab fare, which could bring the direct money cost to \$20. However, this is only a small part of Samantha's opportunity cost. Let's suppose that the time it takes Samantha to find out when and where the movie is playing, hail a cab, travel to the movie theater, wait in line, sit through the previews, watch the movie, and travel back home is three hours—not unrealistic for seeing a movie in Manhattan. Samantha's next best alternative for using her time would be to work on her consulting projects, for which she would earn \$150 per hour. In this case, we can measure the entire opportunity cost of the movie in monetary terms: first, the direct money costs of the movie and cab fare (\$20), and second, the foregone income associated with seeing the movie: ($\$150 \times 3 \text{ hours} = \450)—for a total of \$470!

At such a high price, you might wonder why Samantha would ever decide to see a movie. Indeed, the same reasoning applies to almost everything Samantha does besides work: It is very expensive for Samantha to talk to a friend on the phone, eat dinner, or even sleep. Each of these activities requires her to sacrifice the direct money costs plus another \$150 per hour of foregone income. Would Samantha ever choose to pursue any of these activities? The answer for Samantha is the same as for Jessica or anyone else: yes—if the activity is more highly valued than what is given up. It is not hard to imagine that, after putting in a long day at work, leisure activities would be very important to Samantha—worth the money cost *and* the foregone income required to enjoy them.

Once you understand the concept of opportunity cost and how it can differ among individuals, you can understand some behavior that might otherwise appear strange. For example, why do high-income



In some cases, the entire opportunity cost of a decision can be expressed as a single dollar figure. For example, Samantha's ticket, cab fare, and even the time spent at the movie are all easy to value in dollars (the value of the time is equal to the dollars Samantha could have earned at the next best alternative—working). But what if some part of opportunity cost *cannot* be easily measured in dollars? Then we simply express the opportunity cost as several different things, rather than a single number. For example, suppose that Samantha's next best alternative to the movie was not working, but attending a friend's birthday party instead. Then the opportunity cost of the movie would consist of both the *dollar* cost (ticket plus cab fare) *and* the missed birthday party.

people rarely shop at discount stores like Kmart and instead shop at full-service stores where the same items sell for much higher prices? It's not that high-income people *like* to pay more for their purchases. But discount stores are generally understaffed and crowded with customers, so shopping there takes more time. While discount stores have lower *money* cost, they impose a higher *time* cost. For high-income people, discount stores are actually more costly than stores with higher price tags.

We can also understand why the most highly paid consultants, entrepreneurs, attorneys, and surgeons often lead such frenetic lives, doing several things at once and packing every spare minute with tasks. Since these people can earn several hundred dollars for an hour of work, every activity they undertake carries a correspondingly high opportunity cost. Brushing one's teeth can cost \$10, and driving to work can cost hundreds! By combining activities—making phone calls while driving to work, thinking about and planning the day while in the shower, or reading the morning paper in the elevator—the opportunity cost of these routine activities can be reduced.

And what about the rest of us? As our wages rise, we all try to cram more activities into little bits of free time. Millions of Americans now carry cell phones and use them while waiting for an elevator or walking their dogs. Books on tape are becoming more popular and are especially favored by runners. (Why just exercise when you can also “read” a book?) And for some, vacations have become more exhausting than work, as more and more activities are crammed into shorter and shorter vacation periods.

OPPORTUNITY COST AND SOCIETY

For an individual, opportunity cost arises from the scarcity of time or money. But for society as a whole, opportunity cost arises from a different source: the scarcity of society's *resources*. Our desire for goods is limitless, but we have limited resources to produce them. Therefore,

all production carries an opportunity cost: To produce more of one thing, society must shift resources away from producing something else.

Let's discuss a goal on which we can all agree: better health for our citizens. What would be needed to achieve this goal? Perhaps more frequent medical check-ups for more people and greater access to top-flight medicine when necessary. These, in turn, would require more and better-trained doctors, more hospital buildings and laboratories, and more high-tech medical equipment such as MRI scanners and surgical lasers. In order for us to produce these goods and services, we would have to pull resources—land, labor, and capital—out of producing other things that we also enjoy. The opportunity cost of improved health care, then, consists of all the other goods and services we would have to do without.

PRODUCTION POSSIBILITIES FRONTIERS

Let's build a simple model to help us understand the opportunity cost we must pay for improved health care. To be even more specific, we'll measure production of health care by the *number of lives saved*. This variable is plotted along the horizontal axis in Figure 1. To measure the opportunity cost of health care, we'll make a simplifying assumption: that all goods *other* than life-saving health care can be lumped into a single category, and that we can measure how many units of these

THE PRODUCTION POSSIBILITIES FRONTIER

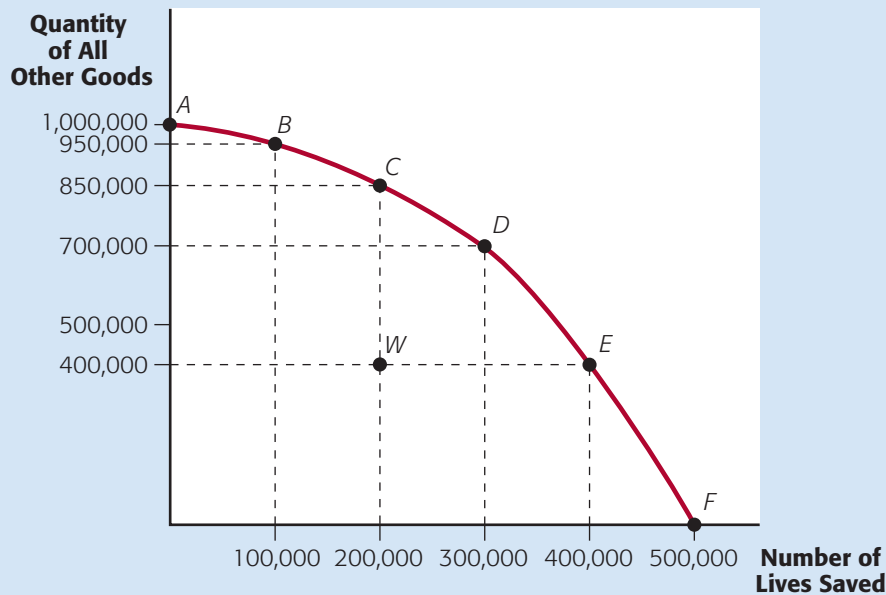


FIGURE 1

Points along a production possibilities frontier show combinations of two goods—here, lives saved and “other goods”—that can be produced using available resources and technology. At point *A*, all resources are used to produce other goods, and no lives are saved. At point *F*, 500,000 lives are saved, but no other goods are produced. The concave, bowed-out shape of the frontier reflects the law of increasing opportunity cost.

“other goods” we’re producing. In Figure 1, the quantity of “other goods” is measured on the vertical axis.

Now look at the curve drawn in Figure 1. It is society’s **production possibilities frontier (PPF)**, *giving the different combinations of goods that can be produced with the resources and technology currently available*. More specifically, this PPF tells us the *maximum quantity* of all other goods we can produce for each number of lives saved and the maximum number of lives saved for each different quantity of other goods. Positions outside the frontier are unattainable with the technology and resources at the economy’s disposal. Society’s choices are limited to points *on* or *inside* the PPF.

Let’s take a closer look at the PPF in Figure 1. Point *A* represents one possible choice for our society: to devote all resources to the production of “other goods” and none to health care. In this case, we would have 1,000,000 units of other goods, but we would have to forego every opportunity to save lives. Point *F* represents the opposite extreme: all available resources devoted to life-saving health care. In that case, we’d save 500,000 lives, but we’d have no other goods.

If points *A* and *F* seem absurd to you, remember that they represent two *possible* choices for society but choices we would be unlikely to make. We want life-saving health care to be available to those who need it, but we also want housing, clothing, entertainment, cars, and so on. So a realistic choice would include a *mix* of health care and movies.

Suppose we desire such a mix, but the economy, for some reason, is currently operating at the undesirable point *A*—no health care, but maximum production of everything else. Then we need to shift some resources from other goods to health care. For example, we could move from point *A* to point *B*, where we’d be saving 100,000 lives. But as a consequence, we’d have to cut back on other goods, producing 50,000 fewer units. The opportunity cost of saving 100,000 lives, then, would be 50,000 units of all other goods.

Production possibilities frontier (PPF) A curve showing all combinations of two goods that can be produced with the resources and technology currently available.

Increasing Opportunity Cost. Suppose we are at point *B*, and now we want to save even more lives. Once again, we shift enough resources into health care to save an additional 100,000 lives, moving from point *B* to point *C*. This time, however, there is an even *greater* cost: Production of other goods falls from 950,000 units to 850,000 units, or a sacrifice of 100,000 units. The opportunity cost of saving lives has risen. You can see that as we continue to save more lives—by increments of 100,000, moving from point *C* to point *D* to point *E* to point *F*—the opportunity cost of producing other goods keeps right on rising, until saving the last 100,000 lives costs us 400,000 units of other goods.

The behavior of opportunity cost described here—the more health care we produce, the greater the opportunity cost of producing still more—applies to a wide range of choices facing society. It can be generalized as the *law of increasing opportunity cost*.

Law of increasing opportunity cost

The more of something that is produced, the greater the opportunity cost of producing one more unit.

According to the law of increasing opportunity cost, the more of something we produce, the greater the opportunity cost of producing even more of it.

The law of increasing opportunity cost causes the PPF to have a *concave* shape, becoming steeper as we move rightward and downward. To understand why, remember (from high school math) that the slope of a line or curve is just the change along the vertical axis divided by the change along the horizontal axis. Along the PPF, as we move rightward, the slope is the change in the quantity of other goods divided by the change in the number of lives saved. This is a negative number, because a positive change in lives saved means a negative change in other goods. The absolute value of this slope is the opportunity cost of saving another life. Now—as we’ve seen—this opportunity cost increases as we move rightward. Therefore, the absolute value of the PPF’s slope must rise as well. The PPF gets steeper and steeper, giving us the concave shape we see in the Figure 1.¹

Why should there be a law of increasing opportunity cost? Why must it be that the more of something we produce, the greater the opportunity cost of producing still more?

Because most resources—*by their very nature*—are better suited to some purposes than to others. If the economy were operating at point *A*, for example, we’d be using all of our resources to produce other goods, including resources that are much better suited for health care. A hospital might be used as a food cannery, a surgical laser might be used for light shows, and a skilled surgeon might be driving a cab or trying desperately to make us laugh with his stand-up routine.

As we begin to move rightward along the PPF, say from *A* to *B*, we shift resources out of other goods and into health care. But we would *first* shift those resources *best suited to health care*—and *least* suited for the production of other things. For example, the first group of workers we’d use to save lives would be those who already have training as doctors and nurses. A surgeon—who would probably not make the best comedian—could now go back to surgery, which he does very well. Similarly, the first buildings we would put to use in the health care industry would be those that were originally built as hospitals and medical offices, and weren’t really doing so well as manufacturing plants, retail stores or movie studios. This is why, at first, the PPF is

¹ You might be wondering if the law of increasing opportunity cost applies in both directions. That is, does the opportunity cost of producing “other goods” increase as we produce more of them? The answer is yes, as you’ll see when you do Problem 2 at the end of this chapter.

very flat: We get a *large* increase in lives saved for only a *small* decrease in other goods.

As we continue moving rightward, however, we shift away from other goods those resources that are less and less suited to life-saving. As a result, the PPF becomes steeper. Finally, we arrive at point *F*, where all resources—no matter how well suited for other goods and services—are used to save lives. A factory building is converted into a hospital, your family car is used as an ambulance, and comedic actor Jim Carrey is in medical school, training to become a surgeon.

The principle of increasing opportunity cost applies to all of society's production choices, not just that between health care and other goods. If we look at society's choice between food and oil, we would find that some land is better suited to growing food and some land to drilling for oil. As we continue to produce more oil, we would find ourselves drilling on land that is less and less suited to producing oil, but better and better for producing food. The opportunity cost of producing additional oil will therefore increase. The same principle applies in choosing between civilian goods and military goods, between food and clothing, or between automobiles and public transportation: The more of something we produce, the greater the opportunity cost of producing still more.

THE SEARCH FOR A FREE LUNCH

This chapter has argued that every decision to produce *more* of something requires us to pay an opportunity cost by producing less of something else. Nobel Prize-winning economist Milton Friedman summarized this idea in his famous remark, "There is no such thing as a free lunch." Friedman was saying that, even if a meal is provided free of charge to someone, society still uses up resources to provide it. Therefore, a "free lunch" is not *really* free: Society pays an opportunity cost by not producing other things with those resources. The same logic applies to other supposedly "free" goods and services. From society's point of view, there is no such thing as a free airline flight, a free computer, or free medical care. Providing any of these things requires us to sacrifice other things, as illustrated by a movement *along* society's PPF.

But what if an economy is not living up to its productive potential, but is instead operating *inside* its PPF? For example, in Figure 1, suppose we are currently operating at point *W*, where the health care system is saving 200,000 lives and we are producing 400,000 units of other goods. Then we can move from point *W* to point *E* and save 200,000 more lives with no sacrifice of other goods. Or, starting at point *W*, we could move to point *C* (more of other goods with no sacrifice in lives saved) or to a point like *D* (more of *both* health care *and* other goods).

As you can see, if we are operating inside the PPF, Friedman's dictum does not apply—there *can* be such a thing as a free lunch! But why would an economy ever be operating inside its PPF? There are two possibilities.

Productive Inefficiency. One reason an economy might be operating inside its PPF is that resources are being wasted. Suppose, for example, that many people who could be outstanding health care workers are instead producing other goods, and many who would be great at producing other things are instead stuck in the health care industry. Then switching people from one job to the other could enable us to have more of *both* health care *and* other goods. That is, because of the mismatch of workers and jobs, we would be *inside* the PPF at a point like *W*. Creating better job matches would then move us to a point *on* the PPF (such as point *E*).

Economists use the phrase *productive inefficiency* to describe the type of waste that puts us inside our PPF.

Productive inefficiency A situation in which more of at least one good can be produced without sacrificing the production of any other good.

A firm, industry, or an entire economy is productively inefficient if it could produce more of at least one good without pulling resources from the production of any other good.

The phrase *productive efficiency* means the absence of any productive *inefficiency*. For example, if the computer industry is producing the maximum possible number of computers with the resources it is currently using, we would describe the computer industry as productively efficient. In that case, there would be no way to produce any more computers without pulling resources from the production of some other good. In order for an entire *economy* to be productively efficient, there must be no way to produce more of *any* good without pulling resources from the production of some other good.

Although no firm, industry, or economy is ever 100 percent productively efficient, cases of gross inefficiency are not as common as you might think. When you study microeconomics, you'll learn that business firms have strong incentives to identify and eliminate productive inefficiency, since any waste of resources increases their costs and decreases their profit. When one firm discovers a way to eliminate waste, others quickly follow.

For example, empty seats on an airline flight represent productive inefficiency. Since the plane is making the trip anyway, filling the empty seat would enable the airline to serve more people with the flight (produce more transportation services) without using any additional resources (other than the trivial resources of the airline meal). Therefore, more people could fly without sacrificing any other good or service. When American Airlines developed a computer model in the late 1980s to fill its empty seats by altering schedules and fares, the other airlines followed its example very rapidly. And when—in the late 1990s—a new firm called Priceline.com enabled airlines to auction off empty seats on the Internet, several airlines jumped at the chance, and others quickly followed. As a result of this—and similar efforts to eliminate waste in personnel, aircraft, and office space—many cases of productive inefficiency in the airline industry were eliminated.

The same sorts of efforts have eliminated some easy-to-identify cases of productive inefficiency in all types of industries: banking, telephone service, Internet service providers, book publishers, and so on. There are certainly instances of inefficiency that remain (an example appears at the end of this chapter). But on the whole, if you search the economy for a free lunch due to productive inefficiency, you won't find as many hearty meals as you might think.

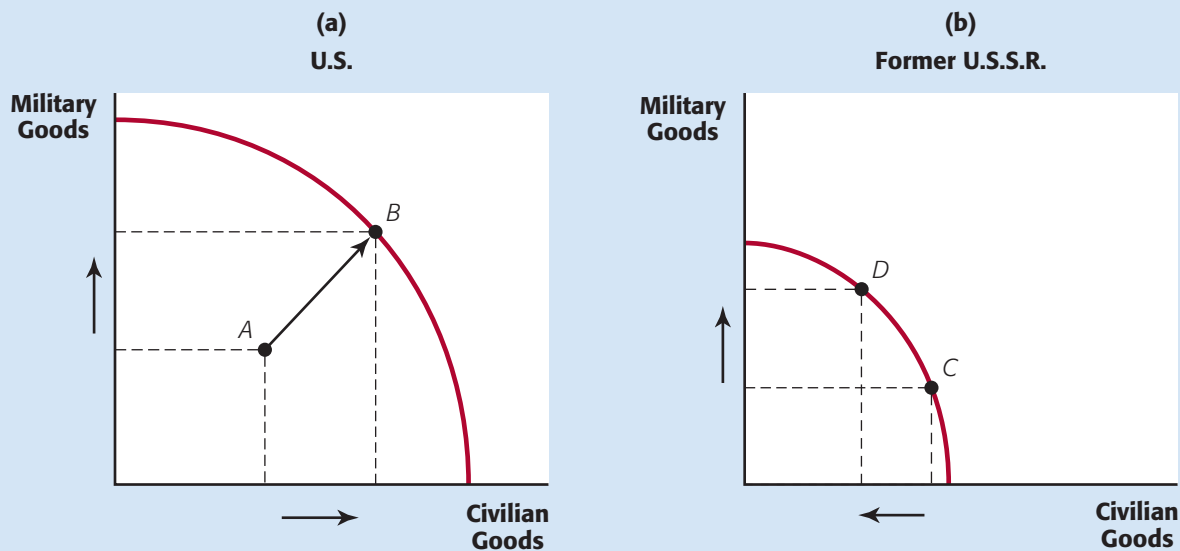
Recessions. Another situation in which an economy operates inside its PPF is a *recession*—a slowdown in overall economic activity. During recessions, many resources are idle. For one thing, there is widespread *unemployment*—people *want* to work but are unable to find jobs. In addition, factories shut down, so we are not using all of our available capital or land either. An end to the recession would move the economy from a point *inside* its PPF to a point *on* its PPF—using idle resources to produce more goods and services without sacrificing anything.

This simple observation can help us understand, in part, why the United States and the Soviet Union had such different economic experiences during World War II. In the Soviet Union, the average standard of living deteriorated considerably as the war began, but when the United States entered the war, living standards improved slightly. Why?

Figure 2 helps to solve this puzzle. The PPF in Figure 2 is like the PPF in Figure 1. But this time, instead of pitting “health care” against “all other goods,” we look at society's choice between *military* goods and *civilian* goods. When the United States

PRODUCTION AND UNEMPLOYMENT

FIGURE 2



At the onset of World War II, the U.S. economy was in a recession with high unemployment. This is shown by point A in panel (a), which is *inside* the production possibilities frontier. War production eliminated the unemployment as the United States moved onto its PPF at point B with more military goods *and* more civilian goods. The Soviet Union, by contrast, began the war with fully employed resources. It could increase military production only by sacrificing civilian goods and moving along its PPF from point C to point D.

entered the war in 1941, it was still suffering from the Great Depression—the most serious and long-lasting economic downturn in modern history, which began in 1929 and hit most of the developed world. For reasons you will learn when you study macroeconomics, joining the allied war effort helped end the Depression in the United States and moved our economy from a point like A, *inside* the PPF, to a point like B, *on* the frontier. Military production increased, but so did the production of civilian goods. Although there were shortages of some consumer goods, the overall result was a rise in the material well-being of the average U.S. citizen.

In the Soviet Union, things were very different. In the 1930s, the Soviet economy—which was internationally isolated—was able to escape entirely the effects of the depression that plagued the rest of the world. Thus, before the war, it was already operating on or near its PPF, at a point like C.² Entering the war—which meant an increase in military production—required a movement *along* its PPF, to a point like D. For the Soviet Union, the drop in civilian production—and the resulting drop in living standards—was the opportunity cost that had to be paid in order to fight the war.³

² Because its economic system caused major productive inefficiencies, some would argue that the Soviet Union was never actually on or even near its PPF. In Figure 2, however, we take the Soviet economic system as a given. Being on the PPF means the economy is producing the maximum civilian output for any given quantity of military output *and* for the given Soviet economic system.

³ There is another explanation for the decline in living standards in the Soviet Union, and it, too, can be illustrated with PPFs. Unlike the United States, large parts of the Soviet Union were decimated during World War II, decreasing the land and capital available for production of any kind. Similarly, the Soviet loss of human life was staggering—about 20 times greater than the loss of American lives. These huge decreases in land, labor, and capital shifted the Soviet PPF significantly *inward*—with fewer resources, civilian production would have to be smaller for any given level of military production.

An economic downturn, such as the Great Depression of the 1930s, does seem to offer a clear-cut free lunch. But eliminating a recession is not *entirely* costfree. When you study macroeconomics, you will see that while a variety of government policies can help to cure or avoid recessions, these same policies risk creating other problems of their own. Of course, we may feel that it is worth paying the cost to end a recession, but there is, nevertheless, a cost. Once again, a truly free lunch is not so easy to find.

ECONOMIC SYSTEMS

As you read these words—perhaps sitting at home or in the library—you are experiencing a very private moment. It is just you and this book; the rest of the world might as well not exist. Or so it seems. . . .

Actually, even in this supposedly private moment, you are connected to the rest of the world in ways you may not have thought about. In order for you to be reading this book, the authors had to write it. Someone (his name is Dennis Hanseman) had to edit it, to help make sure that all necessary material was covered and explained as clearly as possible. Someone else had to prepare the graphics. Others had to run the printing presses and the binding machines, and still others had to pack the book, ship it, unpack it, put it on a store shelf, and then sell it to you.

And there's more. People had to manufacture all kinds of goods: paper and ink, the boxes used for shipping, the computers used to keep track of inventory, and so on. It is no exaggeration to say that thousands of people were involved in putting this book in your hands.

And there is still more. The chair or couch on which you are sitting, the light shining on the page, the heat or the air conditioning in the room, the clothes you are wearing—all these things that you are using right now were *produced by somebody else*. So even now, as you sit alone reading this book, you are economically linked to others in hundreds—even thousands—of different ways.

Take a walk in your town or city, and you will see even more evidence of our economic interdependence: People are collecting garbage, helping schoolchildren cross the street, transporting furniture across town, constructing buildings, repairing roads, painting houses. Everyone is producing goods and services for *other people*.

Why is it that so much of what we consume is produced by other people? Why are we all so heavily dependent on each other for our material well-being? Why doesn't each of us—like Robinson Crusoe on his island—produce our own food, clothing, housing, and anything else we desire? And how did it come about that *you*—who did not produce any of these things yourself—are able to consume them?

These are all questions about our *economic system*—the way our economy is organized. Ordinarily, we take our economic system for granted, like the water that runs out of our faucets. But now it's time to take a closer look at the plumbing—to learn how our economy serves so many millions of people, enabling them to survive and prosper.

SPECIALIZATION AND EXCHANGE

If we were forced to, most of us could become economically *self-sufficient*. We could stake out a plot of land, grow our own food, make our own clothing, and build our own homes. But in no society is there such extreme self-sufficiency. On the contrary, every economic system over the past 10,000 years has been characterized by two features: (1) **specialization**, in which each of us concentrates on a limited number of pro-

Specialization A method of production in which each person concentrates on a limited number of activities.

ductive activities, and (2) **exchange**, in which most of what we desire is obtained by trading with others, rather than producing for ourselves.

Specialization and exchange enable us to enjoy greater production, and higher living standards, than would otherwise be possible. As a result, all economies exhibit high degrees of specialization and exchange.

There are three reasons why specialization and exchange enable us to enjoy greater production. The first has to do with human capabilities: Each of us can learn only so much in a lifetime. By limiting ourselves to a narrow set of tasks—fixing plumbing, managing workers, writing music, or designing Web pages—we are each able to hone our skills and become experts at one or two things, instead of remaining amateurs at a lot of things. It is easy to see that an economy of experts will produce more than an economy of amateurs.

A second gain from specialization results from the time needed to switch from one activity to another. When people specialize, and thus spend more time doing one task, there is less unproductive “downtime” from switching activities.

Before considering the third gain from specialization, it is important to note that these first two gains—acquiring expertise and minimizing downtime—would occur even if all workers were identical. To see why, let’s consider an extreme example. Suppose that three identical triplets—Sheri, Gerri, and Keri—decide to open up their own photocopy shop. They quickly discover that there are three primary tasks to be accomplished each day: making photocopies, dealing with customers, and servicing the machines.

Suppose first that the triplets decide *not* to specialize. Each time a customer walks in, *one* triplet will take the order, make the copies, collect the money, make the change, and give a receipt. In addition, each time a machine runs out of paper or ink, the triplet who is using the machine must remedy the problem. You can see that there will be a great deal of time spent going back and forth between the counter, the copy machines, and the supply room. Moreover, none of the triplets will become an expert at servicing the machines, dealing with customers, or making photocopies. As a result of the downtime between tasks and the lack of expertise, the triplets will not be able to make the maximum possible number of copies or handle the maximum possible number of customers each day.

Now, let’s rearrange production to take advantage of specialization. We’ll put Sheri at the counter, Gerri at the photocopy machine, and Keri keeping the machines in working order. Suddenly, all of that time spent going back and forth is now devoted to more productive tasks. Moreover, Sheri becomes an expert at working the cash register, since she does this all day long. Gerri becomes an expert at making copies, figuring out the quickest ways to select the proper settings, position originals, and turn pages. And Keri learns how to quickly diagnose and even anticipate problems with the machines. Each task is now performed by an expert. You can see that specialization increases the number of copies and customers that the triplets can handle each day, even though there is no difference in their basic abilities or talents.

Adam Smith first explained these gains from specialization in his book *An Inquiry into the Nature and Causes of the Wealth of Nations*, published in 1776. Smith explained how specialization within a pin factory dramatically increased the number of pins that could be produced there. In order to make a pin . . .

One man draws out the wire, another straightens it, a third cuts it, a fourth points it, a fifth grinds it at the top for receiving the head; to make the head requires three distinct operations; to put it on is a [separate] business, to

Exchange The act of trading with others to obtain what we desire.



Economics is a subject that has benefited from specialization and the division of labor. To get a feel for the many different subjects that economists investigate, take a look at the *Journal of Economic Literature's* classification system at <http://www.econlit.org/elcasbk.htm>.

whiten the pins is another; it is even a trade by itself to put them into the paper; and the important business of making a pin is, in this manner, divided into about eighteen distinct operations, which, in some manufactories, are all performed by distinct hands.

Smith went on to observe that 10 men, each working separately, might make 200 pins in a day, but through specialization, they were able to make 48,000! What is true for a pin factory or a photocopy shop can be generalized to the entire economy: Even when workers are identically suited to various tasks, total production will increase when workers specialize.

Of course, in the real world, workers are *not* identically suited to different kinds of work. Nor are all plots of land, all natural resources, or all types of capital equipment identically suited for different tasks. This observation brings us to the *third* source of gains from specialization.

Further Gains to Specialization: Comparative Advantage. Imagine a shipwreck in which there are only two survivors—let’s call them Maryanne and Gilligan—who wash up on opposite shores of a deserted island. Initially they are unaware of each other, so each is forced to become completely self-sufficient.

On one side of the island, Maryanne finds that it takes her one hour to pick one quart of berries or to catch one fish, as shown in the first row of Table 1. On the other side of the island, Gilligan—who is less adept at both tasks—requires an hour and a half to pick a quart of berries and three hours to catch one fish, as listed in the second row of the table. Since both castaways would want some variety in their diets, we can assume that each would spend part of the day catching fish and part picking berries.

Suppose that, one day, Maryanne and Gilligan discover each other. After rejoicing at the prospect of human companionship, they decide to develop a system of production that will work to their mutual benefit. Let’s rule out any gains from specialization that might arise from minimizing downtime or from becoming an expert, as occurred in the photocopy shop example. Will it still pay for these two to specialize? The answer is yes, as you will see after a small detour.

Absolute Advantage: A Detour. When Gilligan and Maryanne sit down to figure out who should do what, they might fall victim to a common mistake: basing their decision on *absolute advantage*. An individual has an **absolute advantage** in the production of some good when he or she can produce it using *fewer resources* than another individual can. On the island, the only resource being used is labor time, so the reasoning might go as follows: Maryanne can pick one quart of berries more quickly than Gilligan (see Table 1), so she has an *absolute advantage* in berry picking. It seems logical, then, that Maryanne should be the one to pick the berries.

But wait! Maryanne can also catch fish more quickly than Gilligan, so she has an absolute advantage in fishing as well. If absolute advantage is the criterion for

Absolute advantage The ability to produce a good or service, using fewer resources than other producers use.

TABLE 1

LABOR REQUIREMENTS FOR BERRIES AND FISH

	1 Quart of Berries	1 Fish
Maryanne	1 hour	1 hour
Gilligan	1½ hours	3 hours

assigning work, then Maryanne should do *both* tasks. This, however, would leave Gilligan doing nothing, which is certainly *not* in the pair's best interests. What can we conclude from this example? That absolute advantage is an unreliable guide for allocating tasks to different workers.

Comparative Advantage. The correct principle to guide the division of labor on the island is comparative advantage:

A person has a comparative advantage in producing some good if he or she can produce it with a smaller opportunity cost than some other person can.

Notice the important difference between absolute advantage and comparative advantage: You have an *absolute* advantage in producing a good if you can produce it using fewer *resources* than someone else can. But you have a *comparative* advantage if you can produce it with a smaller *opportunity cost*. As you'll see, these are not necessarily the same thing.

Table 2 shows the opportunity cost for each of the two castaways to produce berries and fish. For Maryanne, catching one fish takes an hour, time that could instead be used to pick one quart of berries. Thus, for her, the opportunity cost of one fish is one quart of berries. Similarly, her opportunity cost of one quart of berries is one fish. These opportunity costs are listed in the first row of Table 2. For Gilligan, catching one fish takes three hours, time that he could instead use to pick two quarts of berries. The opportunity cost of one fish for Gilligan, then, is two quarts of berries, and the opportunity cost of one quart of berries is one-half of a fish. (Of course, no one catches half a fish unless they are fishing with a machete, but we can still use this number to represent a rate of opportunity cost.) Comparing the two numbers, we see that Maryanne has the lower opportunity cost for one fish, so she has a *comparative advantage* in catching fish. But when we turn our attention to berry picking, we see that it is Gilligan who has the lower opportunity cost—half a fish. Therefore, Gilligan—who has an *absolute* advantage in nothing—has a *comparative* advantage in berry picking.

Let's see what happens as the two decide to move toward specializing according to comparative advantage. What happens each time Gilligan decides to catch one fewer fish? Table 2 tells us that he frees up enough time to pick 2 quarts of berries. We can write the results for Gilligan's production this way:

Gilligan: Fish ↓ 1 ⇒ Berries ↑ 2

Table 2 also tells us that each time Maryanne decides to catch one additional fish, she must sacrifice shift time away from berry picking, sacrificing 1 quart of berries:

Maryanne: Fish ↑ 1 ⇒ Berries ↓ 1

Comparative advantage The ability to produce a good or service at a lower opportunity cost than other producers.



Even castaways do better when they specialize and exchange with each other, instead of trying to be self-sufficient.

	Opportunity Cost of:	
	1 Quart of Berries	1 Fish
For Maryanne	1 fish	1 quart of berries
For Gilligan	½ fish	2 quarts of berries

TABLE 2

OPPORTUNITY COSTS

Now, what happens to total production on the island each time the pair moves toward producing according to comparative advantage? As you can see, Maryanne makes up for the fish that Gilligan is no longer catching. But Gilligan *more than makes up* for the quart of berries that Maryanne isn't picking. In fact, each time the two move toward specialization, fish production remains unchanged, whereas berry production increases. The gains continue until Maryanne is spending all of her work time fishing, and Gilligan is spending all of his work time picking berries.

Since—by producing according to comparative advantage—total production on the island increases, total *consumption* can increase, too. Gilligan and Maryanne can figure out some way of trading fish for berries that makes each of them come out ahead. In the end, each of the castaways will enjoy a higher standard of living when they specialize and exchange with each other, compared to the level they'd enjoy under self-sufficiency.

What is true for our shipwrecked island dwellers is also true for the entire economy:

Total production of every good or service will be greatest when individuals specialize according to their comparative advantage. This is another reason why specialization and exchange lead to higher living standards than does self-sufficiency.

When we turn from our fictional island to the real world, is production, in fact, consistent with the principle of comparative advantage? Indeed, it is. A journalist may be able to paint her house more quickly than a housepainter, giving her an *absolute* advantage in painting her home. Will she paint her own home? Except in unusual circumstances, no, because the journalist has a *comparative* advantage in writing news articles. Indeed, most journalists—like most college professors, attorneys, architects, and other professionals—hire house painters, leaving themselves more time to practice the professions in which they enjoy a comparative advantage.

Even comic book superheroes seem to behave consistently with comparative advantage. Superman can no doubt cook a meal, fix a car, chop wood, and do virtually *anything* faster than anyone else on the earth. Using our new vocabulary, we'd say that Superman has an absolute advantage in everything. But he has a clear comparative advantage in catching criminals and saving the known universe from destruction, which is exactly what he spends his time doing.

Specialization in Perspective. The gains from specialization, whether they arise from developing expertise, minimizing downtime, or exploiting comparative advantage, can explain many features of our economy. For example, college students need to select a major and then, upon graduating, to decide on a specific career. Those who follow this path are rewarded with higher incomes than those who dally. This is an encouragement to specialize. Society is better off if you specialize, since you will help the economy produce more, and society rewards you for this contribution with a higher income.

The gains from specialization can also explain why most of us end up working for business firms that employ dozens, or even hundreds or thousands, of other employees. Why do these business firms exist? Why isn't each of us a *self-employed* expert, exchanging our production with other self-employed experts? Part of the answer is that organizing production into business firms pushes the gains from specialization still further. Within a firm, some people can specialize in working

with their hands, others in managing people, others in marketing, and still others in keeping the books. Each firm is a kind of minisociety within which specialization occurs. The result is greater production and a higher standard of living than we would achieve if we were all self-employed.

Specialization has enabled societies everywhere to achieve standards of living unimaginable to our ancestors. But, if it goes too far, it can have a downside as well. In the old film *Modern Times*, Charlie Chaplin plays a poor soul standing at an assembly line, attaching part number 27 to part number 28 thousands of times a day. In the real world, specialization is rarely this extreme. Still, it has caused some jobs to be repetitive and boring. In some plants, workers are deliberately moved from one specialty to another to relieve boredom.

Of course, maximizing our material standard of living is not our only goal. In some instances, we might be better off *increasing* the variety of tasks we do each day, even if this means some sacrifice in production and income. For example, in many societies, one sex specializes in work outside the home and the other specializes in running the home and taking care of the children. Might families be better off if children had more access to *both* parents, even if this meant a somewhat lower family income? This is an important question. While specialization gives us material gains, there may be *opportunity costs* to be paid in the loss of other things we care about. The right amount of specialization can be found only by balancing the gains against these costs.

RESOURCE ALLOCATION

It was only 10,000 years ago—a mere blink of an eye in human history—that the Neolithic revolution began and human society switched from hunting and gathering to farming and simple manufacturing. At the same time, human wants grew beyond mere food and shelter to the infinite variety of things that can be *made*. Ever since, all societies have been confronted with three important questions:

1. *Which* goods and services should be produced with society's resources?
2. *How* should they be produced?
3. *Who* should get them?

Together, these three questions constitute the problem of **resource allocation**. The way a society chooses to answer these questions—that is, the method it chooses to allocate its resources—will in part determine the character of its economic system.

Let's first consider the *which* question. Should we produce more health care or more movies, more goods for consumers or more capital goods for businesses? Where on its production possibilities frontier should the economy operate? As you will see, there are different methods societies can use to answer these questions.

The *how* question is more complicated. Most goods and services can be produced in a variety of different ways, each method using more of some resources and less of others. For example, there are many ways to dig a ditch. We could use *no capital at all* and have dozens of workers digging with their bare hands. We could use *a small amount of capital* by giving each worker a shovel and thereby use less labor, since each worker would now be more productive. Or we could use *even more capital*—a power trencher—and dig the ditch with just one or two workers. In every economic system, there must always be some mechanism that determines how goods and services will be produced from the infinite variety of ways available.

Resource allocation A method of determining which goods and services will be produced, how they will be produced, and who will get them.

Finally, the *who* question. Here is where economics interacts most strongly with politics. There are so many ways to divide ourselves into groups: men and women, rich and poor, workers and owners, families and single people, young and old . . . the list is endless. How should the products of our economy be distributed among these different groups and among individuals within each group?

Determining *who* gets the economy's output is always the most controversial aspect of resource allocation. Over the last half-century, our society has become more sensitized to the way goods and services are distributed, and we increasingly ask whether that distribution is fair. For example, men get a disproportionately larger share of our national output than women do, whites get more than African-Americans and Hispanics, and middle-aged workers get more than the very old and the very young. As a society, we want to know *why* we observe these patterns (a positive economic question) and what we should do about them (a normative economic question). Our society is also increasingly focusing on the distribution of particular goods and services. Should scarce donor organs be rationed to those who have been waiting the longest, so that everyone has the same chance of survival? Or should they be sold to the highest bidder, so that those able to pay the most will get them? Should productions of Shakespeare's plays be subsidized by the government to permit more people—especially more poor people—to see them? Or should the people who enjoy these plays pay the full cost of their production?

The Three Methods of Resource Allocation. Throughout history, there have been three primary mechanisms for allocating resources. In a **traditional economy**, resources are allocated according to the long-lived practices of the past. Tradition was the dominant method of resource allocation for most of human history and remains strong in many tribal societies and small villages in parts of Africa, South America, Asia, and the Pacific. Typically, traditional methods of production are handed down by the village elders, and traditional principles of fairness govern the distribution of goods and services.

Economies in which resources are allocated largely by tradition tend to be stable and predictable. But they have one serious drawback: They don't grow. With everyone locked into the traditional patterns of production, there is little room for innovation and technological change. Traditional economies are therefore likely to be stagnant economies.

In a **command economy**, resources are allocated by explicit instructions from some higher authority. *Which* goods and services should we produce? The ones we're *ordered* to produce. *How* should we produce them? The way we're *told* to produce them. *Who* will get the goods and services? Whoever the authority *tells* us should get them.

In a command economy, a government body *plans* how resources will be allocated. That is why command economies are also called **centrally planned economies**. But command economies are disappearing fast. Until a few years ago, examples would have included the former Soviet Union, Poland, Rumania, Bulgaria, Albania, and many others. Beginning in the late 1980s, all of these nations have abandoned central planning. The only examples left are Cuba, China, and North Korea, and even these economies—though still dominated by central planning—are moving away from it.

The third method of allocating resources—and the one with which you are no doubt most familiar—is “the market.” In a **market economy**, neither long-held traditions nor commands from above guide our economic behavior. Instead, people are

Traditional economy An economy in which resources are allocated according to long-lived practices from the past.

Command or centrally planned economy An economic system in which resources are allocated according to explicit instructions from a central authority.

Market economy An economic system in which resources are allocated through individual decision making.

largely free to do what they want with the resources at their disposal. In the end, resources are allocated as a result of individual decision making. *Which* goods and services are produced? Whichever ones producers *choose* to produce. How are they produced? However producers *choose* to produce them. *Who* gets these goods and services? Anyone who *chooses* to buy them.

There are, of course, limitations on freedom of choice in a market economy. Some restrictions are imposed by government to ensure an orderly, just, and productive society. We cannot kill, steal, or break contracts—even if that is our desire—without suffering serious consequences. And we must pay taxes to fund government services. But the most important limitations we face in a market economy arise from the overall scarcity of resources.

This last point is crucial: In a market system, individuals are not simply free to do what they want. Rather, they are constrained by the resources they control. And in this respect, we do not all start in the same place in the economic race. Some of us—like the Rockefellers and the Kennedys—have inherited great wealth; some—entrepreneur Bill Gates, the novelist Toni Morrison, and the actress Julia Roberts—have inherited great intelligence, talent, or beauty; and some, such as the children of successful professionals, are born into a world of helpful personal contacts. Others, unfortunately, will inherit none of these advantages. In a market system, those who control more resources will have more choices available to them than those who control fewer resources. Still, in spite of the limitations imposed by government and the constraints imposed by limited resources, the market relies heavily on individual freedom of choice to allocate resources.

But wait . . . isn't there a problem here? People acting according to their own desires, without the firm hand of command or tradition to control them? This sounds like a recipe for chaos! How, in such a free-for-all, are resources actually *allocated*?

The answer is contained in two words: *markets* and *prices*.

The Nature of Markets. The market economy gets its name from something that virtually always happens when people are free to do what they want with the resources they possess. Inevitably, people decide to specialize in the production of one or a few things—often organizing themselves into business firms—and then sellers and buyers *come together to trade*. A **market** is a collection of buyers and sellers who have the potential to trade with one another.

Market A group of buyers and sellers with the potential to trade with each other.

In some cases, the market is *global*—that is, the market consists of buyers and sellers who are spread across the globe. The market for oil is an example of a global market, since buyers in any country can buy from sellers in any country. In other cases, the market is local. Markets for restaurant meals, haircuts, and taxi service are examples of local markets.

Markets play a major role in allocating resources by forcing individual decision makers to consider very carefully their decisions about buying and selling. They do so because of an important feature of every market: the *price* at which a good is bought and sold.

The Importance of Prices. A **price** is *the amount of money a buyer must pay to a seller for a good or service*. Price is not always the same as *cost*. In economics, as you've learned in this chapter, cost means *opportunity cost*—*all* that is sacrificed to buy the good. While the price of a good is a *part* of its opportunity cost, it is not the only cost. For example, the price does not include the value of the time sacrificed to buy something. Buying a new jacket will require you to spend time traveling to and

Price The amount of money that must be paid to a seller to obtain a good or service.

from the store, trying on different styles and sizes, and waiting in line at the cash register.

Still, in most cases, the price of a good is a significant part of its opportunity cost. For large purchases such as a home or automobile, the price will be *most* of the opportunity cost. And this is why prices are so important to the overall working of the economy: they confront individual decision makers with the costs of their choices.

Consider the example of purchasing a car. Because you must pay the price, you know that buying a new car will require you to cut back on purchases of other things. In this way, the opportunity cost to *society* of making another car is converted to an opportunity cost *for you*. If you value a new car more highly than the other things you must sacrifice for it, you will buy it. If not, you won't buy it.

Why is it so important that people face the opportunity costs of their actions? The following thought experiment can answer this question. Imagine that the government passed a new law: When anyone buys a new car, the government will reimburse that person for it immediately. The consequences would be easy to predict. First, on the day the law was passed, everyone would rush out to buy new cars. Why not, if cars are free? The entire stock of existing automobiles would be gone within days—maybe even hours. Many people who didn't value cars much at all, and who hardly ever used them, would find themselves owning several—one for each day of the week, or to match the different colors in their wardrobe. Others who weren't able to act in time—including some who desperately needed a new car for their work or to run their households—would be unable to find one at all.

Over time, automobile companies would step up their production to meet the surge in demand for cars, and then we would face another problem: the government's yearly "automobile budget," which would be hundreds of billions of dollars. Ultimately, we would all bear the cost of the increased car production, since the government would have to raise taxes. But we would pay as *taxpayers*, not as car owners. And our hefty tax bill would be supporting some rather frivolous uses for cars. Chances are, we would all be worse off because of this new policy. By eliminating a price for automobiles, and severing the connection between the opportunity cost of producing a car and the individual's decision to get one, we would have created quite a mess for ourselves.

When resources are allocated by the market, and people must pay for their purchases, they are forced to consider the opportunity cost to society of their individual actions. In this way, markets are able to create a sensible allocation of resources.

Resource Allocation in the United States. The United States has always been considered the leading example of a market economy. Each day, millions of distinct items are produced and sold in markets. Our grocery stores are always stocked with broccoli and tomato soup, and the drugstore always has Kleenex and aspirin—all due to the choices of individual producers and consumers. The goods that are traded, the way they are traded, and the price at which they trade are determined by the traders themselves. No direction from above is needed to keep markets working.

But even in the United States, there are numerous cases of resource allocation *outside* the market. For example, families are important institutions in the United States, and many economic decisions are made within them. Families tend to oper-

ate like traditional villages, not like market economies. After all, few families charge prices for goods and services provided inside the home.

Our economy also allocates some resources by command. Various levels of government collect, in total, about one-third of our incomes as taxes. We are *told* how much tax we must pay, and those who don't comply suffer serious penalties, including imprisonment. Government—rather than individual decision makers—spends the tax revenue. In this way, the government plays a major role in allocating resources—especially in determining which goods are produced and who gets them.

There are also other ways, aside from strict commands, that the government limits our market freedoms. Regulations designed to protect the environment, maintain safe workplaces, and ensure the safety of our food supply are just a few examples of government-imposed constraints on our individual choice.

What are we to make, then, of resource allocation in the United States? Markets are, indeed, constrained. But for each example we can find where resources are allocated by tradition or command, or where government restrictions seriously limit some market freedom, we can find hundreds of examples where individuals make choices according to their own desires. The things we buy, the jobs at which we work, the homes in which we live—in almost all cases, these result from market choices. The market, though not pure, is certainly the dominant method of resource allocation in the United States.

RESOURCE OWNERSHIP

So far, we've been concerned with how resources are allocated. Another important feature of an economic system is how resources are *owned*. The owner of a resource—a parcel of land, a factory, or one's own labor time—determines how it can be used and receives income when others use it. And there have been three primary modes of resource ownership in human history.

Under *communal* ownership, resources are owned by everyone—or by no one, depending on your point of view. They are simply there for the taking; no person or organization imposes any restrictions on their use or charges any fees. It is hard to find economies with significant communal ownership of resources. Karl Marx believed that, in time, all economies would evolve toward communal ownership, and he named this predicted system **communism**. In fact, none of the economies that called themselves Marxist (such as the former Soviet Union) ever achieved Marx's vision of communism. This is not surprising: Communal ownership on a broad scale can work only when individuals have no conflicts over how resources are used. Therefore, communism requires the end of *scarcity*—an unlikely prospect in the foreseeable future.

Nevertheless, there are examples of communal ownership on a smaller scale. Traditional villages maintain communal ownership of land and sometimes cattle. In some of the cooperative farms in Israel—called *kibbutzim*—land and capital are owned by all the members. Often there is a single television, a single kitchen, and a single children's playroom—all communally owned. Conflicts may result when individuals differ over how these resources should be used, but these conflicts are resolved by consensus, rather than by decree or by charging fees for their use.

Closer to home, most families operate on the principle of communal ownership. The house, television, telephone, and food in the refrigerator are treated as if owned jointly. More broadly, who “owns” our sidewalks, streets, and public beaches? No one does, really. In practice, all citizens are free to use them as much and as often as they would like. This is essentially communal ownership.

Communism A type of economic system in which most resources are owned in common.

Socialism A type of economic system in which most resources are owned by the state.

Capitalism A type of economic system in which most resources are owned privately.

Economic system A system of resource allocation and resource ownership.



The Center for International Comparisons at the University of Pennsylvania (<http://pwt.econ.upenn.edu/>) is a good source of information on the performance of economies around the world.

Under **socialism**, the *state* owns most of the resources. The prime example is the former Soviet Union, where the state owned all of the land and capital equipment in the country. In many ways, it also owned the labor of individual households, since it was virtually the only employer in the nation and unemployment was considered a crime.

State ownership also occurs in nonsocialist economies. In the United States, national parks, state highway systems, military bases, public colleges and universities, and government buildings are all state-owned resources. Over a third of the land in the country is owned by the federal government. The military, even under our current volunteer system, is an example in which the state owns the labor of soldiers—albeit for a limited period of time.

Finally, the third system. When most resources are owned *privately*—as in the United States—we have **capitalism**. Take the book you are reading right now. If you turn to the title page, you will see the imprint of South-Western College Publishing Company. This is a *private* company, owned by another company—Thomson Learning—that, in turn, is owned by *private* individuals. These individuals, in the end, own the facilities of South-Western: the buildings, the land under them, the office furniture and computer equipment, and even the reputation of the company. When these facilities are used to produce and sell a book, the private owners receive the income, mostly in the form of company profits. Similarly, the employees of South-Western are private individuals. They are *selling* a resource they own—their labor time—to South-Western, and they receive income—wages and salaries—in return.

The United States is one of the most capitalistic countries in the world. True, there are examples of state and communal ownership, as we've seen. But the dominant mode of resource ownership in the U.S. is *private* ownership. Resource owners keep *most* of the income from supplying their resources, and they have broad freedom in deciding how their resources are used.

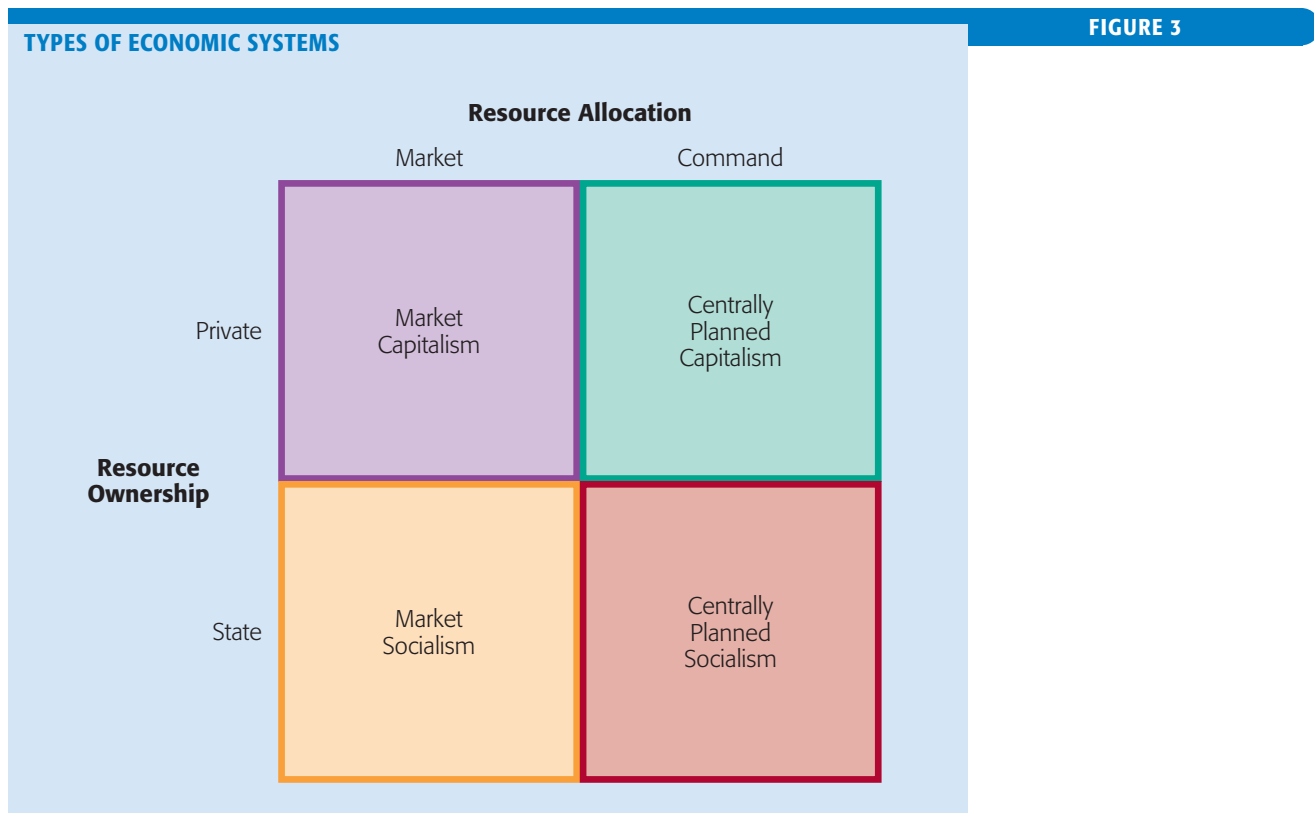
TYPES OF ECONOMIC SYSTEMS

We've used the phrase *economic system* a few times already in this book. But now it's time for a formal definition.

An economic system is composed of two features: a mechanism for allocating resources and a mode of resource ownership.

Let's leave aside the rare economies in which communal ownership is dominant and those in which resources are allocated primarily by tradition. That leaves us with four basic types of economic systems, indicated by the four quadrants in Figure 3. In the upper left quadrant, we have *market capitalism*. In this system, resources are *allocated* primarily by the market and *owned* primarily by private individuals. Today, most nations have market capitalist economies, including all of the countries of North America and Western Europe, and most of those in Asia, Latin America, and Africa.

In the lower right quadrant is *centrally planned socialism*, under which resources are mostly allocated by command and mostly owned by the state. This *was* the system in the former Soviet Union and the nations of Eastern Europe until the late 1980s. But in less than a decade, these countries' economies have gone through cataclysmic change, moving from the lower right quadrant to the upper left. That is, these nations have simultaneously changed both their method of resource allocation and their systems of resource ownership.



Although market capitalism and centrally planned socialism have been the two paramount economic systems in modern history, there have been others. The upper right quadrant represents a system of *centrally planned capitalism*, in which resources are owned by private individuals, yet allocated by command. In the recent past, countries such as Sweden and Japan—where the government has been more heavily involved in allocating resources than in the United States—have flirted with this type of system. Nations at war—like the United States during World War II—also move in this direction, as governments find it necessary to direct resources by command in order to ensure sufficient military production.

Finally, in the lower left quadrant is *market socialism*, in which resources are owned by the state yet allocated by the market mechanism. The possibility of market socialism has fascinated many social scientists, who believed it promised the best of both worlds: the freedom and efficiency of the market mechanism and the fairness and equity of socialism. There are, however, serious problems—many would say “unresolvable contradictions”—in trying to mix the two. The chief examples of market socialism in modern history were short-lived experiments—in Hungary and Yugoslavia in the 1950s and 1960s—in which the results were mixed at best.

Economic Systems and This Book. In this book, you will learn how market capitalist economies operate. This means that the other three types of economic systems in Figure 3 will be, for the most part, ignored. Until 10 years ago, these statements would have been accompanied by an apology that would have gone something like

this: “True, much of the world is characterized by alternative economic systems, but there is only so much time in one course . . .”

In the past decade, however, the world has changed dramatically: About 400 million people have come under the sway of the market as their nations have abandoned centrally planned socialism; another billion or so are being added as China changes course. The study of modern economies is now, more than ever before, the study of market capitalism.

Understanding the Market. The market is simultaneously the most simple and the most complex way to allocate resources. For individual buyers and sellers, the market is simple. There are no traditions or commands to be memorized and obeyed. Instead, we enter the markets we *wish* to trade in, and we respond to prices there as we *wish* to, unconcerned about the overall process of resource allocation.

But from the economist’s point of view, the market is quite complex. Resources are allocated indirectly, as a *by-product* of individual decision making, rather than through easily identified traditions or commands. As a result, it often takes some skillful economic detective work to determine just how individuals are behaving and how resources are being allocated as a consequence.

How can we make sense of all of this apparent chaos and complexity? That is what economics is all about. And you will begin your detective work in Chapter 3, where you will learn about the most widely used model in the field of economics: the model of supply and demand.

Using the THEORY



ARE WE SAVING LIVES EFFICIENTLY?

In the chapter, you learned that if resources are being wasted, we will operate *inside* our PPF rather than on the PPF. In that case, by eliminating the productive inefficiency, we would free up resources. Some of the resources could be used to save more lives and some to produce more of other goods. In Figure 1, this would move us from a point like *W* to a point like *D*, where we end up saving more lives *and* having more of other goods.

But there could also be productive inefficiency in the saving of human lives. If that is the case—if it is possible to save more lives without devoting any additional resources to doing so—then we would, once again, be operating inside our PPF. And once again, we could have a free lunch—save more lives *and* have more of other goods—by eliminating the inefficiency.

Some economists have argued that we do, indeed, waste significant amounts of resources in our life-saving efforts. How have they come to such a conclusion?

The first thing to remember is that saving a life—no matter how it is done—requires the use of resources. For any life-saving action we might take—putting another hundred police on the streets, building another emergency surgery center, or running an advertising campaign to encourage healthy living—we need certain quantities of resources, and a certain number of lives would be saved. In a market economy, resources sell at a price. This allows us to use the dollar cost of a life-saving method to measure the value of the resources used up by that method.

Moreover, we can compare the “cost per year of life saved” of different methods. For example, in the United States we currently spend about \$253 million on heart transplants each year and thereby add about 1,600 years to the lives of heart patients. Thus, the cost per year of life saved from heart transplants is $\$253,000,000/1,600 = \$158,000$ (rounded to the nearest thousand).

TABLE 3

THE COST OF SAVING LIVES

Method	Cost per Life-Year Saved
Brief physician antismoking intervention:	
Single personal warning from physician to stop smoking	\$150
Sickle cell screening and treatment for African-American newborns	\$236
Intensive physician anti-smoking intervention:	
Physician identification of smokers among their patients; 3 physician counseling sessions; 2 further sessions with smoking-cessation specialists; and materials—nicotine patch or nicotine gum	\$2,587
Mammograms: Once every 3 years, for ages 50–64	\$2,700
Mammograms: Annually, for ages 50–64	\$108,401
Exercise electrocardiograms as screening test:	
For 40-year-old males	\$124,374
Heart transplants	\$157,821
Mammograms: Annually, for age 40–49	\$186,635
Exercise electrocardiograms as screening test:	
For 40-year-old females	\$335,217
Heart Transplants	\$157,821
Seat belts on school buses	\$2,760,197
Anti-terrorist screening at airports	\$8,000,000
Asbestos ban in automatic transmissions	\$66,402,402

Sources: Electrocardiograms: Charles E. Phelps, *Health Economics*, 2nd ed. (Reading, MA: Addison-Wesley, 1997). Regular exercise: L. Goldman, A. M. Garber, S. A. Grover, & M. A. Hlatky (1996). Task Force 6. Cost-effectiveness of assessment and management of risk factors (Bethesda Conference). *JACC*, 27(5), 1020–1030. Anti-smoking intensive intervention: *Journal of the American Medical Association*, Dec. 3, 1997. Anti-smoking brief intervention: Malcolm Law and Jin Ling Tang, "An Analysis of the Effectiveness of Interventions Intended to Help People Stop Smoking," *Archives of Internal Medicine*, 1995; 155: pp. 1933–1941, and authors' calculations to convert "per life saved" to "per year of life saved." Annual mammograms: Kent Jeffreys, "Progressive Environmentalism: Principles for Regulatory Reform (Policy Report No. 194), National Center for Policy Analysis, June 1995. Benzene emission controls: Tammy O. Tengs et al., "Five Hundred Life-Saving Interventions and their Cost-Effectiveness," *Risk Analysis*, 1994. All other figures: Tammy O. Tengs, "Dying Too Soon: How Cost-Effectiveness Analysis Can Save Lives," School of Social Ecology, University of California, Irvine, NCPA Policy Report No. 204, May 1997. Anti-terrorist screening at airports: Robert W. Hahn, "The Cost of Anti-terrorist Rhetoric," *The Cato Review of Business and Government*, Dec. 17, 1996, and authors' calculations to convert "per life saved" to "per year of life saved."

Table 3 lists several of the methods we currently use to save lives in the United States. Some of these methods reflect legal or regulatory decisions (such as the ban on asbestos) and others reflect standard medical practices (such as annual mammograms for women over 50). Other methods are used only sporadically (such as seat belts in school buses). You can see that the cost per life saved ranges widely—from \$150 per year of life saved for a physician warning a patient to quit smoking, to over \$66,000,000 per year of life saved from the ban on asbestos in automatic transmissions.

The table indicates that some life-saving methods are highly efficient. For example, our society probably exhausts the potential to save lives from brief physician anti-smoking intervention. Most doctors *do* warn their smoking patients to quit.

But the table also indicates some serious productive *inefficiencies* in life saving. For example, screening and treating African-American newborns for sickle cell anemia is one of the least costly ways of saving a year of life in the United States—only

\$236 per year of life saved. Nevertheless, 20 percent of African-American newborns do *not* get this screening at all. Similarly, intensive intervention to discourage smoking is far from universal in the U.S. health care system, even though it has the relatively low cost of \$2,587 per year of life saved.

To get an idea of what this kind of productive inefficiency means, let's do some thought experiments. First, let's imagine that we shift resources from heart transplants to *intensive* antismoking efforts. Then for each year of life we decided *not* to save with heart transplants, we would free up \$157,821 in medical resources. If we applied those resources toward intensive antismoking efforts, at a cost of \$2,587 per year of life saved, we could then save an additional $\$157,821/\$2,587 = 61$ life years. In other words, we could increase the number of life-years saved without any increase in resources flowing to the health care sector, and therefore, without any sacrifice in other goods and services. A free lunch!

But why pick on heart transplants? Our ban on asbestos in automobile transmissions—which requires the purchase of more costly materials with greater quantities of scarce resources—costs us about \$66 million for each life-year saved. Suppose these funds were spent instead to buy the resources needed to provide women aged 40 to 49 with annual mammograms (currently *not* part of most physicians' recommendations). Then for each life-year lost to asbestos, we'd save $\$66\text{ million}/186,635 = 354$ life years from earlier detection of breast cancer.

The most surprising entry in the table may be the cost of the new antiterrorist screening procedures at airports, introduced in the late 1990s. The number relies on many critical assumptions. One is that, without current screening procedures, the number of fatalities from terrorist incidents on airlines would equal the rate we've had in the recent past—an average of 37 fatalities per year. If the rate would have *increased* without the new procedures, then the new procedures are actually saving more lives than assumed by economic studies, and the cost per life-year saved would be lower. On the other hand, the dollar figure assumes that current policies will be 100 percent effective in preventing fatal terrorist incidents. If this assumption is incorrect, the cost per life-year saved would be higher.

The largest component of the cost of antiterrorist screening is the increase in time it takes for the airlines to process luggage and passengers. This means a greater opportunity cost of time for passengers, who—on average—must arrive at the airport half an hour earlier. Suppose we value time at \$44 per hour. (This is not unreasonable, since higher-income people—especially business travelers—take more flights than lower-income people.) Then, each half-hour delay carries an opportunity cost for passengers of \$22. Multiplying that cost by 400 million annual passenger trips gives us a total opportunity cost of time of $\$22 \times 400\text{ million} = \8.8 billion per year. This is by far the largest cost of the new antiterrorist screening procedures. Together with about \$200 million of annual direct costs for equipment and personnel we get a total of about \$9 billion per year, which implies an expenditure of \$8 million per life-year saved.

What would happen if we applied this \$9 billion to other life-saving methods? You can answer that question on your own, using Table 3. You will see that there are, indeed, more efficient ways of spending our money.

Or are there?

It may be that these studies have left out a lot. For example, *why* do we spend so much on fighting airline terrorism when very few have died from it? The answer might be that the public exaggerates the risk. And if they do, then there are benefits from responding to this risk that go beyond the actual number of life-years saved.

For example, it may be that the new, enhanced safety procedures have convinced tens of thousands—maybe even hundreds of thousands—of travelers to fly rather than use other, slower forms of transportation. These travelers *save* time with the new screening procedures. Further, many travelers—who would otherwise experience serious anxiety while flying—no doubt benefit from increased peace of mind after seeing how carefully the airlines are trying to prevent terrorist attacks. How much is increased peace of mind worth to travelers? It's hard to say, but it should not be ignored.

One could make similar arguments about many environmental regulations, such as the ban on asbestos in auto transmissions. While their cost per life-year saved is exorbitant, they may have substantial—if intangible—benefits besides saving lives. (Can you imagine what some of these benefits might be?)

What can we conclude from all this? That life saving in the United States is no doubt plagued with productive inefficiencies. But the extent of the inefficiency is harder to measure than it appears at first glance.

S U M M A R Y

One of the most fundamental concepts in economics is *opportunity cost*. The opportunity cost of any choice is what we give up when we make that choice. At the individual level, opportunity cost arises from the scarcity of time or money; for society as a whole, it arises from the scarcity of resources—land, labor, and capital. To produce and enjoy more of one thing, we must shift resources away from producing something else. The correct measure of cost is not just the money price we pay, but the opportunity cost: everything we give up when we make a choice. The *law of increasing opportunity cost* tells us that the more of something we produce, the greater the opportunity cost of producing still more.

In a world of scarce resources, each society must have an economic system—its way of organizing economic activity.

All *economic systems* feature *specialization*, where each person and firm concentrates on a limited number of productive activities—and *exchange*, through which we obtain most of what we desire by trading with others. Specialization and exchange enable us to enjoy higher living standards than would be possible under self-sufficiency.

Every economic system determines how resources are owned and how they are allocated. In a market capitalist economy, resources are owned primarily by private individuals and allocated primarily through markets. Prices play an important role in markets by forcing decision makers to take account of society's opportunity cost when they make choices.

K E Y T E R M S

opportunity cost
production possibilities frontier (PPF)
law of increasing opportunity cost
productive inefficiency

specialization
exchange
absolute advantage
comparative advantage
resource allocation
traditional economy

command economy
centrally planned economy
market economy
market
price
communism

socialism
capitalism
economic system

R E V I E W Q U E S T I O N S

1. “Warren Buffett is one of the world's wealthiest men, worth billions of dollars. For someone like Buffet, the principle of opportunity cost simply doesn't apply.” True or false? Explain.
2. What are some reasons why a country might be operating inside its production possibilities frontier (PPF)?
3. Why is a PPF concave—that is, bowed out from the origin? Be sure to give an *economic* explanation.
4. What are three distinct reasons why specialization leads to a higher standard of living?

5. What is the difference between comparative advantage and absolute advantage? Which is more important from an economic viewpoint?
6. List the three questions any resource allocation mechanism must answer. Briefly describe the three primary methods of resource allocation that have evolved to answer these questions.
7. What are the three primary ways in which resources are *owned*? Briefly describe each of them.
8. Why can't the United States economy be described as a *pure market capitalist economy*?
9. True or false?: "Resource allocation and resource ownership are essentially the same thing. Once you know who owns the resources in an economy, you also know by what mechanism those resources will be allocated." Explain your answer.

P R O B L E M S A N D E X E R C I S E S

1. Suppose that you are considering what to do with an upcoming weekend. Here are your options, from least to most preferred: (1) Study for upcoming midterms; (2) fly to Colorado for a quick ski trip; (3) go into seclusion in your dorm room and try to improve your score on a computer game. What is the opportunity cost of a decision to play the computer game all weekend?
2. Redraw Figure 1, but this time identify a different set of points along the frontier. Starting at point *F* (500,000 lives saved, zero production of other goods), have each point you select show equal increments in the quantity of other goods produced. For example, point *H* should correspond to 200,000 units of other goods, point *J* to 400,000 units, point *K* to 600,000 units, and so on. Now observe what happens to the opportunity cost of "200,000 more units of other goods" as you move leftward and upward along this PPF. Does the law of increasing opportunity cost apply to the production of "all other goods"? Explain briefly.
3. How would a technological innovation in life saving—say, the discovery of a cure for cancer—affect the PPF in Figure 1? How would a technological innovation in the production of *other* goods—say, the invention of a new kind of robot that speeds up assembly-line manufacturing—affect the PPF?
4. You and a friend have decided to work jointly on a course project. Frankly, your friend is a less than ideal partner. His skills as a researcher are such that he can review and outline only two articles a day. Moreover, his hunt-and-peck style limits him to only 10 pages of typing a day. On the other hand, in a day you can produce six outlines or type 20 pages.
 - a. Who has an absolute advantage in outlining, you or your friend? What about typing?
 - b. Who has a comparative advantage in outlining? In typing?
 - c. According to the principle of comparative advantage, who should specialize in which task?
5. Suppose that one day, Gilligan (the castaway) eats a magical island plant that turns him into an expert at everything. In particular, it now takes him just half an hour to pick a quart of berries, and 15 minutes to catch a fish.
 - a. Redo Tables 1 and 2 in the chapter.
 - b. Who—Gilligan or Maryanne—has a comparative advantage in picking berries? In fishing? When the castaways discover each other, which of the two should specialize in which task?
 - c. Can *both* castaways benefit from Gilligan's new abilities? How?

C H A L L E N G E Q U E S T I O N

1. Suppose that an economy's PPF is a straight line, rather than a bowed out, concave curve. What would this say about the nature of opportunity cost as production is shifted from one good to the other?

EXPERIENTIAL EXERCISES

1. The transitional economies of Eastern Europe are often in the news as they shift from central planning to more of a market orientation. Take a look at the World Bank's Transition Newsletter at <http://www.worldbank.org/html/prddr/trans/WEB/trans.htm>. Choose one of these economies and try to determine how smoothly its transition is proceeding. What problems is that nation encountering? Do the problems seem to relate mostly to resource allocation, to resource ownership, or both?



2. The ability to measure the true cost of a choice is a skill that will pay you great dividends. Using Infotrac or a recent issue of the *Wall Street Journal*, try to find an article that discusses a decision some firm has made. Then review this chapter's section on "The Concept of Opportunity Cost." Finally, make a list of the kinds of cost involved in the firm's decision. Identify each item in your list as an explicit cost or an implicit cost.



CHAPTER

3

SUPPLY AND DEMAND

CHAPTER OUTLINE

Markets

- Defining the Good or Service
- Buyers and Sellers
- The Geography of the Market
- Competition in Markets
- Supply, Demand, and Market Definition

Demand

- The Law of Demand
- The Demand Schedule and the Demand Curve
- Changes in Quantity Demanded
- Changes in Demand

Supply

- The Law of Supply
- The Supply Schedule and the Supply Curve
- Changes in Quantity Supplied
- Changes in Supply

Putting Supply and Demand Together

What Happens When Things Change?

- An Ice Storm Hits the Northeast:
 - A Decrease in Supply
- Internet Entrepreneurs Get Rich:
 - An Increase in Demand
- The Market for Day Care:
 - Changes in Both Supply and Demand

The Four-Step Procedure

Using the Theory: Anticipating a Price Change

Father Guido Sarducci, a character on the early *Saturday Night Live* shows, once observed that the average person remembers only about five minutes worth of material from college. He therefore proposed the “Five Minute University,” where you’d learn only the five minutes of material you’d actually remember, and dispense with the rest. The economics course would last only 10 seconds, just enough time for students to learn to recite three words: “supply and demand.”

Of course, there is much more to economics than these three words. Still, Sarducci’s observation had some truth. Many people *do* regard the phrase “supply and demand” as synonymous with economics. But surprisingly few people actually understand what the phrase means. In a debate about health care, poverty, recent events in the stock market, or the high price of housing, you might hear someone say, “Well, it’s just a matter of supply and demand,” as a way of dismissing the issue entirely. Others use the phrase with an exaggerated reverence, as if supply and demand were an inviolable physical law, like gravity, about which nothing can be done. So what does this oft-repeated phrase really mean?

First, supply and demand is just an economic model—nothing more and nothing less. It’s a model designed to explain *how prices are determined in a market system*. Why has this model taken on such an exalted role in the field of economics? Because prices themselves play such an exalted role in the economy. In a market system, once the price of something has been determined, only those willing to pay that price will get it. Thus, prices determine which households will get which goods and services and which firms will get which resources. If you want to know why the cell phone industry is expanding while the video rental industry is shrinking, or why homelessness is a more pervasive problem in the United States than hunger, you need to understand how prices are determined. In this chapter, you will learn how the model of supply and demand works and how to use it. You will also learn about the strengths and limitations of the model. It will take more time than Guido Sarducci’s 10-second economics course, but in the end you will know much more than just three little words.

MARKETS

Put any compound in front of a chemist, ask him what it is and what it can be used for, and he will immediately think of the basic elements—carbon, hydrogen, oxygen, and so on. These elements are the basic building blocks of the materials we see in our world, and they help chemists make sense of what would otherwise appear rather chaotic.

Similarly, ask an economist almost any question about the economy, and he will immediately think about *markets*. As you learned in the last Chapter, the word *market* has a special meaning in economics.

A market is a group of buyers and sellers with the potential to trade.

Economists think of the economy as a collection of markets. In each one, the buyers and sellers will be different, depending on what is being traded. There is a market for oranges, another for automobiles, another for real estate, and still others for corporate stocks, French francs, and anything else that is bought and sold.

And this is where the choices begin. A market, as you'll soon see, is an important part of a supply and demand model, like a wing is an important part of a model airplane. And just as we can choose to make a wing out of balsa wood or plastic or metal—depending on our purpose—so, too, we have many choices when we define a market.

DEFINING THE GOOD OR SERVICE

Suppose we're interested in analyzing the computer industry in the United States. Should we define our market very broadly ("the market for computers"), very narrowly ("laptops under four pounds") or something in between ("portable personal computers")? Our choice will depend on the specific question we are trying to answer.

For example, if our goal is to predict how many households will be connected to the Internet by the year 2005, it would be best to combine all computers into one broad category, treating them all as if they were a single good. Economists call this process **aggregation**—combining a group of distinct things into a single whole. It would not do us much good to *disaggregate* computers into different types—desktops, laptops, handheld, faster than 450 Mhz, etc.—because such distinctions have little to do with Internet access and would only get in the way.

But suppose instead we are asking a different question: Why do laptops always cost more than desktops with similar computing power? Then we should use a slightly narrower definition of the product, aggregating all *laptops* together into one good, and all desktops together into another, and then looking at the markets for *each* of these more narrowly defined goods.

How broadly or narrowly we define a good or service is one of the choices that distinguishes *macroeconomics* from *microeconomics*. In macroeconomics, goods and services are aggregated to the highest levels. Macro models even lump all consumer goods—dishwashers, cell phones, blue jeans, and so forth—into the single category "consumption goods" and view them as if they are traded in a single, broadly defined market, "the market for consumption goods." Similarly, instead of recognizing different markets for shovels, bulldozers, computers, and factory buildings, macro models analyze the market for "capital goods." Defining goods in this very broad way allows macroeconomists to take an overall view of the economy without getting bogged down in the details.

Aggregation The process of combining distinct things into a single whole.

In microeconomics, by contrast, we are interested in more disaggregated goods. Instead of asking how much we'll spend on *consumer goods*, a microeconomist might ask how much we'll spend on *health care* or *video games*. Although microeconomics always involves some aggregation—combining different brands of laptop computers into one category, for example—in microeconomics, the process stops before it reaches the highest level of generality.

BUYERS AND SELLERS

A market is composed of the buyers and sellers that trade in it. But who, exactly, *are* these buyers and sellers?

When you think of a seller, your first image might be of a business. Indeed, in many markets, you'd be right: The sellers *are* business firms. Examples are markets for restaurant meals, airline travel, clothing, banking services, and video rentals. But businesses aren't the only sellers in the economy. In many markets, *households* are important sellers. For example, households are the primary sellers in labor markets, such as the markets for Web page designers, for accountants, and for factory workers. Households are also important sellers in markets for used cars, residential homes, and rare artworks. Governments, too, are sometimes important sellers. For example, state governments are major sellers in the market for education through state universities (such as the University of California, the University of Minnesota, and St. Louis Community College).

What about the other side of the market? When you think of *buyers*, your first thought may be “people” like yourself, or “households.” Indeed, many goods and services are bought primarily by households: college education, movies, housing, clothing, and so on. But here, too, the stereotype doesn't always fit. In labor markets, businesses and government agencies are the primary buyers. Businesses and government are also important buyers of personal computers, automobiles, and airline transportation.

As you can see, the buyers in a market can be households, business firms, or government agencies. And the same is true of sellers. Sometimes, it's important to recognize that all three groups are on both sides of a market. But not always. Once again, it depends on our purpose.

When the purpose is largely educational, greater simplification is permitted. For example, to understand *how* the price of paperback books is determined, we would in most cases assume that households are the only buyers. True, business firms and government libraries also buy paperback books. But including these buyers would only complicate our model, without changing any of our conclusions about price. On the other hand, if we wanted to precisely forecast the revenues of booksellers from paperback books, it would be dangerous to ignore orders from businesses and government libraries.

THE GEOGRAPHY OF THE MARKET

While a market itself is not an actual location, the participants in a market *do* live within some geographic area. When we speak of the geography of a market, we mean the geographic area within which the buyers and sellers are located.

It might appear that our choice of geography follows logically from the particular good or service we are analyzing. For example, think about crude oil. It is routinely transported across international waters and is freely traded among buyers and sellers in many different countries. So the market for oil should be a market of *global* buyers and sellers, right?

Not necessarily. Suppose we want to explain why oil is cheaper in the United States than in France? Then we'd need to define a *pair* of markets for oil and see how the price is determined in each one. In one market, global oil producers sell to buyers in France, and in another, the same producers sell to buyers in the United States. In each of these markets, global sellers trade with *national* buyers.

On the other hand, if we want to explain and forecast *world oil prices*, we'd gain little by distinguishing between French and American buyers. In this case, both sellers and buyers would be global.

In defining a market, we must choose the geographic area within which buyers and sellers are located. The buyers can be spread around the globe, or they can be a national, regional, or local group. The same is true of sellers. The geographic definition we choose depends on the specific question we are trying to answer.

COMPETITION IN MARKETS

A final issue in defining a market is how individual buyers and sellers view the price of the product. In many cases, individual buyers or sellers have an important influence over the market price. For example, in the market for cornflakes, Kellogg's—an individual *seller*—simply sets its price every few months. It can raise the price and sell fewer boxes of cereal, or lower the price and sell more. In the market for windshield wiper motors, Ford Motor Company—an individual *buyer*—can influence the price by negotiating special deals, or merely changing the number of motors it buys. The market for breakfast cereals and the market for windshield wiper motors are examples of *imperfectly competitive* markets.

In imperfectly competitive markets, individual buyers or sellers have some influence over the price of the product.

But now think about the national market for wheat. Can an individual seller have any impact on the market price? Not really. On any given day, there is a going price for wheat—say, \$5.80 per bushel. If a farmer tries to charge more than that—say, \$5.85 per bushel—he won't sell any wheat at all! His customers will instead go to one of his many competitors and buy the identical product from them. Each wheat farmer must take the price of wheat as a “given.”

The same is true of wheat *buyers*: If one tries to negotiate a lower price with a producer, he'd be laughed off the farm. “Why should I sell my wheat to you for \$5.75 per bushel, when there are others who will pay me \$5.80?” Accordingly, each buyer must take the market price as a given.

The market for wheat is an example of a *perfectly competitive market*.

In perfectly competitive markets (or just competitive markets), each buyer and seller takes the market price as a given.

What makes some markets imperfectly competitive and others perfectly competitive? You'll learn the complete answer when you are well into your study of *microeconomics*. One hint is that in perfectly competitive markets, there are many small buyers and sellers, and the product is standardized, like wheat. Imperfectly competitive markets, by contrast, have either a few large buyers or sellers, or else the product differs in important ways among different sellers.



The Inomics search engine is devoted solely to economics (<http://www.inomics.com/query/show?what=welcome>). Use it to investigate topics related to supply and demand.

Imperfectly competitive market A market in which a single buyer or seller has the power to influence the price of the product.

Perfectly competitive market A market in which no buyer or seller has the power to influence the price.

In the real world, perfectly competitive markets are rare. However, many markets come *close enough* that we can choose to view them as perfectly competitive. Think of the market for fast-food hotdogs in a big city. On the one hand, every hotdog stand is slightly different from every other. And each might be able to raise its price a bit above its competitors without losing all of its customers. For example, if his competitors are charging \$1.50 for a hotdog, the individual vendor might be able to charge \$1.60 or \$1.70. In these ways, the market for sidewalk hot dogs resembles *imperfect* competition.

But because there are so many other hotdog vendors in a big city, and because they are not *that* different from one another, no vendor can deviate too much from the going price of \$1.50. A vendor that charges \$1.80 or \$1.90, for example, might soon find himself without a business. So in some ways, the market is close to perfect competition.

How, then, do we decide whether to consider a market—such as the market for big-city hotdogs—as perfectly or imperfectly competitive? You won't be surprised to hear that it depends on the question we want to answer. If we want to explain why there are occasional price wars among hotdog vendors, or why some of them routinely charge higher prices than others, viewing the market as perfectly competitive would not work. To answer *these* questions, an individual seller's influence over his or her own price is important.

But if we want to know why hotdogs are cheaper than most other types of fast foods, the simplest approach is to view the market for hotdogs as perfectly competitive. True, each hotdog vendor does have *some* influence over the price. But that influence is so small, and the prices of different sellers are so similar, that our assumption of perfect competition works pretty well.

SUPPLY, DEMAND, AND MARKET DEFINITION

The supply and demand model—which explains how prices are determined in a market system—is a very versatile model. It can be applied to very broadly defined goods (the market for food) or very narrowly defined goods (the market for Granny Smith apples). Households, business firms, or government agencies can appear in any combination on the buying side or the selling side. The buyers and sellers can reside within a small geographic area or be dispersed around the world.

But there is only one restriction that is always implicit in any supply and demand analysis: We must always assume that the market is perfectly competitive.

The supply and demand model is designed to explain how prices are determined in perfectly competitive markets.

Does this mean we can only use the model when sellers and buyers have *no influence at all* over their price? Not really. As you've seen, perfect competition is a matter of degree, rather than an all-or-nothing characteristic. While there are very few markets in which sellers and buyers take the price as completely given, there are many markets in which a *narrow range* of prices is treated as a given (as in the market for hotdogs). In these markets, supply and demand often provides a good approximation to what is going on. This is why it has proven to be the most versatile and widely used model in the economist's tool kit. Neither laptop computers nor orange juice is traded in a perfectly competitive market. But ask an economist to tell you why the cost of laptops decreases every year, or why the price of orange juice rises after a freeze in Florida, and he or she will invariably reach for supply and demand to find the answer.

Supply and demand are like two blades of a scissors: The demand blade tells us how much of something buyers want to buy, and the supply blade tells us how much sellers want to sell. To analyze a market, we need both blades—and they must both be sharp. In this and the next section, we will be sharpening those blades, learning separately about supply and demand. Then, when we have a thorough understanding of each one, we'll put them together—and put them to use. Let's start with demand.

DEMAND

When you come to a market as a buyer, what is your goal? In the most general terms, it's to make yourself as well off as possible. Then why don't you try to buy up everything you can in every possible market? After all, you'd be better off if you had more clothes, more airline travel, a bigger home or apartment, a faster Internet connection. . . . If your goal is to make yourself as well off as possible, you should try to grab up all these things. Right?

Not really. Because in addition to having a goal, you also face *constraints*. First, everything you want to buy has a *price*. Second, you have a limited income with which to buy things. As a result of these two constraints—prices and your limited income—whenever you decide to buy something, you must give up something else that you *could have bought* instead. That is, every purchase carries an opportunity cost. (Even if you have more income each year than you spend, you still pay an opportunity cost when you buy something because you will *save* less that year.)

Both the goals and the constraints of buyers like you play a role in determining the demand side of a market. That is why we do *not* define the quantity of a product demanded as how much a buyer would *like* to have if he could snap his fingers and just have it. Rather, it's how much he would actually *choose* to buy given the constraints that he faces.

An individual's quantity demanded of any good is the total amount that individual would choose to buy at a particular price.

When we turn our attention to demand in the market as a whole, we define a similar concept.

The market quantity demanded of any good is the total amount that all buyers in the market would decide to buy at a particular price.

Notice two very important things about this definition. First, it refers to buyers' *choices*, not to the amount that buyers will *actually* buy. Will buyers, in fact, be *able* to buy what they decide to buy? Or will they be frustrated in their attempts because sellers are not supplying enough? This is a very important question but one that can't be answered until buyers and sellers—demand *and* supply—come together in the market. That will happen a little later in this chapter.

Second, notice that the influence of price is stressed in the definition of quantity demanded. This is for a good reason. The supply and demand model, you recall, is designed to explain how *prices* are determined in perfectly competitive markets. It seems natural, then, to begin our exploration of demand with the influence of prices.

THE LAW OF DEMAND

How does a change in price affect quantity demanded? You probably know the answer to this already: When something is more expensive, people buy less of it.

Individual's quantity demanded

The total amount of a good an individual would choose to purchase at a given price.

Market quantity demanded

The total amount of a good that all buyers in the market would choose to purchase at a given price.

This common observation applies to walnuts, air travel, magazines, education, and virtually everything else that people buy. For all of these goods and services, price and quantity are *negatively related*—that is, when price rises, quantity demanded falls; when price falls, quantity demanded rises. This negative relationship is observed so regularly in markets that economists call it the *law of demand*.

Law of demand As the price of a good increases, the quantity demanded decreases.

The law of demand states that when the price of a good rises and everything else remains the same, the quantity of the good demanded will fall.

Read that definition again, and notice the very important words “everything else remains the same.” The law of demand tells us what would happen *if* all the other influences on buyers’ choices remained unchanged, and only one influence—the price of the good—changed.

This is an example of a common practice in economics. In the real world, many variables change *simultaneously*. But to understand the economy, we must understand the effect of each variable *separately*. Imagine that you were trying to discover which headache remedy works best for you. You wouldn’t gain much information if you took an Advil, a Tylenol, and an aspirin tablet all at the same time. Instead, you should take just *one* of these pills the next time you get a headache and observe its effects. To understand the economy, we go through the same process—conducting mental experiments in which only one thing changes at a time. The law of demand tells us what happens when we change *just* the price of the good, and assume that all other influences on buyers’ choices remain constant.

THE DEMAND SCHEDULE AND THE DEMAND CURVE

To make our discussion more concrete, let’s look at a specific market: the market for real maple syrup in Wichita, Kansas. In this market, the buyers are all residents of Wichita, whereas the sellers (to be considered later) are maple syrup producers in the United States or Canada.

Demand schedule A list showing the quantities of a good that consumers would choose to purchase at different prices, with all other variables held constant.

Table 1 shows a hypothetical **demand schedule** for maple syrup in this market. This is *a list of different quantities demanded at different prices, with all other variables that affect the demand decision assumed constant*. For example, the demand schedule tells us that when the price of maple syrup is \$2.00 per bottle, the quantity demanded will be 6,000 bottles per month. Notice that the demand schedule obeys the law of demand: As the price of maple syrup increases, the quantity demanded falls.

Now look at Figure 1. It shows a diagram that will appear again and again in your study of economics. In the figure, each price-and-quantity combination in Table 1 is represented by a point. For example, point *A* represents the price \$4.00 and quantity 4,000, while point *B* represents the pair \$2.00 and 6,000. When we

TABLE 1

DEMAND SCHEDULE FOR MAPLE SYRUP IN WICHITA	Price (per Bottle)	Quantity Demanded (Bottles per Month)
	\$1.00	7,500
	2.00	6,000
	3.00	5,000
	4.00	4,000
	5.00	3,500

THE DEMAND CURVE

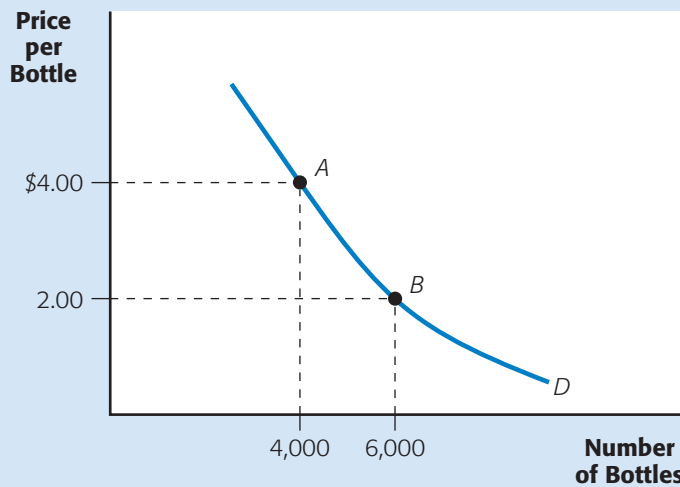


FIGURE 1

The downward-sloping demand curve, D , shows the quantity of maple syrup that would be purchased at each price, holding constant all other variables affecting demand. At \$4.00 per bottle, 4,000 bottles of syrup are demanded (point A). At \$2.00 per bottle, 6,000 bottles are demanded (point B).

connect all of these points with a line, we obtain the famous *demand curve*, labeled with a D in the figure.

The market demand curve (or just demand curve) shows the relationship between the price of a good and the quantity demanded, holding constant all other variables that affect demand. Each point on the curve shows the total quantity that buyers would choose to buy at a specific price.

The demand curve for maple syrup in Figure 1—like virtually all demand curves we might observe—follows the law of demand: A rise in the price of the good causes a decrease in the quantity demanded. Graphically, the law of demand tells us that demand curves slope downward.

CHANGES IN QUANTITY DEMANDED

Markets are affected by a variety of different events. Some events will cause us to *move along* the demand curve for a good. Other events will cause the entire demand curve to *shift*. It is crucial to distinguish between these two very different effects on demand, and economists have adopted a language convention that helps us keep track of the distinction.

Let's go back to Figure 1. There, you can see that if the price of maple syrup rises from \$2.00 to \$4.00 per bottle, the number of bottles demanded falls from 6,000 to 4,000. This is a movement *along* the demand curve, from point B to point A , and we call it a *decrease in quantity demanded*. More generally,

a change in a good's price causes us to move along the demand curve. We call this a change in quantity demanded. A rise in price causes a leftward movement along the demand curve—a decrease in quantity demanded. A fall in price causes a rightward movement along the demand curve—an increase in quantity demanded.

Market demand curve The graphical depiction of a demand schedule; a curve showing the quantity of a good or service demanded at various prices, with all other variables held constant.

Change in quantity demanded A movement along a demand curve in response to a change in price.

TABLE 2

INCREASE IN DEMAND FOR MAPLE SYRUP IN WICHITA

Price (per Bottle)	Original Quantity Demanded (Bottles per Month)	New Quantity Demanded After Increase in Income (Bottles per Month)
\$1.00	7,500	9,500
2.00	6,000	8,000
3.00	5,000	7,000
4.00	4,000	6,000
5.00	3,500	5,500

CHANGES IN DEMAND

Whenever we draw a demand curve, we are always assuming something about the other variables that affect buyers' choices. For example, the demand curve in Figure 1 might tell us the quantity demanded at each price, *assuming* that average household income in Wichita is \$40,000. In the real world, of course, the average household income in Wichita might change—say, from \$40,000 to \$45,000. What would happen? With more income, we would expect households to buy more of *most* things, including maple syrup. This is illustrated in Table 2. At the original income level, households would choose to buy 6,000 bottles of maple syrup if the price is \$2.00 per bottle. But after income rises, they would choose to buy 8,000 bottles at that same price. The same holds for any other price for maple syrup: after income rises, households will choose to buy more than before. In other words, *the entire relationship between price and quantity demanded has changed.*

Figure 2 plots the new demand curve from the quantities in the third column of Table 2. The new demand curve lies to the *right* of the old curve. For example, at a price of \$2.00, the old demand curve told us that the quantity demanded was 6,000 bottles (point B). But after the increase in income, buyers would want to buy 8,000 bottles at that price (point C). Notice that the rise in household income has *shifted the demand curve to the right*. We call this an *increase in demand*, because the word *demand* means the entire relationship between price and quantity demanded.

More generally,

Change in demand A shift of a demand curve in response to a change in some variable other than price.

a change in any determinant of demand—except for the good's price—causes the demand curve to shift. We call this a change in demand. If buyers choose to purchase more at any price, the demand curve shifts rightward—an increase in demand. If buyers choose to purchase less at any price, the demand curve shifts leftward—a decrease in demand.



Language is important when speaking about demand. If you say, "People demand more maple syrup," you might mean that we are moving along the demand curve, like the move from point A to point B in Figure 1. Or you might mean that the entire demand curve has shifted, like the shift from D_1 to D_2 in Figure 2.

To avoid confusion (and mistakes on exams!), always use the special language that distinguishes between these two cases. When we *move along* the demand curve, we call it a *change in quantity demanded*. A change in quantity demanded is always caused by a change in the good's price. But when the entire demand curve shifts, we call it a *change in demand*. A change in demand is always caused by a change in something *other* than the good's price.

Now let's look at the different variables that can cause demand to change and shift the demand curve.

Income and Wealth. Your **income** is what you earn over a period of time—say, \$3,000 per month or \$36,000 per year. Your **wealth**—if you are fortunate enough to have some—is

A SHIFT OF THE DEMAND CURVE

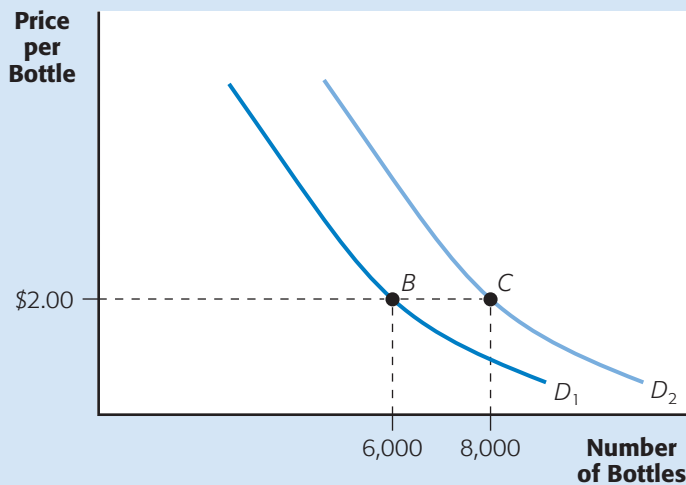


FIGURE 2

A change in any influence on demand besides the price of the good causes the entire demand curve to shift. An increase in income, for example, causes the demand for maple syrup, a normal good, to shift from D_1 to D_2 . At each price, more bottles are demanded after the shift.

the total value of everything you own (cash, bank accounts, stocks, bonds, real estate, valuable artwork, or any other valuable property) minus everything you owe (home mortgage, credit card debt, auto loan, student loans, and so on).

You've already seen (in Table 2 and Figure 2) how an increase in income would increase the demand for maple syrup. And while income and wealth are different things, they have similar effects on demand. If someone's wealth increases—say, through inheritance or an increase in the value of their stocks or bonds—they tend to respond just as if their income had increased, even if their income remains unchanged.

A rise in either income or wealth increases the demand for most goods. We call these **normal goods**. Housing, airline travel, health club memberships and maple syrup are all examples of normal goods.

The demand for most goods (normal goods) is positively related to income or wealth. A rise in either income or wealth will increase demand for these goods, and shift the demand curve to the right.

But not all goods' demand curves behave this way. For some goods—called **inferior goods**—a rise in income or wealth will *decrease* demand. Ground chuck is one example. It's a cheap source of protein, but not most people's idea of a fine dining experience. Higher income or wealth would enable consumers of ground chuck to afford more steaks, decreasing their demand for ground chuck. For similar reasons, Greyhound bus tickets, low-rent housing units, and single-ply paper towels are probably inferior goods. For all of these goods, an increase in consumers' income or wealth would decrease demand, shifting the demand curve to the left.

Prices of Related Goods. A **substitute** is a good that can be used in place of another good and that fulfills more or less the same purpose. For example, many people use maple syrup to sweeten their pancakes, but they could use a number of other

Income The amount that a person or firm earns over a particular period.

Wealth The total value of everything a person or firm owns, at a point in time, minus the total value of everything owed.

Normal good A good that people demand more of as their income rises.

Inferior good A good that people demand less of as their income rises.

Substitute A good that can be used in place of some other good and that fulfills more or less the same purpose.

things instead: honey, sugar, fruit, or jam. Each of these can be considered a substitute for maple syrup.

When the price of a substitute rises, people will choose to buy *more* of the good itself. For example, when the price of jam rises, some jam users will switch to maple syrup, and the demand for maple syrup will increase. In general,

when the price of a substitute rises, the demand for a good will increase, shifting the demand curve to the right.

Of course, if the price of a substitute falls, we have the opposite result: Demand for the original good decreases, shifting its demand curve to the left.

There are countless examples in which a change in a substitute's price affects demand for a good. A rise in the price of postage stamps would increase the demand for electronic mail. A drop in the rental price of videos would decrease the demand for movies at theaters. In each of these cases, we assume that the price of the substitute is the only price that is changing.

A **complement** is the opposite of a substitute: It's used *together with* the good we are interested in. Pancake mix is a complement to maple syrup, since these two goods are used frequently in combination. If the price of pancake mix rises, some consumers will switch to other breakfasts—bacon and eggs, for example—that *don't* include maple syrup. The demand for maple syrup will decrease.

A rise in the price of a complement decreases the demand for a good, shifting the demand curve to the left.

This is why we expect a higher price for automobiles to decrease the demand for gasoline and a lower price for movie tickets to increase the demand for movie theater popcorn.

Population. As the population increases in an area, the number of buyers will ordinarily increase as well, and the demand for a good will increase. The growth of the U.S. population over the last 50 years has been an important reason (but not the only reason) for rightward shifts in the demand curves for food, rental apartments, telephones, and many other goods and services.

Expectations. Expectations of future events—especially future changes in a good's price—can affect demand. For example, if buyers expect the price of maple syrup to rise next month, they may choose to purchase more *now* to stock up before the price hike. The demand curve would shift to the right. If people expect the price to drop, they may postpone buying, hoping to take advantage of the lower price later. This would shift the demand curve leftward.

Expectations are particularly important in the markets for financial assets such as stocks and bonds and in the market for real estate. People want to buy more stocks, bonds, and real estate when they think their prices will rise in the near future. This shifts the demand curves for these items to the right.

Tastes. Suppose we know the number of buyers in Wichita, their expectations about the future price of maple syrup, the prices of all related goods, and the average levels of income and wealth. Do we have all the information we need to draw the demand curve for maple syrup in Wichita? Not really. Because we do not yet know how consumers there *feel* about maple syrup. How many of them eat break-

Complement A good that is used *together with* some other good.

fast? Of these, how many eat pancakes or waffles? How often? How many of them *like* maple syrup, and how much do they like it? And what about all of the other goods and services competing for Wichita consumers' dollars: How do buyers feel about *them*?

The questions could go on and on, pinpointing various characteristics about buyers

that influence their attitudes toward maple syrup. The approach of economics is to lump all of these characteristics of buyers together and call them, simply, *tastes*. Economists do not try to explain where these tastes come from or what makes them change. These tasks are left to other social scientists—psychologists, sociologists, and anthropologists. Instead, economists concern themselves with the *consequences* of a change in tastes, whatever the reason for its occurrence.

When tastes change *toward* a good (people favor it more), demand increases, and the demand curve shifts to the right. When tastes change *away* from a good, demand decreases, and the demand curve shifts to the left. An example of this is the change in tastes away from cigarettes over the past several decades. The cause may have been an aging population, a greater concerns about health among people of *all* ages, or successful antismoking advertising. But regardless of the cause, the effect has been to decrease the demand for cigarettes, shifting the demand curve to the left.

Figure 3 summarizes the important variables that affect the demand side of the market, and how their effects are represented with a demand curve. Notice the important distinction between movements *along* the demand curve and *shifts* of the entire curve.

SUPPLY

Now we switch our focus from the buying side to the selling side of the market. When we discussed demand, we noted that each buyer comes to a market with a goal—to make himself as well off as possible. But the buyer also faces a constraint: He must pay for purchases out of a limited income.

A seller, too, comes to a market with a goal—to make as much profit as possible. And if the seller is a business firm (which we'll assume for most of this chapter), it faces an important constraint: Producing output (goods and services) requires the use of inputs. The quantities of those inputs needed are determined by the firm's *production technology*.

A firm's production technology (or just technology) is the set of methods it can use to turn inputs (resources and raw materials) into outputs (goods or services).

Technology The set of methods a firm can use to turn inputs into outputs

Continuing with our example, there are many different ways for a maple syrup farm to produce its output (maple syrup) from its inputs (land, maple trees, labor,

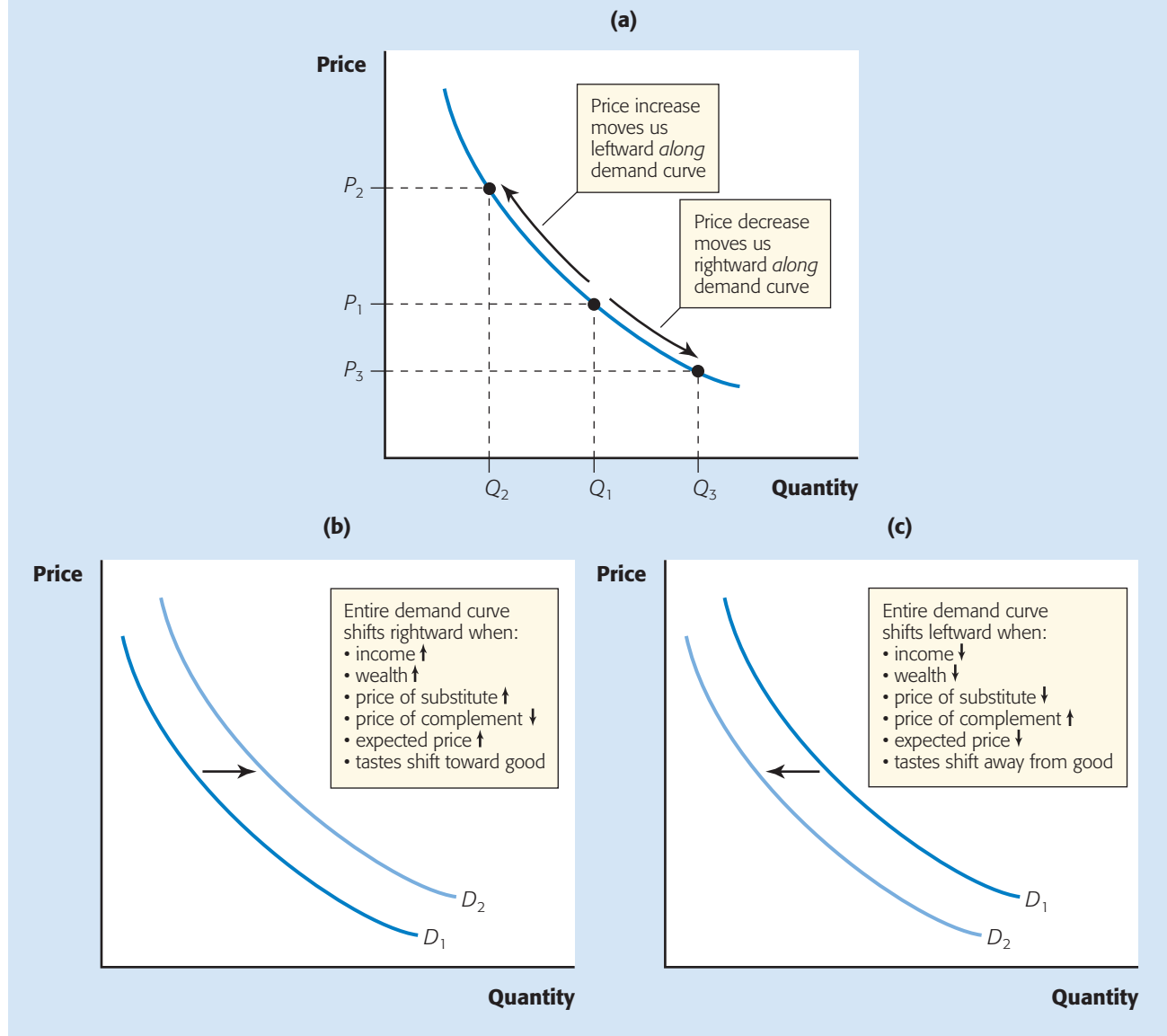


A troubling thought may have occurred to you. Among the variables that shift the demand curve in Figure 3, shouldn't we include the amount supplied by sellers? Or to put the question another way, doesn't supply influence demand?

The answer is no—at least, not directly. The demand curve tells us how much buyers *would choose* to buy at different prices. It provides answers to a series of hypothetical questions: How much maple syrup *would* consumers choose to buy if the price were \$3.00 per bottle? If the price were \$3.50 per bottle? and so on. Sellers' decisions have no effect on the demand curve, since they do not affect the answers to these hypothetical questions.

FIGURE 3

CHANGES IN DEMAND AND IN QUANTITY DEMANDED



capital, fuel, transportation, glass bottles, etc.). The sap can be collected with buckets, bags, plastic tubing, or some combination of these. Syrup evaporators can be fueled with wood, oil, or natural gas, and they can include accessories such as preheaters, reverse osmosis, steam hoods, automatic draw-offs, and more. The syrup can be packaged in glass bottles or plastic bottles or metal tins, and it can be shipped across the country by train, truck, or aircraft. As you can see, there are hundreds if not thousands of different ways to combine inputs to produce a given quantity of maple syrup. Each of these production methods is a part of the known technology of this industry.

A firm's production technology tells us not only what the firm *can* do, it also tells us what it *cannot* do. For example, a firm cannot produce a thousand gallons of maple

syrup per year with only 10 trees, no matter how much labor or equipment it uses, and it cannot produce *any* maple syrup at all using iron ore instead of maple trees.

The known technology in an industry is an important constraint on the firm. Another constraint is that it must *pay a price* for its inputs. Together, the technology of production and the prices of its inputs determine how much it will *cost* the firm to produce different quantities of output.

Finally, every competitive firm faces one more constraint: the market price. The firm is not free to set any price it wants for its output. Rather, it must accept the market price as a given.

In sum,

when a competitive firm comes to a market as a seller, it wants to make the highest possible profit. The firm can choose the level of output it wants to produce, but it faces three constraints: (1) its production technology, (2) the prices it must pay for its inputs, and (3) the market price of its output.

Together, the firm's goal of earning the highest possible profit, and the constraints that it faces, determine the quantity that it will supply in the market.

More specifically,

a firm's quantity supplied of any good is the amount it would choose to produce and sell at a particular price.

And when we turn to the market as a whole:

The market quantity supplied of any good is the amount that all firms in the market would like to produce and sell at a particular price, given the prices they must pay for their inputs, and given any other influences on their selling decisions.

Notice that quantity supplied—like quantity demanded—tells us about sellers' *choices*. The amount that will *actually* be sold will be discussed later, when we put demand and supply together.

THE LAW OF SUPPLY

How does a change in price affect quantity supplied? When a seller can get a higher price for a good, producing and selling it become more profitable. Producers will devote more resources toward its production—perhaps even pulling resources out of other types of production—and increase the quantity of the good they would like to sell. For example, a rise in the price of laptop computers will encourage computer makers to shift resources out of the production of other things (such as desktop computers) and toward the production of laptops.

In general, price and quantity supplied are *positively related*: When the price of a good rises, the quantity supplied will rise as well. This relationship between price and quantity supplied is called the law of supply, the counterpart to the law of demand we discussed earlier.

The law of supply states that when the price of a good rises, and everything else remains the same, the quantity of the good supplied will rise.

Once again, notice the very important words “everything else remains the same.” Although many other variables influence the quantity of a good supplied, the law

Firm's quantity supplied The total amount of a good or service that an individual firm would choose to produce and sell at a given price.

Market quantity supplied The total amount of a good or service that all producers in a market would choose to produce and sell at a given price.

Law of supply As the price of a good increases, the quantity supplied increases.

of supply tells us what would happen if all of them remained unchanged as the price of the good changed.

THE SUPPLY SCHEDULE AND THE SUPPLY CURVE

Let's continue with our example of the market for maple syrup in Wichita. Who are the suppliers in this market? Since maple syrup is easy to transport, any producer on the continent can sell in Wichita. In practice, these producers are located mostly in the forests of Vermont, upstate New York, and Canada. The market quantity supplied is the amount of maple syrup all of these producers together would offer for sale in Wichita at each price for maple syrup.

Supply schedule A list showing the quantities of a good or service that firms would choose to produce and sell at different prices, with all other variables held constant.

Table 3 shows the **supply schedule** for maple syrup in Wichita—a *list of different quantities supplied at different prices, with all other variables held constant*. As you can see, the supply schedule obeys the law of supply: As the price of maple syrup in Wichita rises, the quantity supplied rises along with it. But how can this be? After all, maple trees must be about 40 years old before they can be tapped for syrup, so any rise in quantity supplied now or in the near future cannot come from an increase in planting. What, then, causes quantity supplied to rise as price rises?

Many things. First, with higher prices, firms will find it profitable to tap existing trees more intensively. Second, evaporating and bottling can be done more carefully, so that less maple syrup is spilled and more is available for shipping. Finally, the product can be diverted from other areas and shipped to Wichita instead. For example, if the price of maple syrup rises in Wichita but not in Kansas City, producers would shift deliveries away from Kansas City and toward Wichita.

Now look at Figure 4, which shows a very important curve—the counterpart to the demand curve we drew earlier. In Figure 4, each point represents a price-quantity pair taken from Table 3. For example, point *F* in the figure corresponds to a price of \$2.00 per bottle and a quantity of 4,000 bottles per month, while point *G* represents the price-quantity pair \$4.00 and 6,000 bottles. Connecting all of these points with a solid line gives us the *supply curve* for maple syrup, labeled with an *S* in the figure.

Supply curve A graphical depiction of a supply schedule; a curve showing the quantity of a good or service supplied at various prices, with all other variables held constant.

The supply curve shows the relationship between the price of a good and the quantity supplied, holding constant the values of all other variables that affect supply. Each point on the curve shows the quantity that sellers would choose to sell at a specific price.

Notice that the supply curve in Figure 4—like all supply curves for goods and services—is *upward sloping*. This is the graphical representation of the law of supply.

TABLE 3

SUPPLY SCHEDULE FOR MAPLE SYRUP IN WICHITA

Price (per Bottle)	Quantity Supplied (Bottles per Month)
\$1.00	2,500
2.00	4,000
3.00	5,000
4.00	6,000
5.00	6,500

THE SUPPLY CURVE

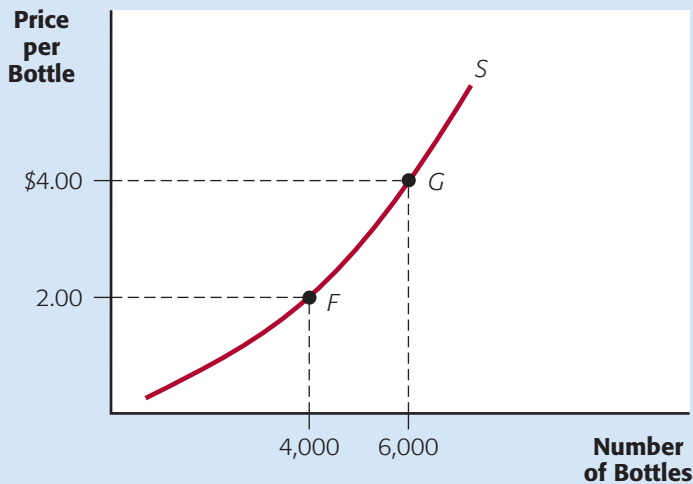


FIGURE 4

The upward-sloping supply curve, S , shows the quantity of a good that firms wish to produce and sell at each price, assuming constant all other variables affecting supply. At \$2.00 per bottle, quantity supplied is 4,000 bottles (point F). At \$4.00 per bottle, quantity supplied is 6,000 bottles (point G).

The law of supply tells us that supply curves slope upward.

CHANGES IN QUANTITY SUPPLIED

Sellers' choices about how much to sell are affected by many different variables. One of these variables—the price of the good—causes sellers to *move along* a given supply curve. The other variables cause the entire supply curve to *shift*. Economists use the same language convention for supply that we discussed earlier for demand. Look once again at Figure 4. Notice that when the price of maple syrup rises from \$2.00 to \$4.00, the number of bottles supplied rises from 4,000 to 6,000. This is a movement *along* the supply curve, from point F to point G , and we call it an *increase in quantity supplied*.

More generally,

a change in a good's price causes us to move along the supply curve. We call this a change in quantity supplied. A rise in price causes a rightward movement along the supply curve—an increase in quantity supplied. A fall in price causes a leftward movement along the supply curve—a decrease in quantity supplied.

Change in quantity supplied A movement along a supply curve in response to a change in price.

CHANGES IN SUPPLY

Both the supply schedule in Table 3 and the supply curve in Figure 4 assume given values for all other variables that might affect supply. For example, the supply curve in Figure 4 might tell us the quantity supplied at each price, *assuming* that maple syrup workers are paid \$10 per hour. But what would happen if these workers' wages fell to \$7 per hour? Then, at any given price for maple syrup, firms would find it more profitable to produce and sell maple syrup, and they would no doubt choose to sell more. This is illustrated in Table 4. For example, at the original wage of \$10, maple syrup producers would choose to sell 6,000 bottles when the price is

TABLE 4

INCREASE IN SUPPLY OF
MAPLE SYRUP IN WICHITA

Price (per Bottle)	Quantity Supplied (Bottles/Month)	Quantity Supplied After Increase in Supply
\$1.00	2,500	4,500
2.00	4,000	6,000
3.00	5,000	7,000
4.00	6,000	8,000
5.00	6,500	8,500

\$4.00. But if they could pay the lower wage of \$7, they would choose to sell 8,000 bottles at that same price of \$4.00 per bottle. The same holds for any other price for maple syrup: After the wage falls, sellers would choose to sell more than before. In other words, *the entire relationship between price and quantity supplied has changed.*

Figure 5 plots the new supply curve from the quantities in the third column of Table 4. The new supply curve lies to the *right* of the old curve. For example, at a price of \$4.00, the old supply curve told us that quantity supplied was 6,000 bottles (point G). But after the decrease in the wage, sellers would choose to supply 8,000 bottles at \$4.00 each (point J). The decrease in maple syrup workers' wages has *shifted the supply curve to the right*. We call this an *increase in supply*.

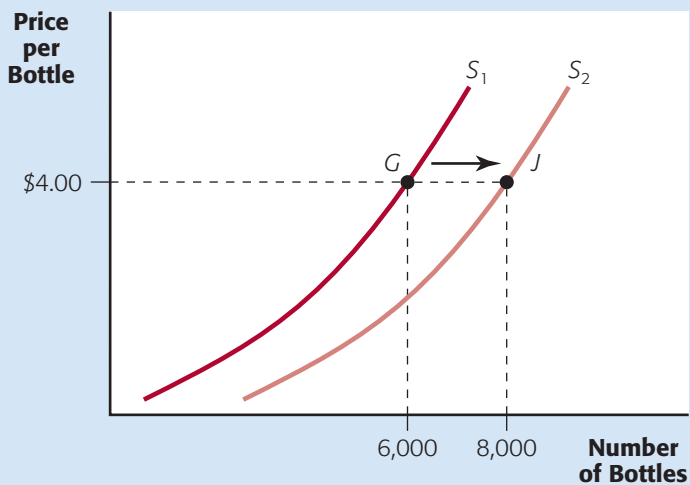
Change in supply A shift of a supply curve in response to some variable other than price.

*A change in any influence on supply—except for the good's price—causes the supply curve to shift. We call this a **change in supply**. When sellers choose to sell more at any price, the supply curve shifts rightward—an increase in supply. When sellers choose to sell less at any price, the supply curve shifts leftward—a decrease in supply.*

FIGURE 5

A SHIFT OF THE SUPPLY CURVE

A change in any nonprice determinant of supply causes the entire supply curve to shift. A decrease in labor costs, for example, causes the supply of maple syrup to shift from S_1 to S_2 . At each price, more bottles are supplied after the shift.



Now let's take a look at the different variables that can cause a change in supply and shift the supply curve.



To avoid confusion, always apply the same language convention for supply that we discussed earlier for demand. When we *move along* the supply curve, we call it a *change in quantity supplied*. A change in quantity supplied is always caused by a change in the good's price. When the entire supply curve shifts, we call it a *change in supply*. A change in supply is caused by a change in something *other* than the good's price.

Prices of Inputs. Producers of maple syrup use a variety of inputs: land, maple trees, evaporators, sap pans, labor, glass bottles, bottling machinery, transportation, and more. A higher price for any of these means a higher cost of producing and selling maple syrup, making it less profitable. As a result, we would expect producers to shift some resources out of maple syrup production, causing a decrease in supply.

In general,

a rise in the price of an input causes a decrease in supply, shifting the supply curve to the left. A fall in the price of an input causes an increase in supply, shifting the supply curve to the right.

Figure 5 has already illustrated one example of this: The supply curve shifted rightward when the wage rate paid to maple syrup workers fell. Now we can see that maple syrup workers are just *one* type of input among many for syrup producers. If the price of bottles, transportation, or any other input were to decrease, it would also shift the supply curve for maple syrup rightward, just as in Figure 5.

Profitability of Alternate Goods. Many firms can switch their production rather easily among several different goods or services, all of which require more or less the same inputs. For example, a dermatology practice can rather easily switch its specialty from acne treatments for the young to wrinkle treatments for the elderly. An automobile producer can—without too much adjustment—switch to producing light trucks. And a maple syrup producer could dry its maple syrup and produce maple *sugar* instead. Or it could even cut down its maple trees and sell maple wood as lumber. These other goods that firms *could* produce are called **alternate goods**.

When an alternate good becomes more profitable to produce—because its price rises, or the cost of producing it falls—the supply curve for the good in question will shift leftward.

In our example, if the price of maple *sugar* rises, and nothing else changes, maple sugar will become more profitable. Producers will devote more of their output to maple sugar, *decreasing* the supply of maple syrup.

Technology. A *technological advance* in production occurs whenever a firm can produce a given level of output in a new and cheaper way than before. For example, the discovery of a surgical procedure called Lasik—in which a laser is used to reshape the interior of the cornea rather than the outer surface—has enabled eye surgeons to correct their patients' vision with fewer follow-up visits and smaller quantities of medication. Similarly, in the late 1990s, several firms—including Ebay, Amazon.com, and Priceline.com—developed new software that enabled people and firms to trade used goods more cheaply over the Internet (compared to

Alternate goods Other goods that a firm could produce, using some of the same types of inputs as the good in question.



The list of variables that shift the supply curve in Figure 6 does not include the amount that buyers want to buy. Is this a mistake? Doesn't demand affect supply?

The answer is no—at least, not directly. The supply curve tells us how much sellers *would choose* to sell at alternative prices. It provides answers to a series of hypothetical questions, such as How much maple syrup would firms choose to sell if the price were \$4.00 per bottle? If the price were \$3.50 per bottle? and so on. Buyers' decisions don't affect the answers to these questions, so they cannot shift the supply curve.

the previous method of running and searching through classified ads). These examples are technological advances because they enable firms to produce the same output (eye surgeries, used goods sales) more cheaply than before.

In maple syrup production, a technological advance might be a new, more efficient tap that draws more maple

syrup from each tree, or a new bottling method that reduces spillage. Advances like this would reduce the cost of producing maple syrup, and producers would want to make and sell more of it at any price.

In general,

cost-saving technological advances increase the supply of a good, shifting the supply curve to the right.

Productive Capacity. A market's productive capacity is determined by the number of producers in the market, and the plant and equipment possessed by each firm. Whenever productive capacity increases, the supply curve shifts rightward, since sellers would choose to sell a greater total quantity at each price. Similarly, a decrease in productive capacity will shift the supply curve leftward. For example, if a sudden blight destroyed maple trees in Vermont, the total productive capacity of maple syrup suppliers would shrink, decreasing the supply of maple syrup to any market. On the other hand, if—over time—more firms moved into the market and started their own maple syrup farms, supply would increase.

Changes in weather can cause sudden changes in productive capacity in many agricultural markets. Good weather increases the productive capacity of all farms in a region, shifting supply curves for their crops to the right. Bad weather destroys crops and decreases productive capacity, shifting supply curves to the left. Natural disasters such as fires, hurricanes, and earthquakes can destroy the productive capacity of *all* industries in a region, thereby causing sudden, dramatic leftward shifts in supply curves.

An increase in sellers' productive capacity—caused by, say, good weather or an increase in the number of firms—shifts the supply curve rightward. A decrease in sellers' productive capacity shifts the supply curve leftward.

Expectations of Future Prices. Imagine that you are the president of Sticky's Maple Syrup, Inc., and your research staff has just determined that the price of maple syrup will soon rise dramatically. What would you do? You should *postpone* producing—or at least selling—your output until later, when the price will be higher and profits will be greater. Applying this logic more generally,

A rise in the expected price of a good will decrease supply, shifting the supply curve leftward.

CHANGES IN SUPPLY AND IN QUANTITY SUPPLIED

FIGURE 6

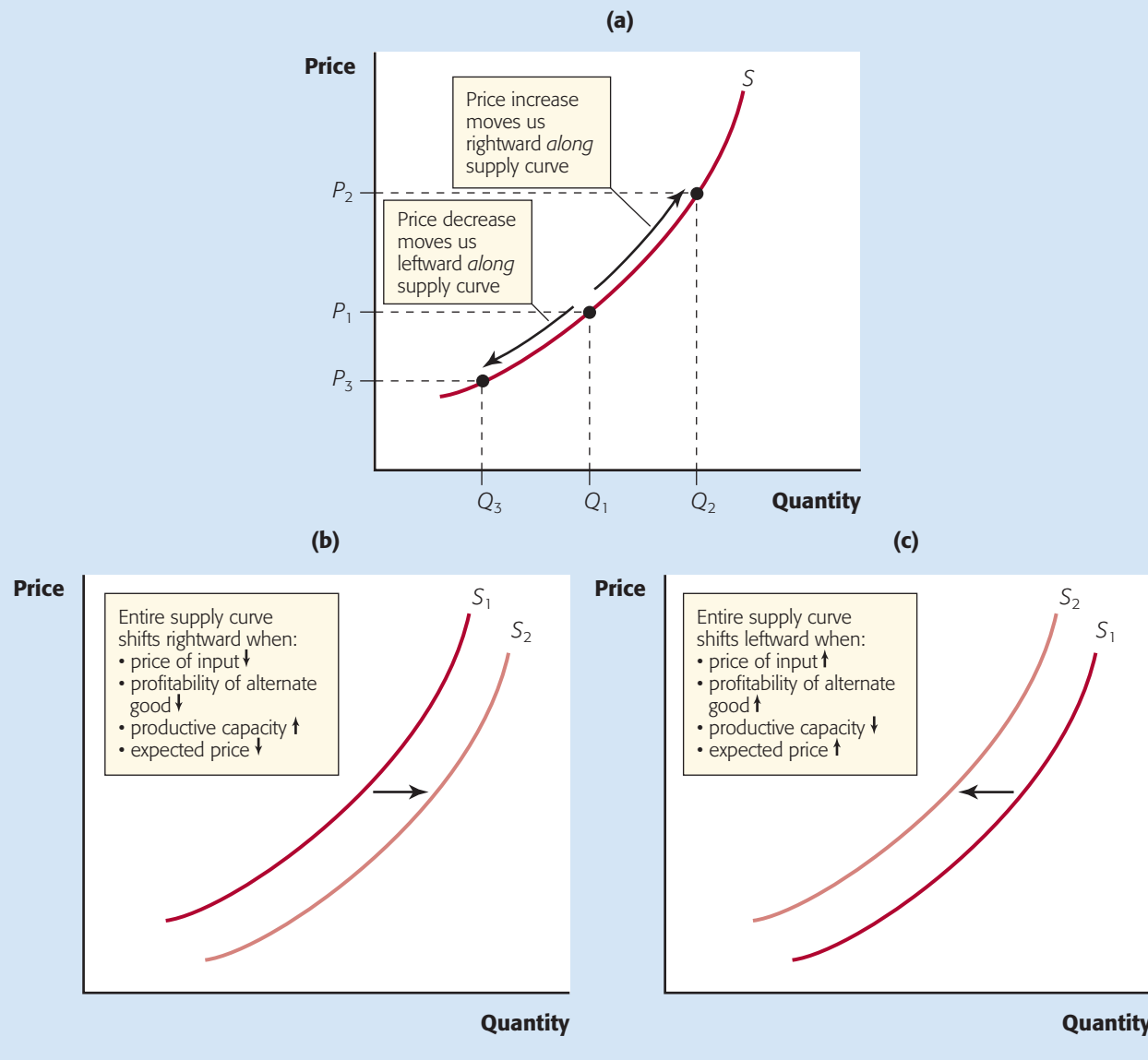


Figure 6 summarizes the different variables that change the supply of a good and shift the supply curve.

PUTTING SUPPLY AND DEMAND TOGETHER

What happens when buyers and sellers, each having the desire and the ability to trade, come together in a market? The two sides of the market certainly have different agendas. Buyers would like to pay the lowest possible price, while sellers would like to charge the highest possible price. Is there chaos when they meet, with



<http://>

Try your hand at a Java-based supply and demand simulation. You can find it at <http://www.openteach.com/javaapplets/econ.html>.

Equilibrium A state of rest; a situation that, once achieved, will not change unless some external factor, previously held constant, changes.

Excess demand At a given price, the excess of quantity demanded over quantity supplied.

buyers and sellers endlessly chasing after each other or endlessly bargaining for advantage, so that trade never takes place? A casual look at the real world suggests not. In most markets, most of the time, there is order and stability in the encounters between buyers and sellers. In most cases, prices do not fluctuate wildly from moment to moment, but seem to hover around a stable value. This stability may be short lived—lasting only a day, an hour, or even a minute in some markets—but still, for this short time, the market seems to be at rest. Whenever we study a market, therefore, we look for this state of rest—a price and quantity at which the market will settle, at least for a while.

Economists use the word *equilibrium* when referring to a state of rest. More formally,

an equilibrium is a situation that, once achieved, will not change unless there is a change in something we have been assuming constant.

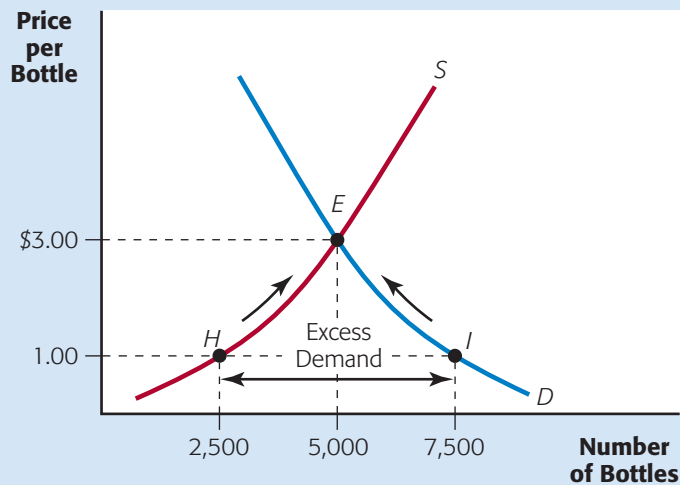
What will be the price of maple syrup in Wichita? And how much will people actually buy each month? We can rephrase these questions as follows: What is the *equilibrium* price of maple syrup in Wichita, and what is the *equilibrium* quantity of maple syrup that will be bought and sold? These are precisely the questions that the supply-and-demand model is designed to answer.

Look at Figure 7, which combines the supply and demand curves for maple syrup in Wichita. We'll use Figure 7 to find the equilibrium in this market through the process of elimination. Let's first ask what would happen if the price of maple syrup in Wichita were \$1.00 per bottle. At this price, we see that buyers would choose to buy 7,500 bottles each week, while sellers would offer to sell only 2,500 per week. There is an **excess demand** of 5,000 bottles. What will happen? Buyers will compete with each other to get more maple syrup than is available, offering to pay a higher price rather than do without. The price will then rise. You can see that \$1.00 per bottle is *not* the equilibrium price, since—if the price *were* \$1.00—it would automatically tend to rise.

FIGURE 7

MARKET EQUILIBRIUM

The intersection of the supply and demand curves at point *E* determines the market price of maple syrup (\$3.00 per bottle) and the number of bottles exchanged (5,000). At a lower price, such as \$1.00 per bottle, buyers would like to purchase more bottles (7,500) than producers are willing to supply (2,500). The resulting excess demand of 5,000 bottles causes the price to rise.



Before we consider other possible prices, let's look more closely at the changes we would see in this market as the price rose. First, there would be a decrease in quantity demanded—a movement along the demand curve leftward from point *I*. At the same time, we would see an increase in quantity supplied—a movement along the supply curve rightward from point *H*. As these movements continued, the excess demand for maple syrup would shrink and, finally—at a price of \$3.00—disappear entirely. At this price, there would be no reason for any further price change, since quantity supplied and quantity demanded would both equal 5,000 bottles per month. There would be no disappointed buyers to offer higher prices. In sum, if the price happens to be below \$3.00, it will rise to \$3.00 and then stay put.

Now let's see what would happen if, for some reason, the price of maple syrup were \$5.00 per bottle. Figure 8 shows us that, at this price, quantity supplied would be 6,500 bottles per month, while quantity demanded would be only 3,500 bottles—an **excess supply** of 3,000 bottles. Sellers would compete with each other to sell more maple syrup than buyers wanted to buy, and the price would fall. Thus, \$5.00 cannot be the equilibrium price.

Moreover, the decrease in price would move us along both the supply curve (leftward) and the demand curve (rightward). As these movements continued, the excess supply of maple syrup would shrink until it disappeared, once again, at a price of \$3.00 per bottle. Our conclusion: If the price happens to be above \$3.00, it will fall to \$3.00 and then stop changing.

You can see that any price higher or lower than \$3.00 is *not* the equilibrium price. If the price is higher than \$3.00, it will tend to drop, and if it is lower, it will tend to rise. You can also see—in Figures 7 and 8—that if the price were exactly \$3.00, there would be neither an excess supply nor an excess demand. Sellers would choose to sell 5,000 bottles per week, and this is exactly the quantity buyers would choose to buy. There would be no reason for the price to change. Thus, \$3.00 must be our sought-after equilibrium price and 5,000 our equilibrium quantity.

Excess supply At a given price, the excess of quantity supplied over quantity demanded.

EXCESS SUPPLY AND PRICE ADJUSTMENT

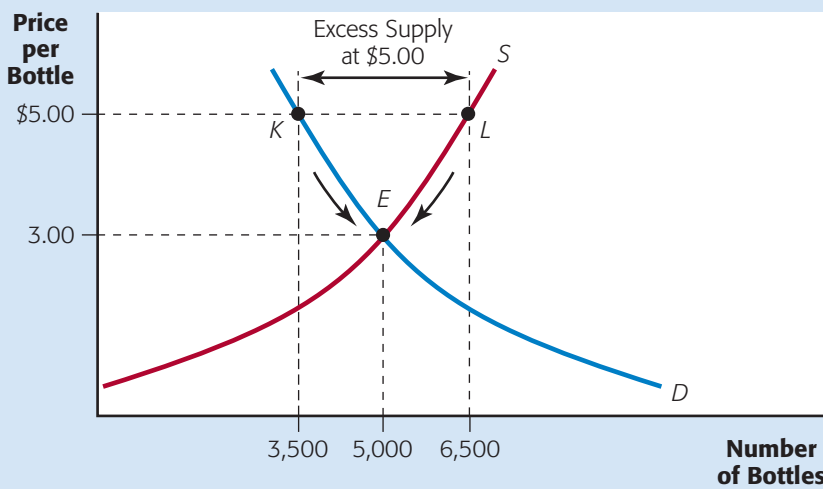


FIGURE 8

At any price above \$3.00 per bottle, the market for maple syrup in Wichita will be out of equilibrium. The excess supply of 3,000 bottles at a price of \$5.00 causes the market price to fall. As the price falls, quantity supplied decreases and quantity demanded increases. At point *E*, the market is back in equilibrium.

No doubt, you have noticed that \$3.00 happens to be the price at which the supply and demand curves cross. This leads us to an easy, graphical technique for locating our equilibrium:

To find the equilibrium price and quantity in a competitive market, draw the supply and demand curves. The equilibrium is the point where the two curves intersect.

The intersection of the supply and demand curves helps us to understand the concept of equilibrium even more clearly. At the intersection, the market is operating on *both* the demand and the supply curves. When the price is \$3.00, buyers and sellers can *actually* buy and sell the quantities they would *choose* to buy and sell at \$3.00. There are no dissatisfied buyers unable to find the goods they want to purchase, nor are there unhappy sellers, unable to find buyers for the products they have brought to the market. This is why \$3.00 is the equilibrium price. In this state of rest, there is a balance between the quantity supplied and the quantity demanded.

But that point of rest will not necessarily be a lasting one, as you are about to see.

WHAT HAPPENS WHEN THINGS CHANGE?

Remember that in order to draw the supply and demand curves in the first place, we had to assume particular values for all the other variables—besides price—that affect demand and supply. If any one of these variables changes, then either the supply curve or the demand curve will shift, and our equilibrium will change as well. Economists are very interested in how and why an equilibrium changes in a market. Let's look at some examples.

AN ICE STORM HITS THE NORTHEAST: A DECREASE IN SUPPLY

In January 1998, New England and Quebec were struck by a severe ice storm. Hundreds of thousands of maple trees were downed, and many more were damaged. In Vermont alone, 10% of the maple trees were destroyed. How did this affect the market for maple syrup in faraway Wichita?

Maple trees are part of the productive capacity of a maple syrup firm, just as factory buildings are part of the productive capacity of a toy manufacturer. And as you learned in this chapter (see Figure 6), a decrease in productive capacity causes a leftward shift of the supply curve in any market in which maple syrup is sold—

including the local market in Wichita.

Figure 9 shows how the ice storm affected this market. Initially, the supply curve for maple syrup in Wichita was S_1 , with the market in equilibrium at Point E . After the ice storm, and the resulting decrease in productive capacity, the supply curve shifted left-



It's tempting to use *upward* and *rightward* interchangeably when describing an increase in demand or supply and to use *downward* and *leftward* when describing a decrease in demand or supply. But be careful! While this interchangeable language works for the demand curve, it does *not* work for the supply curve. To prove this to yourself, look at Figure 6. There you can see that a rightward shift of the supply curve (an increase in supply) is also a *downward* shift of the curve. In later chapters, it will sometimes make sense to describe shifts as upward or downward. For now, it's best to avoid these terms, and stick with *rightward* and *leftward*.

A SHIFT OF SUPPLY AND A NEW EQUILIBRIUM

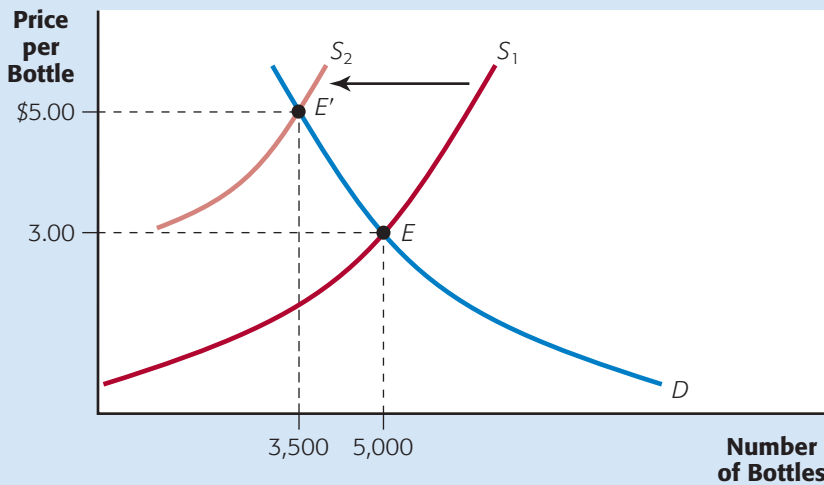


FIGURE 9

An ice storm causes supply to decrease from S_1 to S_2 . At the old equilibrium price of \$3.00, there is now an excess demand. As a result, the price increases until excess demand is eliminated at point E' . In the new equilibrium, quantity demanded again equals quantity supplied. The price is higher, and fewer bottles are produced and sold.

ward—say, to S_2 . The result: a rise in the equilibrium price of maple syrup (from \$3.00 to \$5.00 in Figure 9) and a fall in the equilibrium quantity (from 5,000 to 3,500 bottles).

In this case, it was an ice storm that shifted the supply curve leftward. But suppose, instead, that the wages of maple syrup workers had increased or that evaporators became more expensive or that some maple syrup producers went out of business and sold their farms to housing developers. Any of these changes would have caused the supply curve for maple syrup to shift leftward, increased the equilibrium price and decreased the equilibrium quantity.

More generally,

any change that shifts the supply curve leftward in a market will increase the equilibrium price and decrease the equilibrium quantity in that market.

INTERNET ENTREPRENEURS GET RICH: AN INCREASE IN DEMAND

Since shifts in supply and demand work the same way in *any* market, let's leave Maple syrup for now and look at a different market: housing in San Francisco. In this market, something remarkable has happened recently: The average price of a single-family home¹ increased from \$250,450 in mid-1995 to \$373,750 in mid-1999. In just three and one-half year, the price almost doubled! What explains this dramatic rise in San Francisco housing prices? Supply and demand can give us the answer.

First, let's define the market itself. The sellers are households and real estate companies who own homes in San Francisco. Figure 10 shows their supply curve for



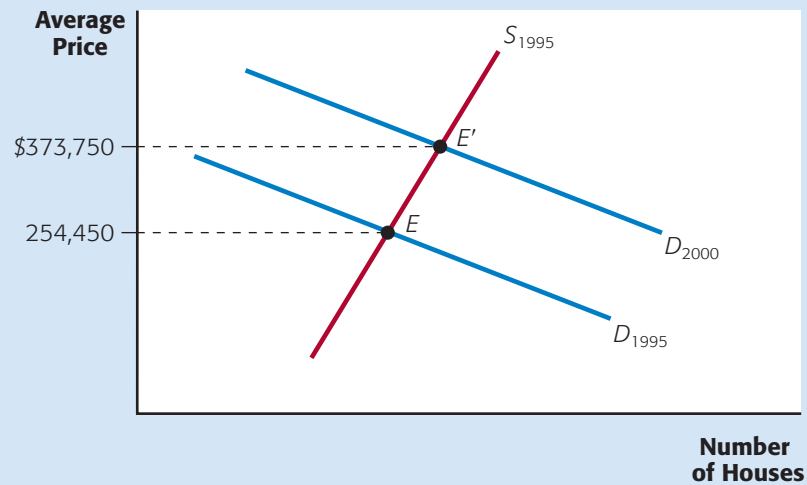
In the late 1990s, an increase in wealth drove up housing prices in San Francisco.

¹ The housing price data is for already-existing, detached homes only. It does not include the price of condominiums or apartments or of newly constructed homes.

FIGURE 10

An increase in household incomes increased demand from D_{1995} to D_{2000} . At the old price of \$254,450, there was an excess demand. As a result, prices rose until excess demand was eliminated at point E' . In the new equilibrium, quantity demanded again equals quantity supplied. The price is higher, and more houses are sold.

A SHIFT OF DEMAND AND A NEW EQUILIBRIUM



housing in 1995, labeled S_{1995} . Notice that this curve sloped upward: A rise in housing prices—with no other change—increases the number of homes offered for sale.²

The demand side of the market consists of households who have the potential to *buy* homes in San Francisco. This includes anyone who works in San Francisco itself or within commuting distance of the city, as well as those who can consider moving to the city (getting jobs or retiring there). The market demand curve in 1995 is represented by the curve D_{1995} . Notice that this demand curve slopes downward: With a higher price (and no other change), buyers would want to buy fewer houses in San Francisco. Point E shows the equilibrium in 1995, the intersection between the demand curve and the supply curve, with an average price of \$254,450.

Now, what happened from 1995 to 1999 that so significantly affected this market? The answer is: the Internet. More specifically, the 1990s was an era in which, by starting up successful companies in a new industry, people could become extremely wealthy in a very short period of time. For example, Pierre Omidyar founded the Internet trading community Ebay in 1995, when his girlfriend wanted to trade Pez dispensers online. The auction idea was a big hit, and by 1999 the 31-year-old Omidyar's wealth was estimated at \$7.8 billion.

This story is not unique. About 200 people with ordinary incomes but extraordinary ideas for new Internet-related companies became *billionaires* in the 1990s. And hundreds of thousands more saw their stocks and stock options rise dramatically in value—doubling, tripling, or quadrupling their wealth within just a few years or less. Disproportionately, the newly rich lived and worked in the Silicon Val-

² The supply curve for housing should slope upward even if we ignore new building activity in the city. To understand why, imagine that you own a home in your current town or city, and that housing prices are rising there. Is there a critical price beyond which you would decide to move elsewhere and cash in on the value of your home? For most people, the answer is yes. After all, even when you own a home, the opportunity cost of continuing to live in an area is the money you *could* have if you sold it and lived elsewhere. As the price of housing rises higher and higher, each additional person who decides to sell his or her home adds to the supply of housing, as shown by the supply curve.

ley area of Northern California—an area within commuting distance of San Francisco. Thus, they were part of the buying side in the housing market there.

As you've learned (see Figure 3), an increase in buyers' wealth causes the demand curve for a normal good—such as housing—to shift rightward. In this case, greater wealth leads people to choose bigger homes and sometimes multiple homes—an increase in the demand for housing. In Figure 10, this is shown as the rightward shift from D_{1995} to D_{1999} , with the equilibrium moving from point E to point E' . And this explains why the price rose from \$254,450 to \$373,750.

More generally,

any change that shifts the demand curve rightward in a market will increase both the equilibrium price and the equilibrium quantity in that market.

Notice that the supply curve has not shifted in Figure 10; it remains at S_{1995} . There *has* been an increase in the quantity of housing supplied (a movement *along* the supply curve), but no change in the *supply* of housing (no shift of the entire curve). Why hasn't the supply curve shifted in Figure 10? Largely because we are dealing with a five-year period—a period too short for significant new construction to change the stock of available housing in a city. While the quantity of housing supplied has increased, it has done so largely because higher prices cause a more intensive use of the *existing* housing stock. This is represented as a movement along the supply curve.

THE MARKET FOR DAY CARE: CHANGES IN BOTH SUPPLY AND DEMAND

So far, we've considered the consequences of a change in a single variable only. But what happens to the market equilibrium when two or more variables change simultaneously? Figure 11 illustrates how we would analyze such a situation, using the market for day care services.

SIMULTANEOUS SHIFTS OF SUPPLY AND DEMAND

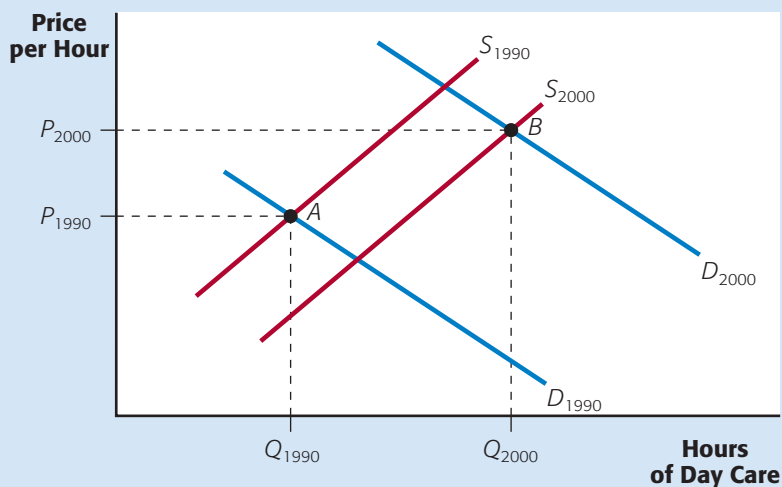


FIGURE 11

As more young mothers sought day care, the demand curve shifted right from D_{1990} to D_{2000} . Simultaneously, more firms entered the market, increasing supply from S_{1990} to S_{2000} . As a result, equilibrium moved from point A to point B . Over the decade, both the quantity of day care services and the price of day care increased.

The story begins in the 1990s, when a variety of factors combined to increase total employment in the United States. Favorable economic conditions drew more and more workers in the labor force, and an increasingly competitive business climate led individuals to work ever-longer hours. By the end of the 1990s, the average American worker spent more time on the job than his or her counterparts in any other developed country. At the same time, legislative reform took thousands of individuals off the welfare reform and into jobs. The implications of these changes were wide-ranging and profound, but here we are interested in just one of them: the use of for-profit day care services.

As more and more women worked outside the home and worked longer hours, they sought care for their preschool children. Many of these women did not have access to the traditional providers of day care—relatives (especially grandparents), friends, or neighbors. So they began to seek day care services in the commercial sector. In terms of our supply and demand model, these changes led a rightward shift of the demand curve for day care services. In Figure 11, the demand curve shifted rightward, from D_{1990} to D_{2000} .

At the same time, many business firms saw an opportunity in these developments. They realized that the labor market trends just described would continue, and perhaps even become stronger. Aiming to earn a profit, these firms set about obtaining office space, hiring teachers, and marketing themselves as high-quality providers of day care services. Large corporations, governments, and nonprofit agencies also got into the act by offering day care services themselves. The effect was to shift the supply curve of day care services to the right from S_{1990} to S_{2000} .

As you can see in Figure 11, the original market equilibrium—in 1990—was determined where the original demand and supply curves intersected at point A. Over the course of the decade, both curves shifted rightward. Therefore, you should not be surprised that the quantity of day care services increased over the decade. But what about the price? In fact, the demand curve shifted farther during the decade than the supply curve did. As a result, the equilibrium in 2000—shown at point B—featured a larger quantity *and* a higher price. What if it were the other way around so that supply increased by more than demand did? In that case, the total quantity exchanged would still have increased, but the price would have fallen.

Figure 11 illustrates just *one* possible combination of simultaneous shifts in supply and demand. But there are others. Table 5 summarizes what we *know* will happen to the equilibrium price (P) and quantity (Q), and what remains uncertain, in each case. For example, to find what happens when demand increases and

TABLE 5

EFFECT OF SUPPLY AND DEMAND SHIFTS ON EQUILIBRIUM PRICE (P) AND QUANTITY (Q)

	Increase in Demand (Rightward Shift)	No Change in Demand	Decrease in Demand (Leftward Shift)
Increase in Supply (Rightward Shift)	$P? Q\uparrow$	$P\downarrow Q\uparrow$	$P\downarrow Q?$
No Change in Supply	$P\uparrow Q\uparrow$	No change in P or Q	$P\downarrow Q\downarrow$
Decrease in Supply (Leftward Shift)	$P\uparrow Q?$	$P\uparrow Q\downarrow$	$P? Q\downarrow$

supply decreases, look at the bottom, leftmost cell: The equilibrium price rises, while the equilibrium quantity might rise, fall, or remain the same.

Remember the advice in Chapter 1—to study economics actively rather than passively. This would be a good time to put down the book, pick up a pencil and paper, and see whether you can *work* with supply and demand curves, rather than just follow along as you read. Try to draw diagrams that illustrate each of the possibilities in Table 5.

THE FOUR-STEP PROCEDURE

In this chapter, we built a model—a supply and demand model—and then used it to analyze price changes in several markets. You may not have noticed it, but we took four distinct Key Steps as the chapter proceeded. Economists take these same four steps to answer almost *any* question about the economy. Why? Because they are so effective in cutting through the chaos and confusion of the economy and helping us see how things really work.

In this book, we'll focus on this *four-step procedure*, which forms the core of economists' unique methodology. And we'll start right now by listing and discussing all four steps.

Key Step 1—Characterize the Market: *Decide which market or markets best suit the problem being analyzed, and identify the decision makers (buyers and sellers) who interact in that market,*



Characterize the Market

In economics, we make sense of the very complex, real-world economy by viewing it as a collection of *markets*. Each of these markets involves a group of *decision makers*—buyers and sellers—who have the potential to trade with each other. At the very beginning of any economic analysis, we must decide which market or markets to look at and how these markets should be *defined*.

To define a market, we must define (a) the thing being traded (such as maple syrup); (b) the decision makers in the market (such as maple syrup producers in New England and Canada on the selling side and households in Wichita on the buying side); and (c) the nature of competition in the market (such as the perfectly competitive markets we've looked at in this chapter). Keep in mind that whenever we draw market supply and demand curves we are treating the market as perfectly competitive, in which each individual buyer and seller treats the price as a given.

Key Step 2—Identify the Goals and Constraints: *Identify the goals that the decision makers are trying to achieve, and the constraints they face in achieving those goals.*



Identify Goals and Constraints

In every market, we assume that each decision maker is trying to achieve a specific goal. Typically, the goal will involve *maximizing some quantity*. Business firms, for example, are usually assumed to maximize profit. Households maximize utility—their well-being or satisfaction. In some cases, however, we might want to recognize that firms or households are actually groups of individuals with different agendas. While a firm's owners might want the firm to maximize profits, the managers might want to consider their own power, prestige, and job security. These goals may conflict, and the behavior of the firm will depend on how the conflict is resolved.

While economists often have spirited disagreements about *what* is being maximized, there is virtually unanimous agreement that, in any economic model, everyone is maximizing *something*. Even the behavior of groups—like the decision makers in a firm or officials of the federal government—is assumed to arise from the behavior of different maximizing individuals, each pursuing his or her own agenda.

In addition to having goals, decision makers also face constraints. Firms are constrained by their production technology, the prices they must pay for their inputs, and the price they can get for their output. Households are constrained by the prices they must pay for their purchases and by their limited incomes. Government agencies are constrained by the prices of the things they buy and by limited budgets. And even entire nations, as a whole, are constrained in their choices by the resources at their disposal.

Find the Equilibrium



Key Step 3—Find the Equilibrium: *Describe the conditions necessary for equilibrium in the market, and a method for determining that equilibrium.*

Once we've defined a market and the goals and constraints of the decision makers there, we can usually find the point at which the market will come to rest—the *equilibrium*. In the perfectly competitive markets we analyzed in this chapter, in which each decision maker takes the price as a given, the equilibrium price is the one at which quantity demanded and quantity supplied are equal. This equilibrium is easy to find on a graph once you've drawn the supply and demand curves. It's simply the point of intersection between the two curves.

But remember: Not all markets are perfectly competitive—or even close to it. When we analyze *imperfectly* competitive markets, we'll have to find the market equilibrium in a different way, as you'll learn when you study microeconomics.

What Happens When Things Change?



Key Step 4—What Happens When Things Change: *Explore how events or government policies change the market equilibrium.*

Almost every economic analysis ends with an exploration of how an event or policy change affects one or more markets. For example, in this chapter, we explored how an ice storm affected the market for maple syrup, how sudden increases in wealth affected the price of homes in San Francisco, and how both supply and demand changes affected the market for day care services.

Do economists really follow this same procedure to analyze almost *any* economic problem? Indeed they do. They use it to answer important *microeconomic* questions. Why does government intervention in a market to lower the price of a good (such as apartment rents) often backfire and sometimes harm the very people it was designed to help? Why do some people earn salaries that are hundreds of times higher than others? Why are economists virtually always skeptical of anyone who says they can “beat” the stock market, even if they have done so in the past? Later in this text, when we turn our attention to these questions, the four-step procedure will play a central role.

Economists also use the procedure to address important *macroeconomic* questions. What causes recessions, and what can we do to prevent them? Why has the United States experienced such low inflation in recent years, and how long can we

expect our recent good fortune to continue? How will the Internet and other new technologies affect the growth rate of the U.S. economy?

In this book, we'll be taking these four Key Steps again and again, every time we want to understand an aspect of the economy. But from now on, you'll recognize the steps as we develop new models, because we'll be calling them to your attention as we use them.

Some of the chapters that follow will concentrate on just one or a few of the steps, while in others, we'll use the entire four-step procedure. To help you keep track, you'll often see icons in the margins of this book that remind you of which of the four steps is being studied. Whenever you see one of these icons, think about how the corresponding step is being used. If you do this, you will soon find yourself thinking like an economist.

You have already seen one of the payoffs to this approach: It can explain how prices are determined in perfectly competitive markets, or in markets that come close to perfect competition. But the four-step procedure takes us even further. It helps us understand how *all* types of markets operate, whether they are perfectly competitive or not. It helps us predict important changes in the economy and prepare for them. And it helps us design government policies to accomplish our social goals and avoid policies that are likely to backfire.

ANTICIPATING A PRICE CHANGE

In the late 1980s, many East Coast colleges purchased expensive equipment that would enable them to switch rapidly from oil to natural gas as a source of heat. The idea was to protect the colleges from a sudden rise in oil prices, like the one they had suffered in the 1970s.

Finally, an event occurred that gave the colleges a change to put their new equipment to use: In the fall of 1990, Iraq invaded Kuwait. As oil prices skyrocketed, the colleges switched from burning oil to burning natural gas. The college administrators expected big savings on their energy bills. But they were in for a shock. When they received the bills from their local utilities, they found that the price of natural gas—like the price of oil—had risen sharply. As a result, they did not save much at all. Many of these administrators were angry at the utility companies and accused them of price gouging. Iraq's invasion of Kuwait, they reasoned, had not affected natural gas supplies at all, so there was no reason for the price of natural gas to rise.

Were the college administrators right? Was this just an example of price gouging by the utility companies who were taking advantage of an international crisis to increase their profits? A simple supply and demand analysis will give us the answer. More specifically, it will enable us to answer two questions: (1) Why did Iraq's invasion of Kuwait cause the price of oil to rise, and (2) Why did the price of natural gas rise as well?

Figure 12 shows supply-and-demand curves in one of the markets relevant to our analysis: the market for crude oil. In this market, oil producers—including those in Iraq and Kuwait—sell to American buyers. Before the invasion, the market was in equilibrium at E with price P_1 and total output Q_1 .

Then came the event that changed the equilibrium: Iraq's invasion and continued occupation of Kuwait—one of the largest oil producers in the world.

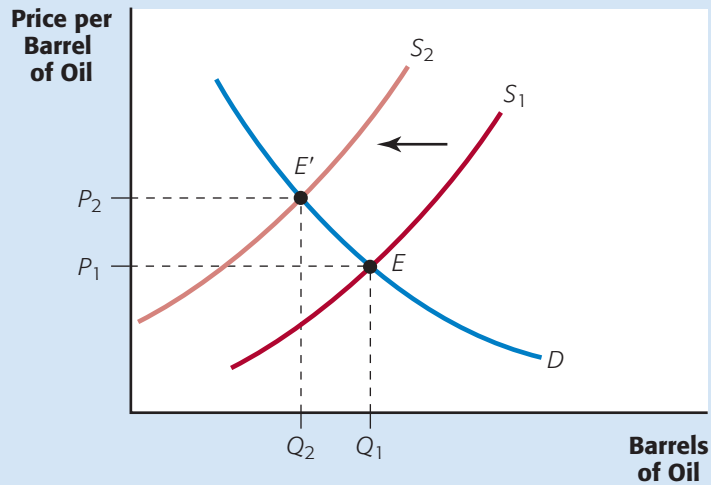
Using the
THEORY



FIGURE 12

Before the Iraqi invasion of Kuwait, the oil market was in equilibrium at point E . The invasion and the resulting embargo on Iraqi oil decreased supply to S_2 . Price increased to P_2 , and the quantity exchanged fell to Q_2 .

THE MARKET FOR OIL



Immediately after the invasion, the United States led a worldwide embargo on oil from both Iraq and Kuwait. As far as the oil market was concerned, it was as if these nations' oil fields no longer existed—a significant decrease in the oil industry's productive capacity. If you look back at Figure 6, you will see that a decrease in productive capacity shifts the supply curve to the left, and this is just what happened. The new equilibrium at E' occurred at a lower quantity and a higher price. This change in the oil market's equilibrium was well understood by most people—including the college administrators—and no one was surprised when oil prices rose.

But what has all this got to do with natural gas prices? Everything, as the next part of our analysis will show.

Figure 13 shows the next market relevant to our analysis: the market for natural gas. In this market, world producers (which did not include Iraq or Kuwait) sell natural gas to American buyers. In this market, the initial equilibrium—before the invasion and before the rise in oil prices—was at point F . How did the invasion affect the equilibrium?

Oil is a *substitute* for natural gas. A rise in the price of a substitute, we know, will increase the demand for a good. (Look back at Figure 3 if you need a reminder.) In this case, the increase in the price of oil caused the demand curve for natural gas to shift rightward. In Figure 13, the price of natural gas rose from P_3 to P_4 .

The administrators were right that the invasion of Kuwait did not affect the supply of natural gas. What they missed, however, was the invasion's effect on the *demand* for natural gas. With a fuller understanding of supply and demand, they could have predicted—*before* investing in their expensive switching equipment—that any rise in oil prices would cause a rise in natural gas prices. Armed with this knowledge, they would have anticipated a much smaller savings in energy costs from switching to natural gas and might have decided that there were better uses for their scarce funds.

THE MARKET FOR NATURAL GAS

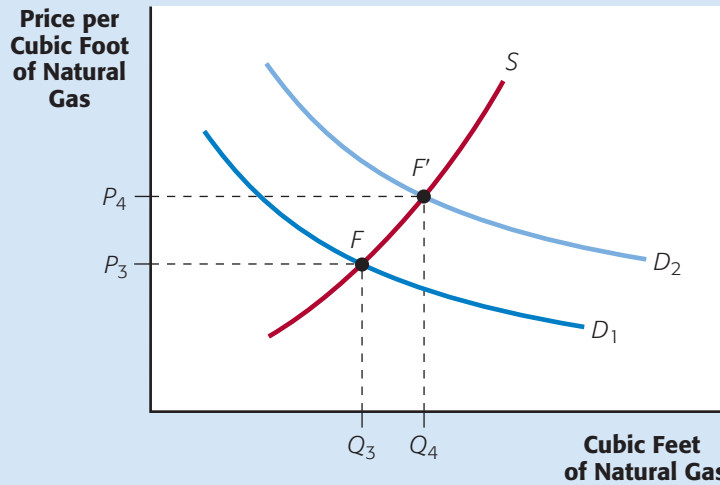


FIGURE 13

Oil is a substitute for natural gas. A rise in the price of oil increases the demand for natural gas. Here, demand for natural gas increases from D_1 to D_2 and the price rises from P_3 to P_4 .

SUMMARY

In a market economy, prices are determined through the interaction of buyers and sellers in *markets*. *Perfectly competitive* markets have many buyers and sellers, and none of them individually can affect the market price. If at least one buyer or seller has the power to influence the price of a product, the market is *imperfectly competitive*.

The model of *supply and demand* explains how prices are determined in perfectly competitive markets. The *quantity demanded* of any good is the total amount buyers would choose to purchase at a given price. The *law of demand* states that quantity demanded is negatively related to price; it tells us

that the *demand curve* slopes downward. The demand curve is drawn for given levels of income, wealth, tastes, and prices of substitute and complementary goods. If any of those factors changes, the demand curve will shift.

The *quantity supplied* of a good is the total amount sellers would choose to produce and sell at a given price. According to the *law of supply*, supply curves slope upward. The supply curve will shift if there is a change in the price of an input, the price of an alternate good, productive capacity, or expectations of future prices.

KEY TERMS

aggregation	market demand curve	complement	alternate goods
imperfectly competitive market	change in quantity demanded	technology	equilibrium
perfectly competitive market	change in demand	firm's quantity supplied	excess demand
individual's quantity demanded	income	market quantity supplied	excess supply
market quantity demanded	wealth	law of supply	
law of demand	normal good	supply schedule	
demand schedule	inferior good	supply curve	
	substitute	change in quantity supplied	
		change in supply	

R E V I E W Q U E S T I O N S

1. How does the way each of the following terms is used in economics differ from the way it is used in everyday language?
 - a. market
 - b. demand
 - c. normal good
 - d. inferior good
 - e. supply
2. What is the difference between *demand* and *quantity demanded*?
3. List and briefly explain the factors that can shift a demand curve and the factors that can shift a supply curve.
4. What is the difference between substitutes and complements? Which of the following pairs of goods are substitutes, which are complements, and which are neither?
 - a. Coke and Pepsi
 - b. Computer hardware and computer software
 - c. Beef and chicken
 - d. Salt and sugar
 - e. Ice cream and frozen yogurt
5. Rank each of the following markets according to how close you think it comes to perfect competition:
 - a. Wheat
 - b. Personal computer hardware
 - c. Gold
 - d. Airline tickets from New York to Kalamazoo, Michigan
6. Is each of the following goods more likely to be *normal* or *inferior*?
 - a. Lexus automobiles
 - b. Secondhand clothes
 - c. Imported beer
 - d. Baby-sitting services
 - e. Recapped tires
 - f. Futons
 - g. Home haircutting tools
 - h. Restaurant meals
7. What does the term *equilibrium* mean in economics?
8. Explain why the price in a free market will not remain above or below equilibrium for long, unless there is outside interference.
9. Determine whether each of the following will cause a change in demand or a change in supply, and in which direction:
 - a. Input prices increase.
 - b. Income in an area declines.
 - c. The price of an alternate good increases.
 - d. Tastes shift away from a good.
10. In the Using the Theory section at the end of this chapter, three of the Key Steps in the four-step procedure are mentioned explicitly, and one step is implicit.
 - a. Identify the three Key Steps explicitly used in the analysis, and briefly describe *where* each is used.
 - b. For the “missing step,” write a sentence or two to be inserted in the analysis that would describe how the step is used.

P R O B L E M S A N D E X E R C I S E S

1. In the late 1990s, beef—which had fallen out of favor in the 1970s and 1980s—became popular again. On a supply and demand diagram, illustrate the effect of such a change on equilibrium price and quantity in the market for beef.
2. Discuss, and illustrate with a graph, how each of the following events will affect the market for coffee:
 - a. A blight on coffee plants kills off much of the Brazilian crop.
 - b. The price of tea declines.
 - c. Coffee workers organize themselves into a union and gain higher wages.
 - d. Coffee is shown to cause cancer in laboratory rats.
 - e. Coffee prices are expected to rise rapidly in the near future.

3. The following table gives hypothetical data for the quantity of gasoline demanded and supplied in Los Angeles per month.

Price per Gallon	Quantity Demanded Millions of Gallons	Quantity Supplied Millions of Gallons
\$1.20	170	80
\$1.30	156	105
\$1.40	140	140
\$1.50	123	175
\$1.60	100	210
\$1.70	95	238

- Graph the demand and supply curves.
 - Find the equilibrium price and quantity.
 - Illustrate on your graph how a rise in the price of automobiles would affect the gasoline market.
4. How would each of the following affect the market for blue jeans in the United States? Illustrate each answer with a supply and demand diagram.
- The price of denim cloth increases.
 - An influx of immigrants arrives in the United States. (Explicitly state any assumptions you are making.)
 - An economic slowdown in the United States causes household incomes to decrease.

- Indicate which curve shifted—and in which direction—for each of the following.
 - The price of furniture rises as the quantity bought and sold falls.
 - Apartment vacancy rates increase while average monthly rent on apartments declines.
 - The price of personal computers continues to decline as sales skyrocket.
- Draw supply and demand diagrams from two different markets, and label the markets *A* and *B*. Then use your diagrams to illustrate the impact of the following events. In each case, determine what happens to price and quantity in each market.
 - A* and *B* are substitutes, and producers expect the price of good *A* to rise in the future.
 - A* and *B* satisfy the same kinds of desires, and there is a shift in tastes away from *A* and toward *B*.
 - A* is a normal good, while *B* is an inferior good. Incomes in the community increase.
 - A* and *B* are complementary goods. There is a technological advance in the production of good *B*.

C H A L L E N G E Q U E S T I O N S

- Suppose that demand is given by the equation $Q_D = 500 - 50P$, where Q_D is quantity demanded, and P is the price of the good. Supply is described by the equation $Q_S = 50 + 25P$, where Q_S is quantity supplied. What is the equilibrium price and quantity?
- A Wall Street analyst observes the following equilibrium price-quantity combinations in the market for restaurant meals in a city over a four-year period:

Year	P (Thousands of Meals per Month)	Q
1	\$12	20
2	\$15	30
3	\$17	40
4	\$20	50

- She concludes that the market defies the law of demand. Is she correct? Why or why not?
- While crime rates have fallen across the country over the past few years, they have fallen especially rapidly in Manhattan. At the same time, there are some neighborhoods in the New York Metropolitan Area in which the crime rate has remained constant. Using supply and demand diagrams for rental housing, explain how a falling crime rate in Manhattan could make the residents in *other* neighborhoods *worse off*. (Hint: As people from around the country move to Manhattan, what happens to rents there? If someone cannot afford to pay higher rent in Manhattan, what might they do?)

EXPERIENTIAL EXERCISES

1. Visit the *Dismal Scientist* Web page at <http://www.dismal.com>, and find an article that you think involves supply and demand considerations. Once you understand the argument, try to present it, using a graph. What is the market being considered, and who are the suppliers and who are the demanders in this market (Key Step #1)? What are the goals of the decision makers on each side of the market, and what are their constraints (Key Step #2)? Is this a market in which the equilibrium is changing (Key Steps #3 and #4)? Explain.



2. You now have a basic understanding of supply and demand. Find a relevant current article using Infotrac or the *Wall Street Journal* and interpret it, using a supply and demand diagram. Explain at least one situation in which a curve shifts. What caused the shift, and how did it affect price and quantity?

WORKING WITH SUPPLY AND DEMAND

In Chapter 3, you learned how supply and demand enable us to explain how prices are determined, and also how and why they change. But the model can do even more than that. It helps us see what happens when governments intervene in markets to influence prices. And it gives us insights about a variety of social policy issues, ranging from the war against illegal drugs to the design of an effective health care system. This chapter is all about *working with* supply and demand, and applying it in the real world.

In much of the chapter, we'll be focusing our attention on Key Step #4 "What Happens When Things Change." Keep in mind, though, that in order to reach Key Step #4 we've implicitly taken the other steps in our 4-step procedure. That is, when we speak about changes in a market, we've already (implicitly) characterized that market (step #1), identified the goals and constraints of the buyers and sellers in that market (step #2), and found the equilibrium there (step #3). Only then can we ask what happens when things change.

GOVERNMENT INTERVENTION IN MARKETS

The forces of supply and demand are important. They determine prices in many markets. And prices, in turn, force decision makers to consider the opportunity cost to *others* of their individual decisions.

So, three cheers for supply and demand! Or better make that *two* cheers. Because while everyone agrees that having prices is necessary for the smooth functioning of our economy, not everyone is happy with the prices that supply and demand give us. Apartment dwellers often complain that their rent is too high, and farmers complain that the price of their crops is too low.

Responding to this dissatisfaction, governments will sometimes intervene to *change* the price in a market. In some cases (*taxes* and *subsidies*), the government will try to change the equilibrium price. In other cases (*price ceilings* and *price floors*), it will try to prevent the market from reaching its equilibrium. What happens when the government intervenes in a market? Let's see.

CHAPTER OUTLINE

Government Intervention in Markets

- Price Ceilings
- Price Floors
- Taxes

Price Elasticity of Demand

- Calculating Price Elasticity of Demand
- Elasticity and Straight-Line Demand Curves
- Categorizing Goods by Elasticity
- Elasticity and Total Expenditure
- Determinants of Elasticity
- Using Price Elasticity of Demand:

Other Demand Elasticities

- Income Elasticity of Demand
- Cross-Price Elasticity of Demand

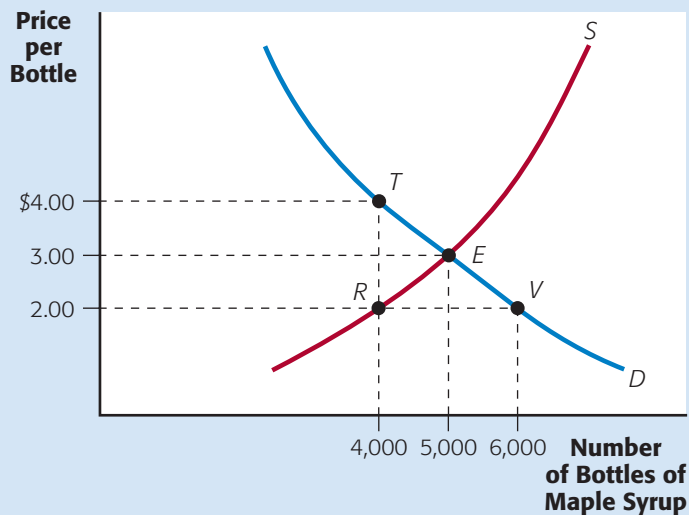
Using the Theory: The Story of Two Markets

- The Market for Food
- Health Insurance and the Market for Health Care

FIGURE 1

A government-imposed price ceiling of \$2.00 per bottle reduces the legal quantity sold to 4,000 bottles, leaving an excess demand of 2,000 bottles. A black market may arise in which scalpers purchase the available 4,000 bottles and sell them (illegally) at the highest price consumers are willing to pay for that quantity—\$4.00 per bottle, determined at point *T* on the demand curve.

A PRICE CEILING IN THE MARKET FOR MAPLE SYRUP



PRICE CEILINGS

Figure 1 shows the market for maple syrup in Wichita, with an equilibrium price of \$3.00 per bottle. Suppose that maple syrup buyers complain to the government that this price is too high. The government responds by imposing a **price ceiling** in this market—a regulation preventing the price from rising above the ceiling.

More specifically, suppose the ceiling is \$2.00 per bottle, and it is strictly enforced. Then producers will no longer be able to charge \$3.00 for maple syrup, but will have to content themselves with \$2.00 instead. In Figure 1, we will move down along the supply curve, from point *E* to point *R*, decreasing quantity supplied from 5,000 bottles to 4,000. At the same time, the decrease in price will move us along the demand curve, from point *E* to point *V*, increasing quantity demanded from 5,000 to 6,000. These changes in quantities supplied and demanded together create an *excess demand* for maple syrup of $6,000 - 4,000 = 2,000$ bottles each month. Ordinarily, the excess demand would force the price back up to \$3.00. But now the price ceiling prevents this from occurring. What will happen?

There is a practical observation about markets that helps us arrive at an answer:

When quantity supplied and quantity demanded differ, the short side of the market—whichever of the two quantities is smaller—will prevail.

This simple rule follows from the voluntary nature of exchange in a market system: No one can be forced to buy or sell more than they want to. With an excess demand, sellers are the short side of the market. Since we cannot force them to sell any more than they want to—4,000 units—buyers will not be able to purchase all they want.

But this is not the end of the story. Because of the excess demand, all 4,000 bottles produced each month will quickly disappear from store shelves, and many buyers will be disappointed. The next time people hear that maple syrup has become available, everyone will try to get there first, and we can expect long lines at stores. In addition, people may have to go from store to store, searching for scarce maple syrup. When we include the *opportunity cost* of the time spent waiting in line or

Price ceiling A government-imposed maximum price in a market.

Short side of the market The smaller of quantity supplied and quantity demanded at a particular price.

shopping around, the ultimate effect of the price ceiling may be a *higher* cost of maple syrup for many consumers.

A price ceiling creates a shortage, and increases the time and trouble required to buy the good. While the price decreases, the opportunity cost may rise.

And there is still more. While the government may be able to prevent maple syrup *producers* from selling above the price ceiling, it may not be able to prevent enterprising individuals from buying maple syrup at the official ceiling price and then reselling it to desperate buyers for a profit. The result is a **black market**, where goods are sold illegally at prices higher than the legal ceiling.

Ironically, the black-market price will typically exceed the original, freely determined equilibrium price—\$3.00 per bottle in our example. To see why, look again at Figure 1. With a price ceiling of \$2.00, sellers supply 4,000 bottles per month. Suppose all of this is bought by people—maple syrup scalpers, if you will—who then sell it at the highest price they can get. What price can they charge? We can use the demand curve to find out. At \$4.00 per bottle (point *T*), the scalpers would just be able to sell all 4,000 bottles. They have no reason, therefore, to charge any less than this.

The unintended consequences of price ceilings—long lines, black markets, and, often, higher prices—explain why they are generally a poor way to bring down prices. Experience with price ceilings has generally confirmed this judgment, so in practice they are rare.

An exception, however, is **rent controls**—city ordinances that specify a maximum monthly rent on many apartments and homes. If you live in a city with rent control, you will be familiar with its consequences. In any case, you may want to reread this section with the market for apartments in mind. How are shortages and long lines manifested? Do rent controls always decrease the cost of apartments to renters? (Think: opportunity cost.) And who are the middlemen—the “apartment scalpers”—who profit in this market?

PRICE FLOORS

Sometimes, governments try to help sellers of a good by establishing a **price floor**—a minimum amount below which the price is not permitted to fall. The most common use of price floors around the world has been to raise prices (or prevent prices from falling) in agricultural markets. Price floors for agricultural goods are commonly called *price support programs*.


In the United States, price support programs began during the Great Depression, after farm prices fell by more than 50% between 1929 and 1932. The Agricultural Adjustment Act of 1933, and an amendment in 1935, gave the president the authority to intervene in markets for a variety of agricultural goods. Over the next 60 years, the United States Department of Agriculture (USDA) put in place programs to maintain high prices for cotton, wheat, rice, corn, tobacco, honey, milk, cheese, butter, and many other farm goods.

Things changed in 1996. In April of that year, Congress passed—and President Clinton signed—the Federal Agriculture Improvement and Reform Act. The new law eliminated many of the government’s price support programs, and dramatically scaled back others. But there were three important exceptions: peanuts, sugar, and dairy products. In these markets, the USDA continues to impose price floors, at least for the time being.

To see how price floors work, let’s look at the market for nonfat dry milk—a market in which the USDA has been supporting prices since 1933. Figure 2 shows

Black market A market in which goods are sold illegally at a price above the legal ceiling.

Rent controls Government-imposed maximum rents on apartments and homes.

 What Happens When Things Change?

Price floor A government-imposed minimum price in a market.

that—before any price floor is imposed—the market is in equilibrium at point *A*. The equilibrium price, 90 cents per pound, corresponds to reasonable estimates of where the price of nonfat dry milk *would be* if there were no government intervention in the market. In the figure, we assume that the equilibrium quantity would be 200 million pounds.

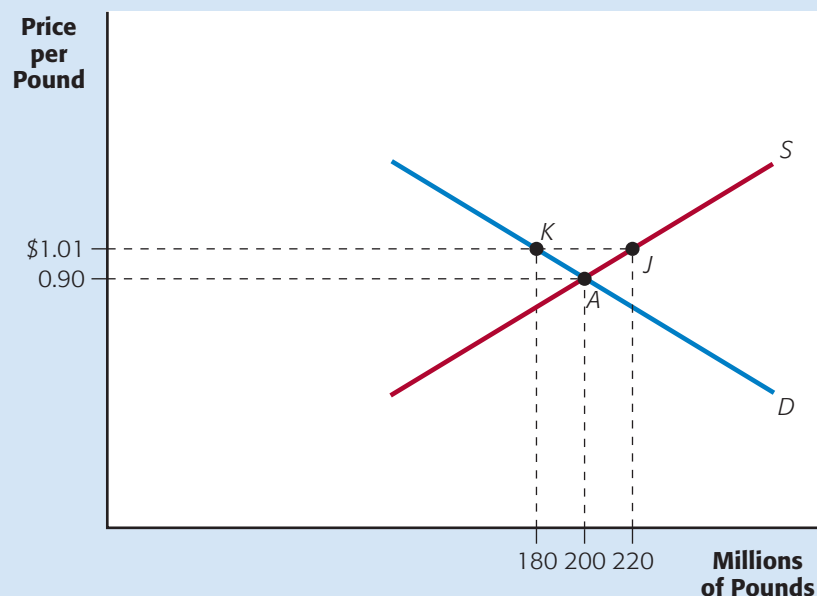
Now let's examine the impact of the current price floor of \$1.01 per pound. At this price, producers want to sell 220 million pounds, while consumers want to purchase only 180 million pounds. There is an excess supply of 220 million – 180 million = 40 million pounds. Our short-side rule tells us that buyers determine the amount actually traded. They purchase 180 million of the 220 million pounds produced, and producers are unable to sell the remainder. The excess supply of 40 million pounds would ordinarily push the market price down to its equilibrium value: \$0.90. What prevents this from happening? Something more than just a government *declaration* of a price floor. After all, if the government merely *declared* that nonfat dry milk must be sold for \$1.01 per pound, producers would have a strong incentive to sell some of their milk for less. Buyers, of course, would be happy to buy at the lower price. How, then, does the government *enforce* its price floor?

With a foolproof strategy. The government simply promises to buy nonfat dry milk from any seller at \$1.01 per pound. With this policy, no supplier would ever sell at any price *below* \$1.01, since it could always sell to the government instead. With the price effectively stuck at \$1.01, private buyers buy 180 million pounds—point *K* on the demand curve in Figure 2. But since quantity supplied is 220 million, at point *J*, the government must buy the excess supply of 40 million pounds each year. In other words, the government maintains the price floor by *buying up* the entire excess supply. This prevents the excess supply from doing what it would ordinarily do: drive the price down to its equilibrium value.

FIGURE 2

A PRICE FLOOR IN THE MARKET FOR NONFAT DRY MILK

If a price floor is established above the market equilibrium price, an excess supply results. Here, the market equilibrium price for nonfat dry milk is \$0.90 per pound, but a floor price was set at \$1.01 per pound. At that higher price, the quantity supplied is 220 million pounds (point *J*), but the quantity demanded is only 180 million pounds (point *K*). Thus, an excess supply of 40 million pounds exists. To maintain its floor price, the government must buy up the entire excess supply at the floor price.



And, indeed, this is what the government has done in markets for many agricultural goods, including nonfat dry milk. Between 1994 and 1997, for example, the USDA had to purchase \$383 million worth of nonfat dry milk to support its price floor.

A price floor creates an excess supply of a good. In order to maintain the price floor, the government must prevent the excess supply from driving down the market price. In practice, the government often accomplishes this goal by purchasing the excess supply itself.

However, purchasing excess supplies of food is expensive, so price floors are usually accompanied by government efforts to *limit* any excess supplies. In the dairy market, for example, the U.S. government has developed a complicated management system to control the production and sale of fluid milk to manufacturers and processors, which helps to limit the government's costs. In other agricultural markets the government has ordered or paid farmers *not* to grow crops on portions of their land, and has imposed strict limits on imports of food from abroad. At the beginning of 2000, these supply limitations were still in use in markets for many types of dairy products, as well as for peanuts and sugar. As you can see, price floors often get the government deeply involved in production decisions, rather than leaving them to the market.

Price floors have many critics—including most economists. They have argued that the government spends too much money buying surplus agricultural products, and the resulting higher prices distort the public's buying and eating habits—often to their nutritional detriment. For example, the General Accounting Office has estimated that from 1986 to 2001, price supports for dairy products have (and will) cost American consumers \$10.4 billion in higher prices. And this does not include the cost of the health effects—such as calcium and protein deficiencies among poor children—due to decreased milk consumption. The irony is that many of the farmers who benefit from price floors are wealthy individuals or large, powerful corporations that do not need the assistance.

The U.S. government responded to these arguments with the reforms of 1996. The full or partial elimination of price floors for many farm products helped to move many of these markets closer to their equilibrium price. But most economists believe that the government did not go far enough. The government continues to prop up prices by restricting imports, and—for a few products like nonfat dry milk—it left price supports in place.



It's tempting to draw a supply and demand diagram with a price floor set *below* the equilibrium price, or a price ceiling *above* the equilibrium price. After all, a floor is usually on the bottom of something, and a ceiling is on the top. Right? In this case, wrong! A price floor set *below* the equilibrium price would have no impact on a market, because the market price would *already* satisfy the requirement that it be higher than the floor. Similarly, a price ceiling set *above* the equilibrium price would have no impact (make sure you understand why). So remember: Always draw an effective price floor *above* the equilibrium price and an effective price ceiling *below* the equilibrium price.

TAXES

Can you think of one product, service, or resource that is not taxed? In the United States, we pay taxes on most of the goods and services we buy as well as on our income and our property. Tax revenues are the primary source of the funds that keep governments operating at the local, state, and federal levels.



What Happens When Things Change?

Excise tax A tax on a specific good or service.

But in addition to providing revenue for government services, taxes also have important effects on markets: They change the behavior of buyers and sellers, and alter the equilibrium price and the equilibrium quantity of goods exchanged. In this section, we'll study a particular kind of tax called an **excise tax**. This is a tax on a specific product. In the United States, excise taxes are imposed on a variety of goods, including cigarettes, gasoline, and airline tickets. In order to see how an excise tax affects a market, we first need to interpret our now-familiar supply curve in a new way.

Remember that a supply curve shows us the quantity of a good that firms would like to sell at each possible price. In terms of the supply curve S in Figure 3, this amounts to choosing a price along the vertical axis, reading over to the supply curve, and then moving down to the horizontal axis to find the corresponding quantity supplied. For instance, at a price of \$0.90, 60 units would be supplied. But there is an equally valid and useful interpretation of the supply curve S : It shows us the *minimum* price per unit at which firms are willing to sell any particular number of units. Under this interpretation, we can choose a quantity—say, 120 units in Figure 3—and read up to the supply curve and then over to the vertical axis to find the corresponding price per unit. So, for example, firms will only supply 120 units if they are paid at least \$1.50 per unit. (We know this because at any price less than \$1.50 per unit, they would supply fewer than 120 units.)

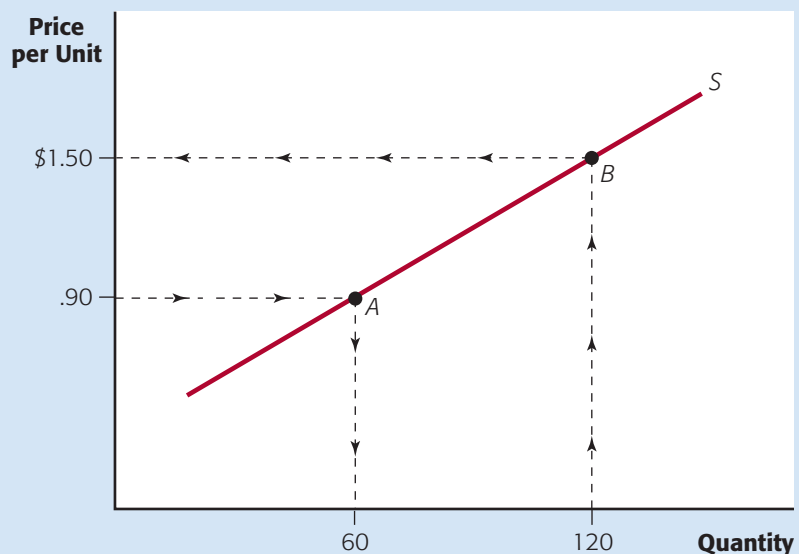
Let's use this new interpretation of the supply curve to study the excise tax on airline tickets. Figure 4 shows the market for international air travel. In the absence of the tax, supply curve S shows the minimum price the airlines must get per ticket in order to supply each number of tickets on the horizontal axis. Without any tax, the equilibrium occurs at point A , with 11.3 million tickets sold each year at a price of \$730 each.

Now suppose that the government imposes an excise tax of \$100 per ticket. This tax is to be collected from the airlines on each ticket they sell and turned over

FIGURE 3

REINTERPRETING THE SUPPLY CURVE

Any supply curve can be interpreted in two different—and equally valid—ways. First, it shows the total quantity of a good or service that all firms in a market will supply at any price. For instance, at \$0.90 per unit, point A shows that firms are willing to supply 60 units. But the supply curve also shows the minimum price per unit at which firms are willing to sell any given quantity. To sell 120 units, firms must receive at least \$1.50 per unit (point B).



THE MARKET FOR INTERNATIONAL AIR TRAVEL

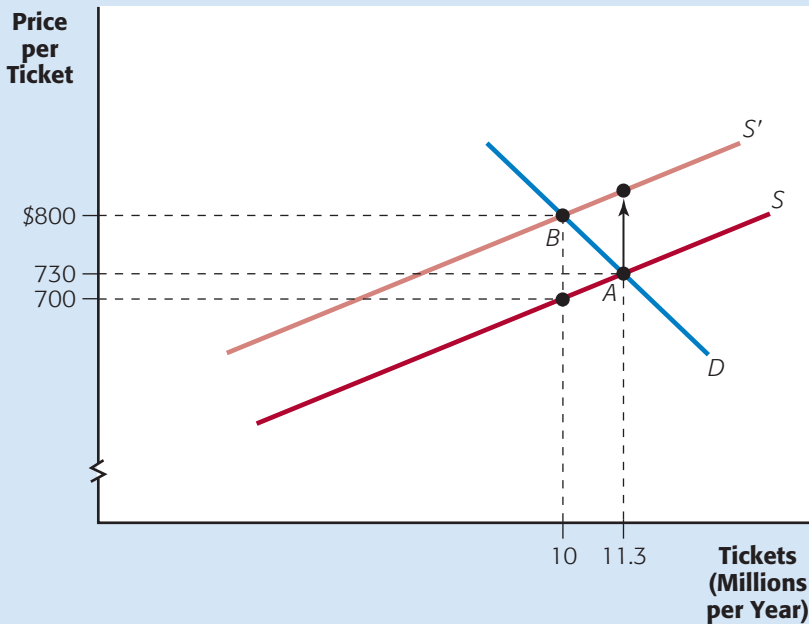


FIGURE 4

The market for international air travel is initially in equilibrium at point *A* with 11.3 million tickets sold annually at a price of \$730 per ticket. If the government imposes an excise tax of \$100 per ticket, the supply curve will shift vertically by \$100—from *S* to *S'*. The new equilibrium is at point *B* where *S'* crosses the unchanged demand curve *D*. At point *B*, consumers purchase 10 million tickets at \$800 each. Of the total revenue of \$800 per ticket, the airlines must pay \$100 per ticket in tax to the government, leaving them with a net revenue of only \$700 per ticket. Thus, after the tax is imposed, consumers end up buying fewer tickets at a higher price, and the airlines sell fewer tickets and receive a lower net price.

to the government. Will the supply curve *S* still represent the airlines' selling behavior in the market? Not at all. For example, look at point *A*. Before the tax, this point told us that the airlines would sell 11.3 million tickets only if they received at least \$730 per ticket. But now, \$100 per ticket must be turned over to the government, so a price of \$730 would leave only \$630 for the airlines. This is not enough to get the airlines to provide 11.3 million tickets. What is the minimum price the airlines must receive in order to provide 11.3 million tickets? The answer is \$830. At that price, they could pay the \$100 tax to the government, and keep \$730 for themselves—just enough to make them supply 11.3 million tickets.

The same argument could be applied to *every* quantity along the supply curve. Whatever the minimum price needed per ticket before the tax, it will be \$100 *greater* after the tax. In other words, the tax creates a *new supply curve* in the market for airline travel. The new supply curve—*S'* in Figure 4—lies \$100 above the original curve.

Note that the new supply curve *S'* tells us the minimum price that the airlines must be *paid* to sell each quantity of tickets. This is the airlines' *gross price*—what they get *before* they pay the tax. But what is the airlines' *net price* per ticket—the amount they actually get to *keep*? To find that, we must deduct the tax—\$100 per ticket—from the gross price at each quantity. That is, at each quantity, the *old* supply curve, which lies \$100 below the new one, tells us the net price—the amount that firms actually keep after paying the tax.

More generally,

an excise tax shifts the market supply curve upward by the amount of the tax. For each quantity supplied, the new, higher supply curve tells us firms' gross price, and the original, lower supply curve tells us the net price.

You can see in Figure 4 that once the excise tax is imposed, point *A* is no longer the equilibrium. With the new supply curve, the equilibrium has moved to point *B*, where the new supply curve intersects the original demand curve. At point *B*, the price consumers must pay is higher—\$800 rather than \$730—and the quantity exchanged is smaller—10 million tickets rather than 11.3 million. But what about the airlines? While their gross price is \$800 (on the new supply curve), their net price is only \$700. Thus, the excise tax has reduced the airlines' net price from \$730 to \$700.

Notice something interesting about the conclusion we've reached using Figure 4: When a tax of \$100 per ticket is put on the market, the price paid by buyers rises, but by *less than* \$100. Thus, buyers are not bearing the full monetary burden of the tax. Similarly, the (net) price received by sellers falls, but by less than the \$100 tax. Sellers are not bearing the full monetary burden of the tax either.

We can conclude that,

an excise tax on a good increases the price paid by consumers, but decreases the (net) price received by sellers. Thus, both buyers and sellers bear part of the burden of paying the tax.

In our example, buyers contribute \$70 of the tax on each ticket (their price rises from \$730 to \$800), and sellers contribute \$30 (their net price falls from \$730 to \$700).

But there is another burden imposed by the tax besides price changes: The *quantity exchanged* decreases as well. Whereas 11.3 million tickets are sold before the tax is imposed, only 10 million are sold afterward. Thus, buyers are harmed because they pay more for each ticket *and* because they buy fewer tickets. Sellers are harmed because they keep less for each ticket sold *and* because they sell fewer tickets.

Are there any rules that determine *how* the burden of an excise tax will be distributed between buyers and sellers? The answer is yes. But to understand these rules, you need to learn one more tool that economists use to analyze markets. That is the subject of the next section.

PRICE ELASTICITY OF DEMAND

Imagine that you are the mayor of one of America's large cities. Every day, the headlines blare about local problems—poverty, crime in the streets, the sorry state of public education, roads and bridges that are falling apart, traffic congestion—and you, as mayor, are held accountable for all of them. Of course, you could help alleviate these problems, if only you had more money to spend on them. But where to get the money?

One day, an aide bounds into your office. "I've got it," he says, beaming. "The perfect solution. We raise mass transit fares." He shows you a sheet of paper on which he's done the calculation: Each year, city residents take 100 million trips on public transportation. If fares are raised by 50 cents, the transit system will take in an additional \$50 million—enough to make a dent in some of the city's problems.

You stroke your chin and think about it. So many issues to balance: fairness, practicality, the political impact. But if you have taken the first week or two of introductory microeconomics, another thought will occur to you: Your aide has made a serious mistake! Public transportation—like virtually everything else that people buy—obeys the law of demand: A rise in price—with no other change—will cause a decrease in quantity demanded. If you raise fares, each *trip* will bring in more revenue, but *there will be fewer trips taken*. If the impact on the number of trips is small, mass transit revenue might rise. But if people begin to abandon mass transit

in droves, the city will be much worse off, actually *losing* revenue, even as those continuing to ride pay higher fares. How can you determine the ultimate impact of the fare hike on the city's revenue?

To answer that question, you would need one more piece of information. And the same information is needed by anyone who needs to know how a change in price affects his revenue: a theater setting ticket prices, a cell phone company setting the price per minute for phone calls, or a doctor deciding on patients' fees. The information you need concerns something that economists call the *price elasticity of demand*, which is a measure of how *sensitive* quantity demanded is to a change in price.

There are many different ways to measure the sensitivity of quantity demanded to price. The elasticity approach—which has proven the most useful—compares the *percentage change in quantity demanded* with the *percentage change in price*.

More specifically:

the price elasticity of demand (E_D) for a good is the percentage change in quantity demanded divided by the percentage change in price:

$$E_D = \frac{\% \Delta Q^D}{\% \Delta P}.$$

For example, if a 2% rise in the price of newspapers causes a 3% drop in the quantity of newspapers demanded, then $E_D = \% \Delta Q^D / \% \Delta P = -3\% / 2\% = -1.5$. We would say, “The price elasticity of demand for newspapers is minus 1.5.”

There are a few things to keep in mind about a price elasticity of demand (or just *elasticity of demand*, for short). First, it will virtually always be a *negative* number: As long as the good obeys the law of demand, a positive change in price ($\% \Delta P > 0$) will cause a negative change in quantity demanded ($\% \Delta Q^D < 0$), so the ratio of the two ($\% \Delta Q^D / \% \Delta P$) must have a minus sign.

Second, an elasticity of demand has a straightforward interpretation: It tells us the percentage change in quantity demanded *for each 1-percent increase* in price. An elasticity of -2.5 , for example, tells us that if price rises by 1 percent, quantity demanded falls by 2.5 percent. If price rises by 2 percent, quantity demanded falls by 5 percent, and so on. In general, the greater the absolute value of the number, the more sensitive quantity demanded is to price: An elasticity of -2.5 means greater price sensitivity than an elasticity of -1 or -0.5 .

Finally, keep in mind that a demand elasticity tells us the response of quantity demanded to a price change *if all other influences on demand remain unchanged*. We are interested in the pure effect of a price change on quantity demanded, uncluttered by changes in other prices, income, tastes, or other variables. Elasticity tells us the change in quantity we *would* observe if just the price of the good changed and nothing else did. In other words,

a price elasticity of demand tells us the percentage change in quantity demanded caused by a 1-percent rise in price as we move along a demand curve from one point to another.

Price elasticity of demand The sensitivity of quantity demanded to price; the percentage change in quantity demanded caused by a 1-percent change in price.

CALCULATING PRICE ELASTICITY OF DEMAND

Suppose that you know the demand curve for a product; that is, you know what quantity consumers in a market would like to buy at each possible price. You would still have one more task in order to calculate a demand elasticity: measuring the *percentage change* in both quantity demanded and price.



It's tempting to calculate an elasticity from simple observation: looking at what actually happened to buyers' purchases after some price changed. But this often leads to serious errors. Elasticity of demand tells us the effect a price change would have on quantity demanded if all other influences on demand remain unchanged. But in the real world, it is unlikely that other influences will remain unchanged in the weeks or months after a price change.

Consider what happened in Baltimore in March 1996, when the city increased mass transit fares by 8 percent. Over the next six months, ridership increased by 4.5 percent. Does this mean that the elasticity of demand for mass transit in Baltimore is positive? Does mass transit violate the law of demand? Not at all. Around the time of the fare hike, the city also made improvements in service and advertised them heavily. This no doubt helped to change tastes in favor of mass transit, shifting the demand curve rightward. If all other influences on demand for mass transit had remained unchanged, ridership would no doubt have fallen.

Economists and statisticians have developed tools to isolate the effect of price changes on quantity demanded when other variables are changing at the same time. If you major in economics, you will learn some of these tools in a course with a title such as econometrics, statistical methods, or quantitative analysis.

Percentage Changes for Elasticities.

A percentage change is *usually* defined as the change in a variable divided by its starting, or base, value. (See the Appendix to Chapter 1 on percentage changes.) But this can create a problem when we use elasticities.

For example, look at Figure 5, which shows a hypothetical monthly demand curve for laptop computers in the United States. As we move from point *A* to point *B* on this curve, the price of an average laptop rises from \$1,000 to \$1,500. The corresponding *percentage* change in price—using our starting price of

\$1,000 as the base price—would be $(\$1,500 - \$1,000)/\$1,000 = 0.50$ or 50 percent. But what if—instead of moving from *A* to *B*—we move from *B* to *A*? Then, instead of increasing from \$1,000 to \$1,500, the price would *decrease* from \$1,500 to \$1,000. In this case, our base price would be \$1,500, and our percentage change in price would now become $(\$1,000 - \$1,500)/\$1,500 = -0.33$, or -33 percent. So our measure of the change in price between two points on the demand curve—and our measure of price elasticity that is based on it—would depend on whether the price was rising or falling over the interval. The same is true of quantity demanded: The percentage change would depend on the *direction* of the change.

In order to ensure that the elasticity of demand over an interval is the same number whether the price increases or decreases over the interval, we adopt a simple convention when calculating elasticities: *The base value used to calculate a percentage change in a variable is always midway between the initial value and the new value.* Thus, if the price rises from \$1,000 to \$1,500, or falls from \$1,500 to \$1,000, we use as our base price the value midway between these two prices, found by calculating their simple average: $(\$1,000 + \$1,500)/2 = \$1,250$. This way, we are using the same base value regardless of the direction that price changes.

More generally, when price changes from any value P_0 to any other value P_1 , we define the percentage change in price as

$$\% \Delta P = \frac{(P_1 - P_0)}{\left[\frac{(P_1 + P_0)}{2} \right]}$$

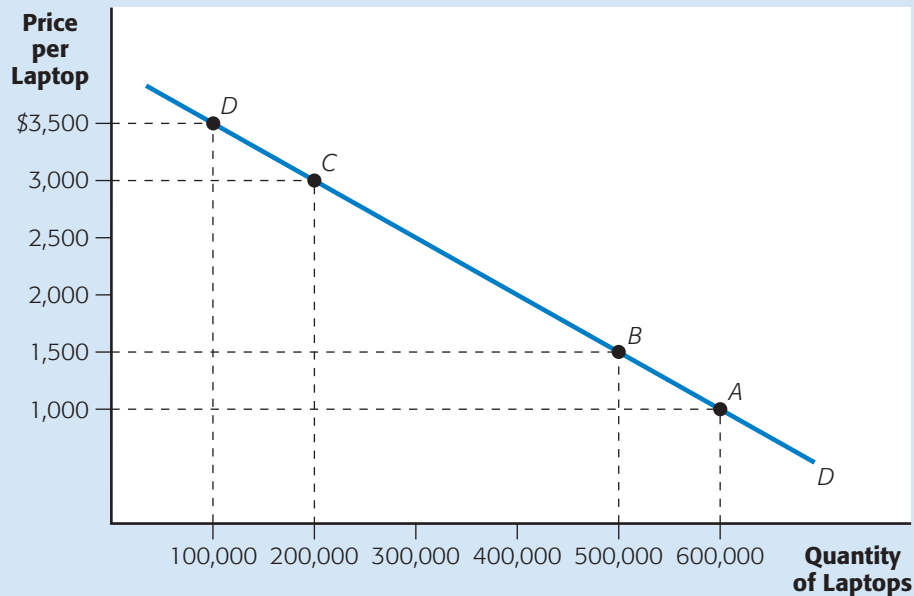
The term in the numerator is the change in price; the term in the denominator is the base price—the midpoint between the two prices. If you plug the preceding numbers into this formula, you'll see that if price rises from \$1,000 to \$1,500, the percentage change in price is $(\$1,500 - \$1,000)/\$1,250 = 0.40$ or 40 percent. If price falls from \$1,500 to \$1,000, the percentage change is $(\$1,000 - \$1,500)/\$1,250 = -0.40$ or -40 percent.

The percentage change in quantity demanded is calculated in a similar way. When quantity demand changes from Q_0 to Q_1 , the percentage change is calculated as

CALCULATING PRICE ELASTICITY OF DEMAND

FIGURE 5

Movement Along Demand Curve	$\% \Delta Q^D$	$\% \Delta P$	Elasticity of Demand
Point A to Point B	$(500,000 - 600,000)/650,000$ = -0.182 or -18.2%	$(\$1,500 - \$1,000)/\$1,250$ = 0.40 or 40%	-18.2%/40% = -0.46
Point C to Point D	$(100,000 - 200,000)/150,000$ = -0.667 or -66.7%	$(\$3,500 - \$3,000)/\$3,250$ = 0.154 or 15.4%	-66.7%/15.4% = -4.33



$$\% \Delta Q^D = \frac{(Q_1 - Q_0)}{\left[\frac{(Q_1 + Q_0)}{2} \right]}$$

Once again, we are using the number midway between the initial and the new quantity demanded as our base quantity.

Using the Formula. Now let's calculate an elasticity of demand for laptop computers using the data in Figure 5. For now, we'll stick to the interval from point A to point B. As price rises from \$1,000 to \$1,500, quantity demanded falls from 600,000 to 500,000. We have

$$\% \Delta Q^D = \frac{(500,000 - 600,000)}{\left[\frac{(500,000 + 600,000)}{2} \right]} = \frac{-100,000}{550,000} = -0.182, \text{ or } -18.2 \text{ percent.}$$

$$\% \Delta P = \frac{(\$1,500 - \$1,000)}{\left[\frac{(\$1,500 + \$1,000)}{2} \right]} = \frac{\$500}{\$1,250} = 0.400, \text{ or } 40.0 \text{ percent.}$$

$$E_D = \frac{-0.182}{0.400} = -0.46.$$

We find that, over the interval from point *A* to *B* in Figure 5, the quantity of laptops demanded falls by 0.46 percent—a little less than half a percent—for each 1-percent increase in price.

A Shortcut. In practice, there is an easier way to calculate elasticity. Starting with the definition

$$E_D = \frac{\% \Delta Q^D}{\% \Delta P}$$

we can substitute in and then rearrange terms as follows:

$$E_D = \frac{\frac{Q_1 - Q_0}{\frac{1}{2}(Q_1 + Q_0)}}{\frac{P_1 - P_0}{\frac{1}{2}(P_1 + P_0)}} = \frac{(Q_1 - Q_0)}{(Q_1 + Q_0)} \times \frac{(P_1 + P_0)}{(P_1 - P_0)}.$$

Applying this shortcut method to our data for laptops, we obtain

$$E_D = \frac{500,000 - 600,000}{500,000 + 600,000} \times \frac{\$1,500 + \$1,000}{\$1,500 - \$1,000} = -0.091 \times 5 = -0.46,$$

which is exactly what we obtained earlier.

ELASTICITY AND STRAIGHT-LINE DEMAND CURVES

In Figure 5, we drew the demand curve for laptops as a straight line. Along this demand curve, each time price rises by \$500, the quantity of laptops demanded decreases by 100,000 per month. This behavior remains constant regardless of the price at which we start. Does this mean that the price elasticity of demand for laptops is the same for any interval along this demand curve? Absolutely not!

To see why, let's compare what happens when the price of laptops rises by \$500 along two different intervals. If we move from *A* to *B*, the price rise of \$500 corresponds to a *percentage* price rise of \$500/\$1,250 = 0.40 or 40 percent. But if we move from *C* to *D*—another \$500 increase in price—the *percentage* rise in price is \$500/\$3,250 = 0.154 or 15.4%. In other words, the same *absolute* price increase corresponds to a smaller *percentage* increase. In general, as we move upward and leftward along a straight-line demand curve, the same absolute increment in price will correspond to smaller and smaller percentage increments in price. Why? Because the base price used to calculate percentage changes keeps rising.

Something similar happens as quantity changes. Whether we move from *A* to *B* or from *C* to *D* quantity demanded falls by the same number: 100,000. But the *percentage* drop in quantity demanded is greater along the interval *C* to *D* because the base quantity there is smaller. In general, as we move upward and leftward along a straight-line demand curve, the same *absolute* decrease in quantity corresponds to larger and larger *percentage* decreases in quantity.

Figure 6 summarizes what we've just discovered about any straight-line demand curve. As we move upward and leftward by equal distances, the percentage change in quantity rises, while the percentage change in price falls. Together, this means that the price elasticity of demand must be getting larger.

Elasticity of demand varies along a straight-line demand curve. More specifically, demand becomes more elastic as we move upward and leftward.

Let's check this result by going back to Figure 5 and calculating the elasticity of demand along the interval from point C to point D.

$$\% \Delta Q^D = \frac{(Q_1 - Q_0)}{\left[\frac{(Q_1 + Q_0)}{2} \right]} = \frac{(100,000 - 200,000)}{150,000} = -0.667, \text{ or } -66.7 \text{ percent.}$$

$$\% \Delta P = \frac{(P_1 - P_0)}{\left[\frac{(P_1 + P_0)}{2} \right]} = \frac{(\$3,500 - \$3,000)}{\$3,250} = 0.154, \text{ or } 15.4 \text{ percent.}$$

$$E_D = \frac{-0.667}{0.154} = -4.33.$$

As expected, demand is more elastic (-4.33) over this interval than it is over the interval from A to B that we calculated earlier (-0.46).

CATEGORIZING GOODS BY ELASTICITY

When the numerical value of the price elasticity of demand is *between 0 and -1.0* , we say that demand is **inelastic**. When demand for a good is inelastic, the absolute value of the elasticity will be smaller than 1.0, that is,

Inelastic demand A price elasticity of demand between 0 and -1 .

ELASTICITY AND STRAIGHT-LINE DEMAND CURVES

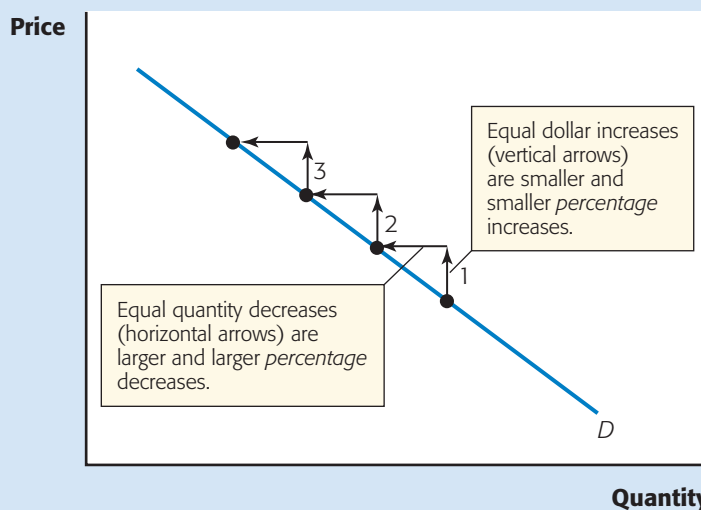


FIGURE 6

Elasticity varies along a straight-line demand curve. As we move in equal increments upward and leftward along the demand curve (indicated by the arrows), the percentage change in quantity demanded rises, while the percentage change in price falls. Therefore, demand becomes more elastic.

$$\left| \frac{\% \Delta Q^D}{\% \Delta P} \right| < 1.0.$$

Or, rearranging, we obtain

$$|\% \Delta Q^D| < |\% \Delta P|.$$

In words, inelastic demand means that the percentage change in quantity demanded will be *smaller* than the percentage change in price, ignoring the sign. For example, if price rises by 4 percent, quantity demanded will fall, but by *less* than 4 percent. When demand is inelastic, quantity demanded is *not* very sensitive to price.

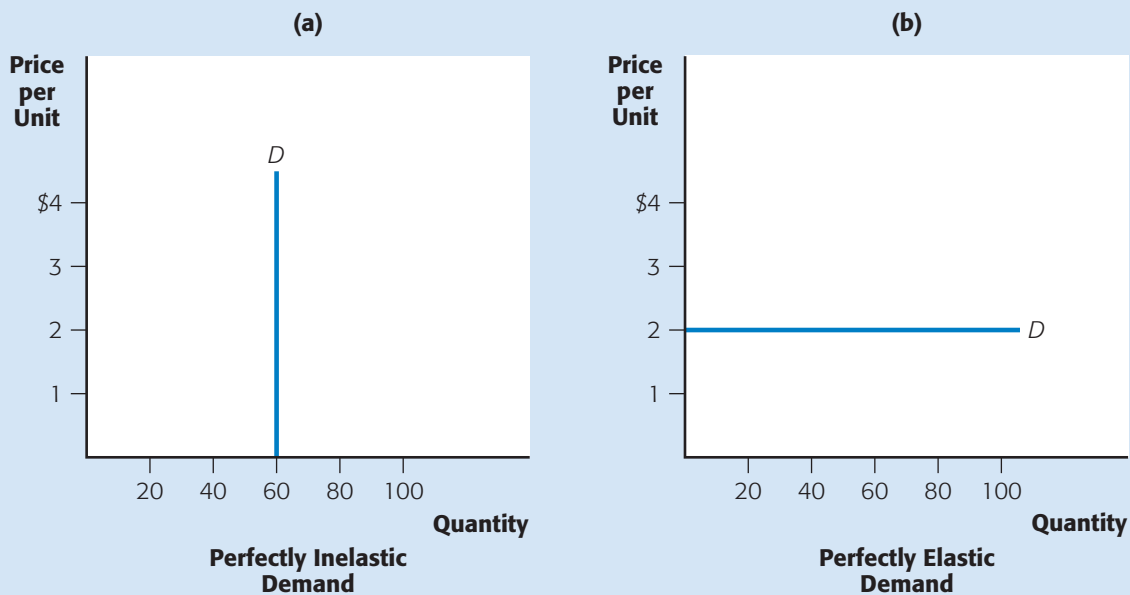
An extreme case of inelastic demand occurs when a change in price causes absolutely no change in quantity demanded at all. In this case, since $\% \Delta Q^D = 0$, the elasticity will equal zero. We call this special case **perfectly inelastic** demand. Panel (a) of Figure 7 shows what the demand curve for a good would look like if demand were perfectly inelastic at every price. The demand curve is vertical: No matter what the price, quantity demanded is the same.

Perfectly inelastic demand is mostly interesting from a theoretical point of view; it is difficult to find examples of goods with zero elasticity of demand in the real world. With zero demand elasticity, the good would have to be one that consumers want only in a fixed quantity. One example might be insulin—the drug needed by diabetics to control their blood sugar. Insulin has no use other than in the management of diabetes. For diabetics, quantity requirements for insulin are quite rigid, and there are no substitutes for its use. A drop in price will not encourage diabetics to use more, nor will a modest rise in price cause diabetics to economize on its use.

Perfectly inelastic demand A price elasticity of demand equal to 0.

FIGURE 7

EXTREME CASES OF DEMAND



The vertical demand curve of panel (a) represents the case of perfectly inelastic demand. At every price, the same quantity is demanded. The horizontal curve in panel (b) represents perfectly elastic demand. A small change in price would lead to an extremely large change in quantity demanded.

When E_D is less than -1.0 , we say that demand is **elastic**. In this case, the absolute value of the elasticity will be greater than 1.0 :

$$\left| \frac{\% \Delta Q^D}{\% \Delta P} \right| > 1.$$

Or, rearranging, we get

$$|\% \Delta Q^D| > |\% \Delta P|.$$

When demand is elastic, the percentage change in quantity demanded is *larger* than the percentage change in price, ignoring the signs. For instance, if price rises by 4 percent, quantity demanded will fall by *more* than 4 percent. Elastic demand means that quantity demanded is *sensitive to price*.

An extreme case of price sensitivity occurs when demand is **perfectly** or **infinitely elastic**. Even the tiniest change in price causes a huge change in quantity demanded, so huge that, for all intents and purposes, we can call the response infinite. When demand is perfectly elastic over every interval, the demand curve will be a horizontal line—as shown in panel (b) of Figure 7. The demand for a single brand of salt may fall into this category. If the price of Morton salt rose a little, while other brands next to it on the supermarket shelf continued to cost the same, virtually everyone would switch to the other brands, causing the quantity of Morton salt demanded to plummet.

Finally, when elasticity of demand is exactly equal to -1 , we have **unitary elasticity**. In this case, $|\% \Delta Q^D| = |\% \Delta P|$, and demand for the good is exactly at the boundary between elastic and inelastic. Many consumer products seem to have price elasticities near -1.0 . In addition, a price elasticity of -1.0 is important as a benchmark case, as you will see a bit later.

Elastic demand A price elasticity of demand less than -1 .

Perfectly (infinitely) elastic demand

A price elasticity of demand approaching minus infinity.

Unitary elastic demand A price elasticity of demand equal to -1 .

ELASTICITY AND TOTAL EXPENDITURE

When the price of a good increases, the law of demand tells us that people will demand less of it. But this does not necessarily mean that they will *spend* less on it. After the price rises, fewer units will be purchased, but each unit will cost more. It turns out that whether total spending on the good rises or falls depends entirely on the price elasticity of demand for the good.

To see this more formally, note that the total expenditure (TE) on a good is defined as

$$TE = P \times Q$$

where P is the price per unit and Q is the quantity purchased. We can use a rule about percentage changes, explained in the Appendix to Chapter 1: *When two numbers are both changing, the percentage change in their product is (approximately) the sum of their individual percentage changes.* Applying this to total expenditure, we can write



You've seen that elasticity changes along a straight-line demand curve. But the result applies more generally as well. Except in special cases (such as those in Figure 7), elasticity can change along *any* demand curve, whether a straight line or a curve. For this reason, you should try to avoid two common mistakes. First, don't describe a "demand curve" as elastic or inelastic; while demand might be elastic along *part* of the demand curve, it might be inelastic along another part of the curve.

Second, don't equate the "flatness" or "steepness" of a demand curve with how elastic or inelastic it is. "Steepness" and "flatness" refer to the slope of a demand curve—the absolute change in one variable divided by another. Elasticity, on the other hand, refers to the percentage change in one variable divided by the percentage change in the other. Slope and elasticity are not the same. A straight-line demand curve, for example, remains equally steep or flat along its entire length. Yet—as you've seen—the elasticity of demand changes as we move along it.

$$\% \Delta TE = \% \Delta P + \% \Delta Q.$$

Now let's assume that P rises by 10 percent. What will happen to total expenditure? If demand is *unitary elastic*, then Q will fall by 10 percent, so we will have

$$\% \Delta TE = 10 \text{ percent} + (-10 \text{ percent}) = 0.$$

The percentage change in total expenditure is zero, meaning that total expenditure does not change at all! If demand is *inelastic*, a 10-percent rise in price will cause quantity demanded to fall by *less* than 10 percent, so we have

$$\% \Delta TE = 10 \text{ percent} + (\text{something less negative than } -10 \text{ percent}) > 0.$$

The percentage change in total expenditure is greater than zero, so total expenditure rises. Finally, if demand is *elastic*, so that Q falls by more than 10 percent, TE will fall:

$$\% \Delta TE = 10 \text{ percent} + (\text{something more negative than } -10 \text{ percent}) < 0.$$

Of course, the results we just obtained for a price increase of 10 percent would hold for any price change—increase or decrease. Our conclusions about elasticity and total expenditure are presented in Table 1. They can be summarized as follows:

Where demand is price inelastic, total expenditure moves in the same direction as price. Where demand is elastic, total spending moves in the opposite direction from price. Finally, where demand is unitary elastic, total expenditure remains the same as price changes.

Let's check the statements in Table 1, using our hypothetical demand curve for laptop computers. The first two columns of Table 2 present familiar price and quantity pairs for laptops, taken from Figure 5. The third column lists total expenditure:

Notice what happens to total expenditure as we move along the demand curve. Demand for laptops, you recall, was inelastic ($E_D = -0.46$) when price rose from \$1,000 to \$1,500. According to the rules in Table 1, we expect a price rise to *increase* total expenditure, and that is exactly what happens: The \$500 rise in price causes total expenditure to increase from \$600 million to \$750 million. When price rose from \$3,000 to \$3,500, however, demand was elastic ($E_D = -4.33$). Our rules tell us that a rise in price should decrease total expenditure. Indeed, the \$500 price hike causes total expenditure to fall from \$600 million to \$350 million.

There is an easy way to see how a change in price changes the total expenditure of buyers, using a graph of the demand curve. Look at Figure 8. At point A , price is \$1,000 per laptop and quantity demanded is 600,000 laptops. Total expenditure is

TABLE 1

EFFECTS OF PRICE CHANGES ON EXPENDITURE

Where demand is:	A price increase will:	A price decrease will:
inelastic ($ E_D < 1$)	increase expenditure	decrease expenditure
unitary elastic ($ E_D = 1$)	cause no change in expenditure	cause no change in expenditure
elastic ($ E_D > 1$)	decrease expenditure	increase expenditure

TABLE 2

Price per Laptop (P)	Quantity Demanded (per Month) (Q)	Total Monthly Expenditure ($P \times Q$)
\$1,000	600,000	\$600 million
\$1,500	500,000	\$750 million
\$3,000	200,000	\$600 million
\$3,500	100,000	\$350 million

EFFECTS OF PRICE CHANGES FOR LAPTOP COMPUTERS

price \times quantity = $\$1,000 \times 600,000 = \600 million. But this is exactly equal to the *area* of the wider rectangle, which has a width of 600,000 and a height of \$1,000. Thus, the area of this rectangle shows total expenditure on the good when price is \$1,000. More generally,

At any point on a demand curve, buyers' total expenditure is the area of a rectangle with width equal to quantity demanded and height equal to price.

Now suppose that price rises from \$1,000 to \$1,500, so we move along the demand curve to point B , where quantity demanded drops to 500,000. Here, total expenditure is $\$1,500 \times 500,000 = \750 million, given by the area of the taller rectangle, with width equal to 500,000 and height equal to \$1,500. You can see that the area of the total expenditure rectangle drawn for price = \$1,500 is larger than the area of the total expenditure rectangle for price = \$1,000. This confirms what we know already from Table 2: The rise in price from \$1,000 to \$1,500 causes total expenditure to increase because demand is inelastic for that price change.

ELASTICITY AND TOTAL EXPENDITURE

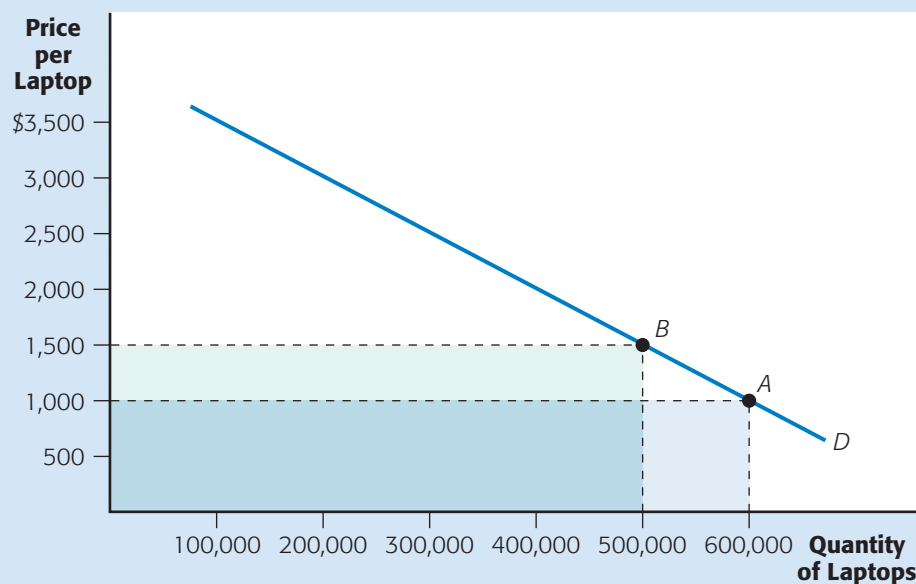


FIGURE 8

Any point along a demand curve defines a rectangle whose area indicates total expenditure on the good. At point A , where price is \$1,000 and 600,000 laptops are demanded, expenditure is \$600 million. At point B , expenditure is \$750 million. Moving from A to B , expenditure increases, so demand must be inelastic over that range.

Finally, there is one important implication of our elasticity–total expenditure rule. What a buyer spends, a seller receives. Therefore, the total amount that consumers spend on a good—which we’ve called *total expenditure*—is also the *total sales revenue* of sellers. This is one reason why knowing the price elasticity of demand for their product can be so important to firms. In some cases, the price elasticity of demand is all a firm needs to forecast its future revenues. Some of the end-of-chapter problems will show you how this is done.

DETERMINANTS OF ELASTICITY

Table 3 lists the price elasticity of demand for several goods and services. Keep in mind that these elasticities are calculated for a specific range of prices that have been observed in the past. If a large price change moved us out of the range of past observations, the elasticity might be very different. For example, although the elasticity of demand for gasoline is -0.20 when the price varies in a range from \$1.00 to \$2.00 per gallon, the elasticity might be very different for price changes in a range from \$10.00 to \$15.00 per gallon, which have never been observed.

Notice that all of the price elasticities of demand are negative: Each of these goods obeys the law of demand. Even cigarettes—which are highly addictive—have an elasticity less than zero: A rise in price reduces the quantity of cigarettes demanded.

You can also see that the calculated elasticities vary widely. Why is it that demands for Tide detergent, Pepsi, and Coke are so elastic, while those for eggs and gasoline are so inelastic? More generally, what determines whether the demand for a good will be elastic or inelastic? Two characteristics seem to be the most important determinants of elasticity: the availability of substitutes, and the importance in the buyers’ budget.

Availability of Substitutes. When the price of a good rises, we look for substitutes. If close substitutes are easy to find, we can cut back on our purchases of the good in question, and demand is more elastic. If close substitutes are difficult to find, we can’t cut back as much, and so demand is less elastic.

This logic helps explain some of the differences in elasticity values found in Table 3. In spite of what the commercials tell us, most of us recognize that Coke is an extremely close substitute for Pepsi (and vice versa). And there are a variety of other *reasonably* close substitutes for Pepsi, such as other carbonated soft drinks, iced tea, or fruit juice. This helps to explain why a 10-percent rise in the price of Pepsi would lead to more than a 20-percent decline in quantity demanded. By contrast, there are fewer close substitutes for eggs—especially if you are baking from a recipe—or for gasoline, especially if you need to drive to work. This helps to explain the relatively low elasticity values for these goods.

Substitutability can be a slippery concept, however, and we need to be careful when we use it. Remember that, in analyzing any problem, the first Key Step of our four-step procedure is to define the market we are dealing with. You may also remember that we can choose to define a market in different ways, depending on the question we want to analyze. But it turns out that the elasticity value we will use in analyzing a problem depends crucially on *how* broadly or narrowly we define the market itself. After all, it is easier to find substitutes for a narrowly defined good (Pepsi) than for a broadly defined good (bottled drinks). Therefore,

the more narrowly we define a good, the easier it is to find substitutes, and the more elastic is the demand for the good. The more broadly we define a good, the harder it is to find substitutes and the less elastic is the demand for the good.

TABLE 3

Specific Brands	Narrow Categories	Broad Categories
Tide Detergent	Transatlantic Air Travel	Recreation
	Tourism in Thailand	
Pepsi	Ground Beef	Clothing
Coke	Pork	Food
	Milk	Imports
	Cigarettes	Transportation
	Electricity	
	Beer	
	Eggs	
	Gasoline	
	Oil	

SOME SHORT-RUN PRICE ELASTICITIES OF DEMAND

Sources: Michael G. Vogt and Chutima Wittayakorn, "Determinants of the Demand for Thailand's Exports of Tourism," *Applied Economics*, Vol. 30, Issue 6, pp. 711–715. Sachin Gupta et al., "Do Household Scanner Data Provide Representative Inferences from Brand Choices? A Comparison with Store Data," *Journal of Marketing Research*, Fall 1996, pp. 383ff. F. Gasmí, J. J. Laffont, and Q. Vuong, "Econometric Analysis of Collusive Behavior in a Soft-Drink Market," *Journal of Economics and Management Strategy*, Summer 1992, pp. 277–311. Richard Blundell, Panos Pashardes, and Guglielmo Weber, "What Do We Learn about Consumer Demand Patterns from Micro Data?" *American Economic Review*, June 1993, pp. 570–597. Michael T. Maloney and Robert E. McCormick, "Setting the Record Straight: The Consumer Wins the Competition," *Citizens for a Sound Economy Foundation*, Issue Analysis No. 46, January 30, 1997. J. L. Sweeney, "The Response of Energy Demand to Higher Prices: What Have We Learned?" *American Economic Review*, May 1984, pp. 31–37. F. Chaloupka, "Rational Addictive Behavior and Cigarette Smoking," *Journal of Political Economy*, August 1991, pp. 722–742; J. M. Cigliano, "Price and Income Elasticities for Airline Travel," *Business Economics*, September 1980, pp. 17–21. M. D. Chinn, "Beware of Econometricians Bearing Estimates," *Journal of Policy Analysis and Management*, Fall 1991, pp. 546–557. M. R. Baye, D. W. Jansen, and Jae-Woo Lee, "Advertising Effects in Complete Demand Systems," *Applied Economics*, October 1992, pp. 1087–1096. Dale M. Heien, "The Structure of Food Demand: Interrelatedness and Duality," *American Journal of Agricultural Economics*, May 1982, pp. 213–221. Gary W. Brester and Michael K. Wohlgenant, "Estimating Interrelated Demands for Meats Using New Measures for Ground and Table Cut Beef," *American Journal of Agricultural Economics*, November 1991, pp. 1182–1194. David R. Henderson, "Do We Need to Go to War for Oil?" *Cato Foreign Policy Briefing*, No. 4, October 24, 1990.

The key is that different things are assumed constant when we use a narrow definition compared with a broader definition. Once we define the good in question, our elasticity calculations always assume that all other prices do not change. Pepsi has a large price elasticity because when the price of this particular soft drink rises, we consider the effect on quantity demanded, assuming that the prices of *all other soft drinks*, including Coke, are not changing. We therefore expect a strong quantity response as consumers switch to these other soft drinks that are now *relatively* cheaper. But suppose we had defined our good more broadly as *carbonated soft drinks*. Now, any price increase would apply to Pepsi, Coke, and *all* soft drinks at the same time. While it is still possible to substitute other drinks in place of soft drinks, it is not as easy as substituting one soft drink for another. So we expect the more aggregated item, soft drinks, to have a much lower price elasticity of demand. (Now look at the elasticity entry for Tide detergent. Suppose the good had instead been defined as "laundry detergent." Would you expect a larger or smaller elasticity value?)

Table 3 also shows that when markets are defined *very* broadly—food rather than ground beef, or transportation rather than transatlantic travel—elasticities of demand tend to be lower. There are very few substitutes for food in general. Although many people can eat less, it is not an easy adjustment to make. The same is true for other broad categories, such as recreation, transportation, and clothing.

The ability to find substitutes for goods also depends on our tastes. Goods that we think of as *necessities*—for example, medical care, food, and housing—are

difficult to find substitutes for. Goods that we think of as *luxuries*—like a trip to Europe or recreation—can be substituted for more easily. We expect necessities to be less price elastic than luxuries, and Table 3 confirms this. The demand for food is less elastic than the demand for recreation, and the demand for milk is less elastic than the demand for transatlantic travel.

In general, the more “necessary” we regard an item, the harder it is to find substitutes, and the less elastic is demand for the good.

But here, too, how broadly or narrowly we define the good makes an important difference. Many goods we would consider necessities when broadly defined (e.g., medical care) become easy-to-substitute-for luxuries when more narrowly defined (e.g., visits to Dr. Hacker). When the price of *all medical care* rises, we expect a relatively small decrease in quantity demanded. But if the price of just *Dr. Hacker’s medical care* rises, the quantity response should be much larger.

Finally, the ease with which we can substitute one good for another will usually depend heavily on the *time horizon* of our analysis. The elasticities in Table 3 are all **short-run elasticities**—in which the quantity response is measured just a short time—say, a few months—after a price change. A **long-run elasticity** measures the quantity response after a year or more has elapsed. In study after study, we find that long-run elasticities are generally larger than short-run elasticities.

Why? Because it is easier for consumers to find substitutes when they have more time to do so. For example, while the *short-run* elasticity for gasoline is relatively low—about -0.2 —most studies show a *long-run* elasticity at least three times as great. This is because some of the adjustments needed to substitute for gasoline—like buying a more fuel-efficient car—take some time. Table 4 lists some of the ways households would adjust to a significant rise in the price of gasoline over the short run and the long run. Notice that the options available in the long run have a greater potential impact on consumers’ demand for gasoline than the options available in the short run.

Other goods show a similar pattern of greater elasticity in the long run than the short run. Estimates of long-run elasticities for cigarettes and electricity— -0.80 and -0.97 , respectively—are each about twice as large as their short-run counterparts in Table 3.

Short-run elasticity An elasticity measured just a short time after a price change.

Long-run elasticity An elasticity measured a year or more after a price change.

TABLE 4

ADJUSTMENTS AFTER A RISE IN THE PRICE OF GASOLINE
Short Run (a few months or less)

Use public transit more often
 Arrange a car pool
 Get a tune-up
 Drive more slowly on the highway
 Eliminate unnecessary trips (use mail order instead of driving to stores; locate goods by phone instead of driving around; shop for food less often and buy more each time)
 If there are two cars, use the more fuel-efficient one

Long Run (a year or more)

Buy a more fuel-efficient car
 Move closer to your job
 Switch to a job closer to home
 Move to a city where less driving is required

It is usually easier to find substitutes for an item in the long run than in the short run. Therefore, demand tends to be more elastic in the long run than in the short run.

Importance in the Buyer's Budget. When a good takes up a large part of your budget, a price change has a large impact on how much money you have left to spend on other goods. For example, most people spend a large fraction of their budget on housing. If the price of housing rises by, say, 10 percent, the impact on people's budgets would be substantial. As a result, people would try hard to economize on housing (move to a smaller apartment, or live with a roommate). We thus expect housing to have a large elasticity of demand.

In general,

the more of their total budgets that households spend on an item, the more elastic is demand for that item.

For example, a trip to Europe would take a big bite out of most people's budgets. A rise in price will therefore make consumers think very carefully about substitutes—traveling to Canada or Mexico, perhaps. This is partly why the demand for transatlantic air travel is so elastic.

For the opposite extreme, consider the case of ordinary table salt. A family with an income of \$50,000 per year will typically spend less than 0.005 percent of it on salt. The price of salt could double—even triple or quadruple or quintuple—and still have virtually no impact on that family's ability to afford other goods. Economically, there is little to be gained by cutting back on salt consumption when its price rises, so we expect it to be relatively price *inelastic*.¹

USING PRICE ELASTICITY OF DEMAND

Knowing the price elasticity of demand for a good and understanding the link between elasticity and total expenditure or revenue is helpful in many different contexts. For example, producers of goods and services—doctors, bakers, theater owners, manufacturers, and others—can use price elasticity of demand to predict how a price change will affect their total sales revenue. And government policy makers can and do use demand elasticities to price many government services, to make tax policy, and to design programs to help the needy. The concept of demand elasticity is even at the center of the debate over the war on drugs in the United States and many other countries, as the next section shows.

The War on Drugs. In 1999, the U.S. government spent about \$18 billion intervening in the market for illegal drugs like cocaine, heroin, and marijuana. Most of this money is spent on efforts to restrict the *supply* of drugs. But many economists argue that society would be better off if antidrug efforts were shifted from the supply side to the demand side of the markets. Why? The answer hinges on the price elasticity of demand for illegal drugs.



What Happens When Things Change?

¹ Earlier, we argued that the demand for one brand of table salt should be perfectly *elastic*. Now, we're suggesting that the demand for salt should be *inelastic*. Is this a contradiction? Not at all. Can you explain why? (*Hint*: Are we defining our market the same way in both statements?)



The war on drugs has focused on decreasing supply.

Look at Figure 9(a), which shows the market for heroin if there were no government intervention. The equilibrium would be at point A , with price P_1 and quantity Q_1 . Total expenditure on heroin would be the area of the shaded rectangle, $P_1 \times Q_1$.

Panel (b) of the figure shows the impact of a policy to restrict supply through any one of several methods, including vigilant customs inspections, arrest and stiff penalties for drug dealers, or diplomatic efforts to reduce drug traffic from producing countries like Colombia and Thailand. The decrease in supply is represented by a leftward shift of the supply curve, establishing a new equilibrium at price P_2 and quantity Q_2 . As you can see, supply restrictions, if they successfully reduce the equilibrium quantity of heroin, will also raise its equilibrium price.

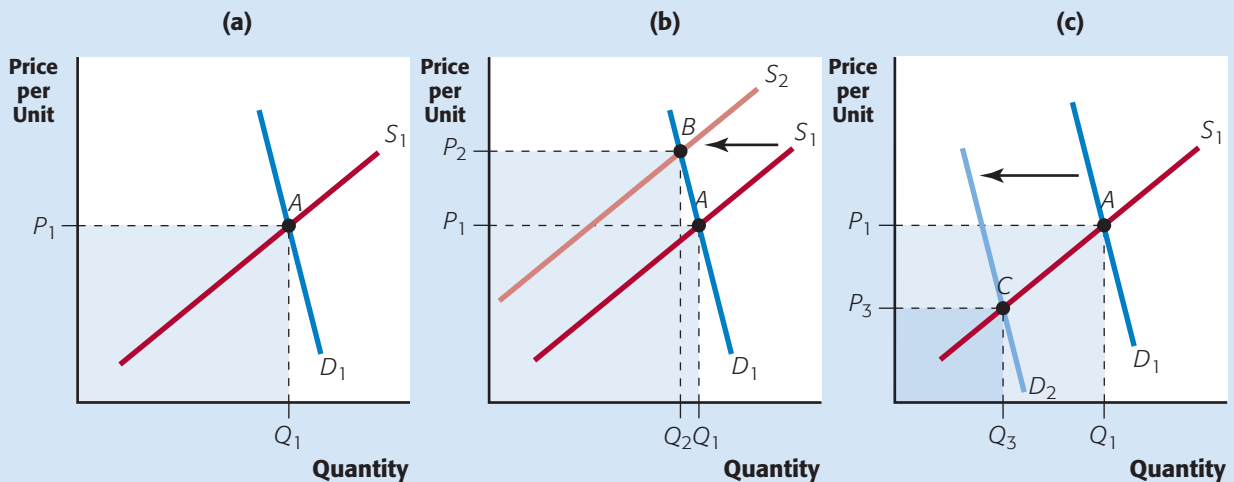
But now let's consider the impact of this policy on the total expenditure on drugs. The demand for addictive drugs such as heroin and cocaine is price *inelastic*. As you've learned, when demand is inelastic, a rise in price will *increase* total expenditure. This means that a policy of restricting the supply of illegal drugs, if successful, will also increase the total expenditure of drug users on their habit. In panel (b), total expenditure rises from the area of the shorter rectangle to the area of the taller one.

The change in total expenditure has serious consequences for our society. Many drug users support their habit through crime. If the total expenditure needed to support a drug habit rises, they may commit more crimes—and more serious ones. And don't forget that the total expenditure of drug users is also the total *revenue* of the illegal drug industry. The large revenues—and the associated larger profits to be made—attract organized as well as unorganized crime and lead to frequent and very violent turf wars.

The same logic, based on the inelastic demand for illegal drugs, has led many economists to advocate a shift of emphasis from decreasing supply to decreasing de-

FIGURE 9

THE WAR ON DRUGS



Panel (a) shows the market for heroin in the absence of government intervention. Total expenditures—and total receipts of drug dealers—are given by the area of the shaded rectangle. Panel (b) shows the effect of a government effort to restrict supply: Price rises, but total expenditure increases. Panel (c) shows a policy of reducing demand: Price falls, and so does total expenditure.

mand. Policies that might decrease the demand for illegal drugs and shift the demand curve leftward include stiffer penalties on drug *users*, heavier advertising against drug use, and greater availability of treatment centers for addicts. In addition, more of the effort against drug sellers could be directed at retailers rather than those higher up the chain of supply. It is the retailers who promote drugs to future users and thus increase demand. Panel (c) illustrates the impact these policies, if successful, would have on the market for heroin. As the demand curve shifts leftward, price *falls* from P_1 to P_3 , and quantity demanded falls from Q_1 to Q_3 . Now, we cannot say whether the drop in quantity will be greater under a demand shift than a supply shift (it depends on the relative sizes of the shifts). But we *can* be sure that a demand-focused policy will have a very different impact on equilibrium price, moving it down instead of up. Moreover, the demand shift will decrease total expenditure on drugs—to the *inner* shaded rectangle—since both price and quantity decrease. This can contribute to a lower crime rate by drug users and make the drug industry less attractive to potential dealers and producers.

Mass Transit. Earlier in this section, you were asked to imagine that you were mayor of a large city considering an increase in mass transit fares. Your assistant advised you to do it, since you would collect more revenue on each commuter trip. But you were worried that raising fares might cause so many more people to stop using mass transit that your revenue would actually decline. Can elasticity help here?

Very much so. Elasticity studies show that the long-run demand for mass transit is inelastic, which tells us that raising the fare would *increase* revenue. More specifically, the long-run elasticity of demand in large cities (those with more than one million inhabitants) averages around -0.36 . In words: A 1-percent increase in fares would decrease ridership by about a third of a percent.

Let's use this elasticity figure to analyze what would happen if New York City raised the price of its subway and bus rides from \$1.50 to \$2.00. Since this would be an increase of about 29 percent, we could expect ridership to change by $0.29 \times -0.36 = -0.103$, or a decrease of about 10 percent. In the late 1990s, commuters took about 1.7 billion trips per year on New York buses and subways, for a total revenue of $1.7 \text{ billion} \times \$1.50 = \$2.55 \text{ billion}$. The price hike would decrease the total number of trips by 10.3 percent, to about 1.53 billion, but also raise the revenue from each trip to \$2.00. Thus, total revenue would be about $1.53 \text{ billion} \times \$2.00 = \$3.06$. Comparing \$2.55 billion with \$3.06 billion, we see that the fare hike would increase total revenue by about half a billion dollars—a substantial increase.

Why, then, doesn't New York raise the mass transit fare to \$2.00? In fact, why stop at \$2.00? If the demand remains inelastic, why not continue to raise fares to \$2.50, or \$3.00, or even higher? In fact, why don't cities across the country raise *their* fares above present levels as well?

The answer is that generating revenue is only *one* goal that city governments consider in pricing mass transit. In addition to obtaining revenue, city officials want to provide an affordable means of transportation to low-income households, to manage traffic congestion on city streets, and to limit pollution of city air. To accomplish these other goals requires a large ridership. A fare increase, even if it would raise total revenue, would decrease total ridership and require the city to sacrifice these other goals. This is what keeps mass transit fares lower than the revenue-maximizing fare.

An Oil Crisis. For the past five decades, the Middle East has been a geopolitical hot spot. And the stakes for the rest of the world are high because the region produces about one-fifth of the world's oil supply. That is why the U.S. military is



What Happens When Things Change?



What Happens When Things Change?

constantly asking “what if” questions and making contingency war plans to respond to hypothetical crisis situations.

And elsewhere in government, economic officials are constantly asking their own set of “what if” questions. One central question is this: If an event in the Middle East were to disrupt oil supplies, what would happen to the price of oil on world markets? Not surprisingly, elasticity plays a crucial role in answering this question.

As you can see in Table 3, the short-run elasticity of demand for oil is about -0.15 . Since a political or military crisis is usually a short-run phenomenon, the short-run elasticity is what we are interested in. But for this problem, we need to use elasticity in a new way. Remember that elasticity tells us the percentage decrease in quantity demanded for a 1-percent increase in price. But suppose we flip the elasticity fraction upside down, to get $1/E_D = \% \Delta P / \% \Delta Q^D$. This number—the inverse of elasticity—tells us the percentage rise in price that would bring about each 1-percent decrease in quantity demanded. For oil, this number is $1/-0.15 = -6.67$. What does this number mean? It tells us that to bring about each 1-percent decrease in world oil demand, oil prices would have to rise by 6.67 percent.

Now we can make reasonable forecasts about the impact of various events on oil prices. Imagine, for example, an event that temporarily removed half of the Middle East’s oil from world markets. And let’s assume a worst-case scenario: No other nation increases its production during the time frame being considered. What would happen to world oil prices?

Since the Middle East produces about 20 percent of the world’s oil, a reduction by half would decrease world oil supplies by 10 percent. It would then require a price increase of $10 \times 6.67 = 66.7$ percent to restore equilibrium to the market. If oil were initially selling at \$20 per barrel, we could forecast the price to rise by $\$20 \times 0.667 = \13.34 per barrel, for a final price of \$33.34.

Why is it so important to forecast the price of oil that might result from a crisis? If you were a heavy industrial user of oil, you would know the answer. But the forecast is also of immense value to government economists, who would use it to help answer *other* questions. These would include macroeconomic questions, such as, How would a \$13.34 per barrel rise in the price of oil affect the U.S. inflation rate? and microeconomic questions, such as, How would a \$13.34 rise in the price of oil affect the number of flights offered by U.S. airlines, and how would it affect the prices they would charge travelers?

What Happens When
Things Change?



Taxes Once Again. Armed with the concept of price elasticity of demand, we can return to an issue discussed earlier in this chapter—an excise tax. Earlier, you learned that such a tax is partly paid by sellers and partly paid by buyers of a good. Now you will learn a general rule that determines how these tax payments are distributed between sellers and buyers.

To see how this works, let’s review our earlier example of the market for international air travel. Look back at Figure 4 (p. 87). Recall that after the tax was imposed, the new equilibrium was determined at point *B* where the gross supply curve—the one that includes the tax of \$100 per ticket—intersected the demand curve. Travelers ended up paying \$800 for each ticket—\$70 more than before. And the airlines ended up getting \$700 for each ticket, \$30 less than before. Thus, although the tax is formally *collected* from the airlines, it is *paid* by both the airlines (\$70 out of each \$100) and their customers (\$30 out of each \$100).

Will the tax always be divided up this way—70 percent paid by consumers and 30 percent by the airlines? No. It depends in large part on the elasticity of demand for airline travel. In our example, the elasticity of demand is approximately -1.3 .

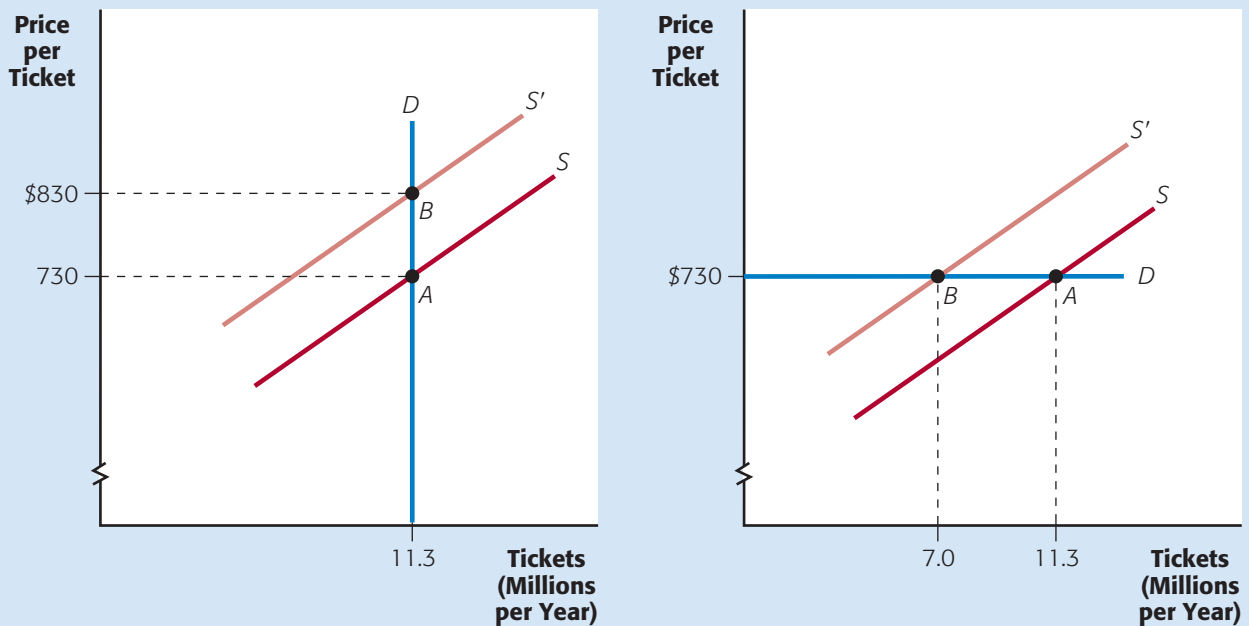
Let's see what would happen if the supply curve remained the same but demand were less elastic. Specifically, what if it was perfectly inelastic?

Panel (a) of Figure 10 shows a perfectly inelastic demand curve along with the same supply curve as in Figure 5. Imposing an excise tax of \$100 per ticket will shift the supply curve upward to S' . We can compare the initial equilibrium at point A with the new equilibrium at point B . Consumers will bear the entire burden of the tax. At B , consumers pay \$830 per ticket—\$100 more than they paid at point A . Firms receive \$830 per ticket, pay \$100 to the government, and end up with the same net price—\$730—as before.

At the other extreme, suppose that the demand curve was perfectly elastic. Once again, the pre-tax equilibrium is at point A , where 11.3 million tickets are sold at a price of \$730 each. And as before, imposition of a \$100 per ticket excise tax shifts the supply curve up vertically by \$100 to S' . The new equilibrium is at point B . Fewer tickets are purchased than before, but this time, the per-unit price paid by consumers remains \$730. It doesn't change at all. The airlines, however, must still pay \$100 per ticket to the government. Thus, the airlines' *net* price—what they get to keep—is only \$630. In this case, all of the excise tax is paid by the airlines.

ELASTICITY AND THE EXCISE TAX ON INTERNATIONAL AIR TRAVEL

FIGURE 10



Who pays an excise tax? The answer depends in part on the elasticity of demand. In panel (a), the demand curve is perfectly inelastic. An excise tax of \$100 per unit shifts the supply curve upward by \$100—from S to S' . In the new equilibrium at point B , the market price is \$100 higher, but the quantity is unchanged. Sellers receive \$830 per unit, pay \$100 to the government, and retain \$730—the same as before. Consumers, however, pay \$100 more per unit for the same number of units they were buying before the tax was imposed. With perfectly inelastic demand, consumers bear the entire burden of the tax.

Panel (b) shows the opposite extreme—the case of perfectly elastic demand. In this case, at point B , consumers end up paying the same price as before the tax was imposed, although they purchase fewer units. Firms receive \$730 per unit but must pay \$100 to the government, so their net revenue is \$630 per unit. Firms end up with \$100 less than before the tax, so they bear the entire burden.

Panels (a) and (b) show extreme cases, in which the entire tax is paid by buyers or by sellers. In most cases, the tax will be shared, as in panel (a). But the extreme cases lead us to a general rule about how a tax will be shared:

*The more elastic the demand curve, the more of an excise tax is paid by sellers.
The more inelastic the demand curve, the more of the tax is paid by buyers.*

OTHER DEMAND ELASTICITIES

In Chapter 3, we saw that other variables besides price influence quantity demanded. We can measure the sensitivity of demand to each of these variables by defining *other* types of demand elasticities. In general, the term *elasticity* measures the percentage change in one variable caused by a 1-percent change in some other variable. But whereas the price elasticity told us about relative movements *along* the demand curve, these other elasticities give us information about how the demand curve *shifts*.

INCOME ELASTICITY OF DEMAND

Recall from Chapter 3 that a change in average household income in a market will shift the demand curve. An *income elasticity* tells us how sensitive demand is to changes in *income*. More specifically,

Income elasticity of demand The percentage change in quantity demanded caused by a 1-percent change in income.

the income elasticity of demand is the percentage change in quantity demanded divided by the percentage change in income, with all other influences on demand remaining constant.

$$E_I = \frac{\% \Delta Q^D}{\% \Delta I}$$

where I is income in the market. More simply, we can interpret this number as *the percentage increase in quantity demanded for each 1-percent rise in income*. For example, if the income elasticity of demand for a certain good is 1.4, then a 1-percent rise in income will increase demand for the good by 1.4 percent, a 2-percent rise in income will increase demand by 2.8 percent, and so on.

Income elasticities and price elasticities of demand differ in several respects. First, a price elasticity of demand measures the effect of changes in the *price* of the good and assumes that other influences on demand, including income, remain unchanged. An income elasticity does just the reverse: It measures the effect on demand we would observe if income changed and all other influences on demand—including the price of the good—remained the same. In other words, instead of letting price vary and holding income constant, now we are letting income vary and holding price constant.

This leads to another difference between price and income elasticities of demand: A price elasticity measures the sensitivity of demand to price as we *move along the demand curve* from one point to another. An income elasticity, by contrast, tells us the relative *shift* in the demand curve—the increase in quantity demanded *at a given price*.

Finally, while a price elasticity is virtually always negative, an income elasticity can be positive or negative. This is because an increase in income will increase the demand for some goods and decrease the demand for others. If you look at the income elasticities in Table 5, you will see both positive and negative numbers.

TABLE 5

Good or Service	Income Elasticity	Good or Service	Income Elasticity	SOME INCOME ELASTICITIES
<i>Narrow Categories</i>		<i>Broad Categories</i>		
Fresh Fruit	1.99	Imports	2.73	
Computers	1.71			
Transatlantic Air Travel	1.40	Transportation	1.79	
College Education	0.55			
Cigarettes	0.50	Recreation	1.07	
Chicken	0.42	Clothing	1.02	
Pork	0.34	Food	0.60 to 0.85	
Fresh Vegetables	0.26			
Tooth Extraction	−0.13 to 0.47			
Ground Beef	−0.20			
Bread	−0.42			
Potatoes	−0.81			

Sources: Erik Brynjolfsson, "Some Estimates of the Contribution of Information Technology to Consumer Welfare," MIT Sloan School, Working Paper #161, Revised, January 1994. Trisha Bezmen and Craig A. Depken, II, "School Characteristics and the Demand for College," *Economics of Education Review*, Vol. 17, No. 2, 1998. F. Chaloupka, "Rational Addictive Behavior and Cigarette Smoking," *Journal of Political Economy*, August 1991, pp. 722–742. J. M. Cigliano, "Price and Income Elasticities for Airline Travel," *Business Economics*, September 1980, pp. 17–21. M. D. Chinn, "Beware of Econometricians Bearing Estimates," *Journal of Policy Analysis and Management*, Fall 1991, pp. 546–557. Willard G. Manning, Jr., and Charles E. Phelps, "The Demand for Dental Care," *Bell Journal of Economics*, Autumn 1979. Dale M. Heien, "The Structure of Food Demand: Interrelatedness and Duality," *American Journal of Agricultural Economics*, May 1982, pp. 213–221. M. R. Baye, D. W. Jansen, and Jae-Woo Lee, "Advertising Effects in Complete Demand Systems," *Applied Economics*, October 1992, pp. 1087–1096. Gary W. Brester and Michael K. Wohlgenant, "Estimating Interrelated Demands for Meats Using New Measures for Ground and Table Cut Beef," *American Journal of Agricultural Economics*, November 1991, pp. 1182–1194.

In Chapter 3, you learned that an increase in income will increase the demand for *normal goods*. These goods have a *positive* income elasticity of demand. When we define goods by broad categories—food, housing, clothing, entertainment, energy, transportation—income elasticity is always positive because an increase in income will always increase demand in each of these categories, even if it decreases spending on particular goods *within* the category. For example, a rise in income may enable you to afford better-quality clothing—so you will buy more high-quality items, and fewer low-quality items—but you almost certainly will end up buying *more clothing* in general. But even when we narrow our definition to specific goods and services—books, CDs, chicken, fresh vegetables, automobiles, and trips to Europe—income elasticities are usually positive. In Table 5, the first six goods have positive income elasticities, as do all of the broad categories.

In Chapter 3, however, you also learned that some goods are *inferior*—demand decreases when income rises. These goods will have a negative income elasticity. While the broad category *travel* is a normal good, and the narrower category *airline travel* is also normal, *bus travel* is an inferior good in many markets. As household income rises, travelers are likely to shift from cheaper (but often less pleasant) bus travel to more expensive (and more pleasant) car and airline travel. Similarly, while food is normal—as are steak, fresh fruit, and sushi—potatoes and ground beef are inferior. As income rises, many households will shift from these cheaper sources of calories to more expensive items. (Why do some studies show that tooth extraction is an inferior good? *Hint*: What are the substitutes for tooth extraction? How much do they cost?)

Normal goods can be further divided into two categories. An **economic necessity** has an income elasticity between zero and one. Since income elasticity is defined

Economic necessity A good with an income elasticity of demand between 0 and 1.

as $\% \Delta Q^D / \% \Delta I$, you can see that when $0 < E_I < 1$, we must have $\% \Delta Q^D < \% \Delta I$. For an economic necessity, a given percentage increase in income causes a *smaller* percentage increase in quantity demanded. The broad category of food is certainly an economic necessity: A 10-percent rise in income will cause the quantity of food demanded to rise, but by less than 10 percent. In fact, using the lower estimate in Table 5, ($E_I = 0.60$), a 10-percent rise in income would increase the demand for food by only 6 percent.

Economic luxury A good with an income elasticity of demand greater than 1.

Goods whose income elasticity is greater than 1.0 are called **economic luxuries**. From the definition of income elasticity, if $E_I > 1$, we must have $\% \Delta Q^D > \% \Delta I$. Thus, when income rises, the quantity demanded of these items will increase by a greater percentage than the rise in income. For example, transportation is an economic luxury: Using the income elasticity in Table 5 ($E_I = 1.79$), we see that a 10-percent rise in income will increase quantity of transportation demanded by about 18 percent.

An interesting implication follows from these definitions: As income rises, the proportion of income spent on economic necessities will fall, while the proportion of income spent on economic luxuries will rise. To see this more clearly, consider Table 6, which shows what would happen to a family's spending on two goods—food and transportation—if its income were to double again and again. We'll use the income-elasticity estimates from Table 5: $E_I = 0.60$ for food, and $E_I = 1.8$ for transportation.

In the table, food is an economic necessity ($E_I < 1$), so that each time income doubles, spending on food increases, but by less than 100 percent. Transportation, by contrast, is an economic luxury ($E_I > 1$), so that each time income doubles, spending on transportation more than doubles. Notice how the percentage of income spent on food continues to fall, while that spent on transportation continues to rise.

To some extent, our definitions of economic necessities and economic luxuries correspond to the more common notions of necessity and luxury. In common speech, a necessity is something that people need. Food, medical care, and housing—each of which people need—also have income elasticities that are less than 1.0. A luxury is considered something desirable but not really necessary. Most of us would regard restaurant meals, opera tickets, trips to Paris, and certainly yachts and caviar as luxuries, and, indeed, each of these items has an income elasticity greater than 1.0.

But it is important to remember that economic necessities and luxuries are categorized by actual consumer behavior and *not* by our judgment of a good's importance to human survival. People can certainly survive without cigarettes. But since cigarettes have an income elasticity between 0 and 1, they are categorized as an economic necessity. Similarly, some of us might think of a computer as a necessity in our lives, and yet—because studies show that the income elasticity of spending on computers is greater than 1.0—we categorize it as an economic luxury.

TABLE 6

INCOME AND SPENDING ON ECONOMIC NECESSITIES AND ECONOMIC LUXURIES	Percent of Income Spent on Food		Percent of Income Spent on Transportation		
	Income	Spending on Food	Spending on Transportation	Spending on Transportation	
	\$10,000	\$ 6,000	60%	\$ 1,000	10%
	\$20,000	\$ 9,600	48%	\$ 2,800	14%
	\$40,000	\$15,360	38%	\$ 7,840	20%
	\$80,000	\$24,576	30%	\$21,952	27%

CROSS-PRICE ELASTICITY OF DEMAND

A cross-price elasticity relates the change in quantity demanded for one good to a price change in another. More formally, we define the **cross-price elasticity of demand** between good X and good Y as:

$$E_{x,y} = \frac{\%Q_x^D}{\%\Delta P_y}$$

In words,

A cross-price elasticity of demand tells us the percentage change in quantity demanded of a good for each 1-percent increase in the price of some other good, all other influences on demand remaining unchanged.

For example, look at the cross-price elasticities reported in Table 7. The cross-price elasticity of Pepsi with the price of Coke is 0.8. This means that when the price of Coke rises by 10 percent, the quantity of Pepsi demanded increases by 8 percent, *all other influences on demand remaining unchanged*. Among the other influences that are assumed to remain unchanged are the price of the good itself (Pepsi), the prices of all related goods *except* Coke, and household income in the market.

As you can see in the table, a cross-price elasticity can be positive or negative, and the sign gives us valuable information about the relationship between the two goods. If $E_{x,y} < 0$, an increase in the price of good Y causes a decrease in quantity demanded for good X. As we know from Chapter 3, this means that goods X and Y are complements. For example, in Table 7, the cross-price elasticity between entertainment and food is negative: A 1-percent rise in the price of food causes a 0.7-percent decrease in the quantity of entertainment demanded. Entertainment and food are complements. This is not surprising: Many forms of entertainment—throwing a party, having a picnic in a state park, or even seeing a movie—are accompanied by spending on food. In the same way, the cross-price elasticity between automobiles and gasoline should be negative: A rise in the price of automobiles will

Cross-price elasticity of demand

The percentage change in the quantity demanded of one good caused by a 1-percent change in the price of another good.

TABLE 7

Products	Cross-Price Elasticity
Margarine with price of butter	1.53
Pepsi with price of Coke	0.80
Coke with price of Pepsi	0.61
Ground beef with price of beef table cuts	0.41
Ground beef with price of poultry	0.24
Electricity with price of natural gas	0.20
Theater with price of all other lively arts	0.12
Entertainment with price of food	-0.72

SOME CROSS-PRICE ELASTICITIES

Sources: F. Gasmí, J. J. Laffont, and Q. Vuong, "Econometric Analysis of Collusive Behavior in a Soft-Drink Market," *Journal of Economics and Management Strategy*, Summer 1992, pp. 277–311. Dale M. Heien, "The Structure of Food Demand: Interrelatedness and Duality," *American Journal of Agricultural Economics*, May 1982, pp. 213–221. Gary W. Brester and Michael K. Wohlgenant, "Estimating Interrelated Demands for Meats Using New Measures for Ground and Table Cut Beef," *American Journal of Agricultural Economics*, November 1991, pp. 1182–1194. E. T. Fuji et al., "An Almost Ideal Demand System for Visitor Expenditures," *Journal of Transport Economics and Policy*, May 1985. C. Hsiao and D. Mountain, "Estimating the Short-Run Income Elasticity of Demand for Electricity by Using Cross-Sectional Categorized Data," *Journal of the American Statistical Association*, June 1985, pp. 259–265.

decrease the quantity of gasoline demanded, especially in the longer run. Similarly, the cross-price elasticities between bread and butter, computers and Internet service, or sunblock lotion and trashy novels are negative: A rise in the price of one item in the pair should decrease the quantity demanded of the other.

If $E_{x,y} > 0$, an increase in the price of good Y causes a decrease in quantity demanded for good X . In this case, goods X and Y are *substitutes*. Most of the cross-price elasticities in Table 7 are positive, indicating that most of the pairs of goods are substitutes rather than complements. For example, the table tells us that margarine and butter are substitutes as are ground beef and poultry.

While the *sign* of the cross-price elasticity helps us distinguish substitutes and complements among related goods, its *size* tells us how *closely* the two goods are related. A large absolute value for $E_{x,y}$ suggests that the two goods are close substitutes or complements, while a small value suggests a weaker relationship.

Butter and margarine seem to be very close substitutes—even closer than Pepsi and Coke. A 10-percent rise in the price of butter will increase the quantity of margarine demanded by about 15 percent. This makes sense, since either good can be substituted for the other in most recipes. While electricity and natural gas are substitutes, they are more distant substitutes than butter and margarine. This, too, makes sense: Natural gas and electricity are exchangeable only in certain uses, and even then, only when the proper equipment is available.

USING THE THEORY: THE STORY OF TWO MARKETS

Using the THEORY



Char-
acterize
the
Market



Identify
Goals
and
Con-
straints



THE MARKET FOR FOOD

Price floors are infrequent in market economies, with one glaring exception: markets for agricultural goods. Almost every government in the world has, at one time or another, experimented with price floors to help keep food prices high. And many governments—including the U.S. government—still have them. What is so special about agricultural markets? Why do governments intervene there so often? What would happen if they did *not* intervene?

Agricultural markets have a rare combination of features affecting supply and demand. The best way to understand these features is to consider the market for food *as a whole*, rather than the market for one particular crop. Why? Because the market forces that affect one type of food product tend to affect virtually *all* food products at the same time. When we combine food products into one category, let's see how these market forces operate on all food products together. In this market, households—constrained by their limited incomes and the market price—choose how much food to buy in order to maximize their well-being. Sellers—constrained by their production technology, the prices of inputs, and the price they can get for food—strive to maximize profit.

What are the unique forces that affect the market for food? First, we find that the *supply* of food is subject to:

1. significant technological advance in the long run, and
2. extreme sensitivity to weather in the short run.

At the same time, the *demand* for food is characterized by another pair of forces:

3. a very low price elasticity, and
4. a very low income elasticity.

To see how farmers are plagued by these features in the long run, let's see how each affects the supply and demand curves. Property (1) tells us that the supply curve for

food tends to shift rightward over time. To see this, think of the effects on output as farmers have shifted from hand plows to horse-drawn plows to tractors. Each of these innovations has caused a significant decrease in the cost per unit of virtually every kind of food, and farmers have been able to produce more food at any given price.

Over the past 50 years, mechanization in farming has led to steady rightward shifts in the supply curve of food. Because of these technological changes, food production has grown much faster than population. That is very good news for the human race, but bad news for the average farmer. When the supply curve for food shifts rightward, the equilibrium price will fall.

In order to prevent the price of food from falling, the demand curve would have to shift rightward by the same distance as the supply curve. We have already pointed out, however, that population growth—one cause of a rightward-shifting demand curve—has fallen short of food-production growth. Aside from population, the only other change that could shift the demand curve continuously rightward would be growth in the average income of the population. But farmers can expect no rescue here either. Notice property (4). Food is characterized by a very low income elasticity; a 1-percent increase in income causes less than a 1-percent increase in the demand for food. Thus, income growth has only weak effects on the demand for food. Since technological change in farming has outpaced both the growth in population and the effects of rising incomes on the demand for food, the shifts in demand have not been able to keep up with the shifts in supply.

Since, in the long run, the rightward shifts in the supply curve seem to outpace the small shifts in the demand curve, let's simplify things by imagining that the demand curve does not shift at all. This is a close approximation to the long-run situation in agricultural markets, and it will enable us to see how elasticities shed light on the problem.

In Figure 11(a), we see that the market is initially in equilibrium at point A. When the supply curve shifts rightward, we move along the demand curve for food, from point A to point B. What happens to the total revenue of farmers as we make this move? You already have the tools to answer this question: Demand for food is price inelastic, as stated in property (3). Therefore, when price decreases, total expenditure on food (the total revenue of farmers) will *fall*. Now we see the ultimate effect of technological progress on farmers: As long as rapid technological progress continues (and as we enter the era of biotechnology, there is every reason to think it will accelerate) the farm sector is doomed to ever-decreasing total revenue. Since new technologies often require large-scale production, only the largest farms will enjoy a decrease in production costs. The result—for the typical small and medium-sized farmer—is a squeeze on profits.

Now let's turn to the short run. Here, the problem is that crop harvests depend heavily on weather patterns, which are very unstable from year to year. If there is good weather, production will be high and the supply of food will be large. As shown in Figure 11(b), the supply curve shifts rightward. If there is bad weather, the supply of food will be much lower, and the supply curve shifts leftward. You can see that under good weather, we have a lower equilibrium price, but higher equilibrium quantity, of food. With bad weather, the equilibrium price is higher, but the equilibrium quantity is lower.

Since the demand for food is price inelastic, a leftward shift in the supply curve, which raises the price of food, will also increase total expenditure on food. So shifts in the supply curve have an ironic effect: Total revenue of farmers is greater when the weather is bad! As long as total costs of production do not differ too greatly under good and bad weather, farm profits will also be higher under bad weather. (This is why farmers actually hope for bad weather—it will create a scarcity of food,



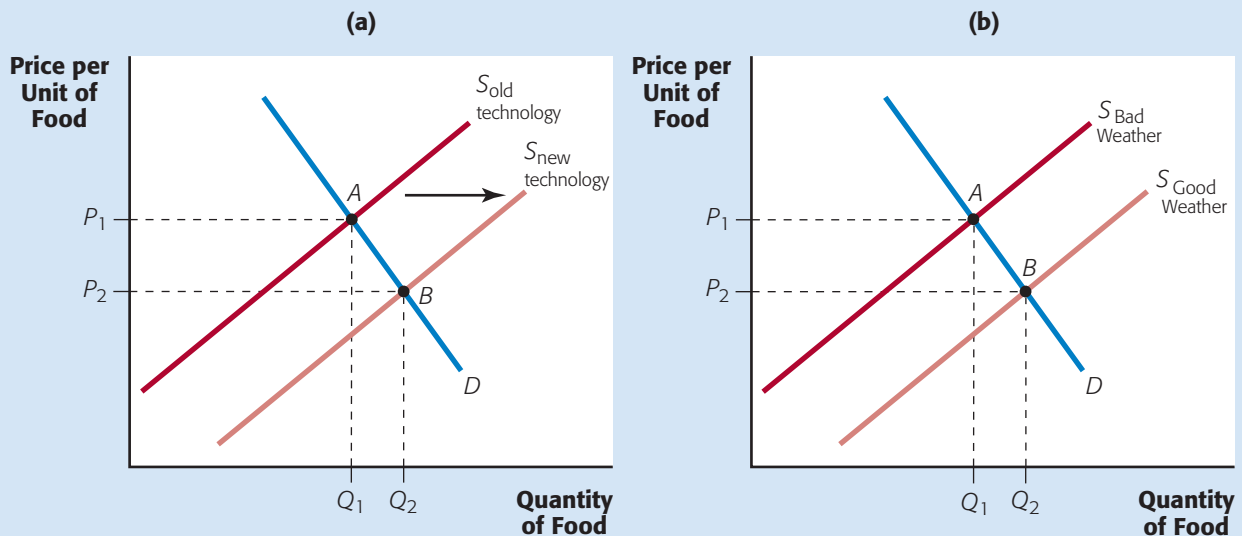
Find the Equilibrium



What Happens When Things Change?

FIGURE 11

THE MARKET FOR FOOD



Over the typical range of prices, the demand for food is inelastic. Panel (a) depicts the market over the long run. Over time, technological changes shift the supply curve to the right, lowering both price and total revenue of farmers. Panel (b) represents the market in the short run. If the weather is good, supply will be greater and price will be lower than when the weather is bad. Farmers' total revenue (and their incomes) are lower when the weather is good.

which will drive the price up and increase their profits—even with lower crop yields.) You can see, then, that farm profits—which depend on unstable and unpredictable weather patterns—will be highly unstable themselves.

Table 8 shows the impact that changing weather has had on the U.S. winter wheat crop and the fate of winter wheat farmers. Notice the large variations in prices and quantities. As winter weather changed from good to bad to worse, the supply curve shifted leftward: Production fell from 1993 to 1994, and again from 1994 to 1995. Notice also that each fall in output was associated with a rise in price. The price rose by about 12 percent from 1993 to 1994 and a whopping 28 percent from 1994 to 1995. Finally, you can see that each price rise is associated with an increase in the total sales revenue of winter wheat farmers. From its low in 1993 to its high in 1995, the value of sales increased by 26 percent.

The trend over these three years was positive for farmers. But it can just as easily turn negative. In mid-1999, for example, the winter wheat crop was expected to

TABLE 8

U.S. WINTER WHEAT PRODUCTION

	1993	1994	1995
Bushels produced	1.75 billion	1.67 billion	1.54 billion
Average price per bushel	\$3.03	\$3.40	\$4.35
Total value of sales	\$5.30 billion	\$5.67 billion	\$6.70 billion

Source: U.S. Department of Agriculture, *Marketing Year Average Prices and Value of Production, by States and United States, 1993, 1994, and 1995*. Production figures calculated by authors as total value of sales divided by average price.

grow by 9 percent above the previous year—from 1.74 to 1.9 billion bushels. As a result, the price was expected to fall dramatically, from \$3.70 in 1998 to \$2.72 in 1999. Farmers were preparing for a very bad year.

Shrinking and unstable incomes are clearly problems for farmers, but are they problems for society? Under ordinary circumstances, the answer would be no. If there is something inherent in farming that makes it a risky or unrewarding type of work, then we would expect farm employment and farm production to shrink. This shrinkage would raise the price of food until those remaining on farms found the job attractive enough to stay. In other words, left to its own devices, the market for farm goods would reach an *equilibrium*—just like any other market.

But farming seems to be special. If the market were allowed to function on its own, the first to leave would be small family farmers, who could not compete against the large conglomerates, which are more technologically advanced and can bear the risk of unstable revenue more easily. And here lies the problem: The notion of the small family farm has tremendous political appeal. In addition, farmers have banded together to form powerful and effective government lobbies, to make sure that agricultural markets do not have to go through the same painful process of adjustment as other markets in the economy. The result has been continual government interference with supply and demand in agricultural markets around the world. As you saw earlier in this chapter, such interference causes problems of its own—the price we pay to protect our farmers from the harsh realities of the market.

HEALTH INSURANCE AND THE MARKET FOR HEALTH CARE

In 1990, health-related expenditures in the United States amounted to about \$700 billion, or 12.2 percent of Gross Domestic Product (GDP). By the end of the decade, the figure had risen to about \$1.3 trillion, or 14.3 percent of GDP. With this rapid rise in spending, the U.S. found itself devoting a larger share of its resources to health care than any other nation in the world. Why such rapid growth—with no end in sight? A variety of explanations have been offered.

On the supply side, scientific breakthroughs have made it possible to treat diseases and conditions that only a few years ago would have proved fatal. These technological changes enable us to live longer, but also raise the cost of keeping a person healthy. On the demand side, as U.S. society ages, it is only natural that spending on health care will increase. After all, as individuals become older, they require more frequent visits to the doctor, have more operations, and may eventually need geriatric care.

Both of these reasons—our longer lives and the use of more expensive types of services—contribute to the rise in health care spending. But there is another reason as well—health insurance.

In the United States most—but by no means all—citizens have some form of health insurance. For the elderly and some of the poor, the insurance is provided by the federal and state governments through the Medicare and Medicaid programs. For others, health insurance comes as a fringe benefit provided by employers. Many of these insurance policies have a special feature called *coinsurance* that involves a sharing of costs between the consumer/patient and the insurance provider. With 30 percent coinsurance, for example, the patient would pay 30 percent of a physician's or hospital bill, and the insurance company would pay the remaining 70 percent.

Let's look at the market for a specific type of health care—annual physical examinations in a large urban area. In this market, buyers—limited by their incomes and the price they must pay for physical exams—interact with physicians



Characterize the Market



Identify Goals and Constraints

Find the Equilibrium



What Happens When Things Change?



and HMOs, who are striving for maximum profit. In the absence of health insurance, there would be a demand for physical exams represented by demand curve D in Figure 12. There would also be a supply of exams provided by doctors and represented by supply curve S . At point A , the two curves intersect to determine an equilibrium price of \$50 per examination and an equilibrium quantity of 100,000 examinations per year.

Now let's examine the effects of health insurance with a 50 percent coinsurance rate. Since consumers now pay only half the cost of any health services they utilize, the effect is to rotate their demand curve upward to D' .

To understand why the demand curve rotates in this way, we just need to reinterpret the demand curve in a way analogous to our reinterpretation of the supply curve earlier, in Figure 3. Usually we think of the demand curve as telling us the quantity buyers will buy at each price. But it also tells us the maximum price that buyers can be charged and still have them buy a given quantity. For example, using the original demand curve D , we see that to get consumers to buy 100,000 examinations, the most they could be charged would be \$100 per exam. If the price is any higher than \$100, people will buy fewer than 100,000 examinations.

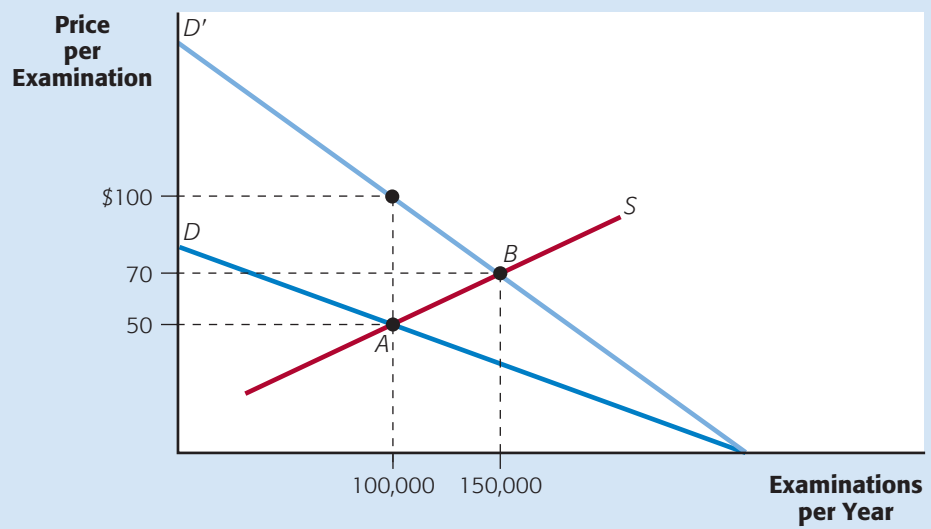
Now, once consumers have a 50 percent coinsurance rate, they would be willing to pay twice as much as before and still buy any given quantity. This is why the demand curve rotates from D to D' . At any quantity along the horizontal axis, the corresponding price along demand curve D' is twice the price along D . For example, at a quantity of 100,000 examinations, the price along D' is \$100 per exam, but half of that would be paid by insurance companies leaving the remaining \$50 to be paid by the consumer.

Once consumers are insured, the new market equilibrium is determined at point B where supply curve S and demand curve D' cross. The effect of insurance is to increase both the quantity of health care provided and the price per unit. In our example, total expenditure increases from $\$50 \times 100,000$ or \$5,000,000, to $\$70 \times 150,000$ or \$10,500,000.

FIGURE 12

At point A , the market for physical examinations is in equilibrium with 100,000 exams provided each year at \$50 each. The introduction of health insurance with a 50 percent coinsurance rate causes the demand curve to rotate upward from D to D' . In the new equilibrium at point B , more exams are provided at a higher price, so that total expenditure on physical examinations has increased.

THE MARKET FOR HEALTH CARE WITH COINSURANCE



The actual amount by which expenditure increases depends in large part on the shapes of demand curve D and supply curve S and on the effective coinsurance rate. For given demand and supply curves, the higher the coinsurance rate, the greater the increase in health care spending. You can see this by imagining the effect of a 10-percent coinsurance rate (which is much closer to reality than our hypothetical 50-percent rate was). With 10-percent coinsurance, each individual pays only one-tenth of the cost of a unit of health care. In that case, the demand curve D' will be much steeper than the curve shown in Figure 12—so much steeper that at any quantity, the price along curve D' could be 10 times as high as the price along D .

Health insurance has definite benefits to our society. Since most people are risk averse, they feel better off and more secure when they are insured. And since some operations and therapies can cost \$100,000 or more, insurance coverage against catastrophic illness can mean the difference between prosperity and bankruptcy for some families. On the other hand, our current health insurance system keeps patients from facing the full opportunity costs of their health care decisions. This can cause people to *overconsume* health care. And some economists believe that health insurance encourages the development of high-cost, low-benefit technologies. At a minimum, health insurance reduces buyers' incentives to monitor their health care expenditures very closely or to shop around for high-quality, low-cost care. Concerns such as these lie at the heart of current debates about the nature of our health care system.

S U M M A R Y

The model of supply and demand is a powerful tool for understanding all sorts of economic events. For example, governments often intervene in markets—either by creating *price ceilings* or *price floors*, or by imposing taxes or subsidies. Supply and demand enables us to predict how these interventions affect the price of a good and the quantity exchanged.

Another powerful tool is the *price elasticity of demand*, defined as the percentage change in quantity demanded divided by the percentage change in price that caused it. In general, price elasticity of demand varies along a demand curve. In the special case of a straight-line demand curve, demand becomes less and less elastic as we move downward and rightward along the curve. Along an elastic portion of any demand curve, a rise in price causes sellers' revenues and consumers' expenditures to fall. Along an *inelastic* portion of any demand curve, a rise in price causes sellers' revenues and consumers' expenditures to increase. Generally speaking, demand for a

good tends to be more elastic: the more narrowly the good is defined, the easier it is to find substitutes for the good, and the greater the share of households' budgets that is spent on the good. And *long-run-elasticities* are almost always larger—in absolute value—than *short-run elasticities*.

The *income elasticity of demand* is the percentage change in quantity demanded divided by the percentage change in income that causes it. For *normal goods*, the income elasticity of demand is positive. For *inferior goods*, the income elasticity is negative.

The *cross-price elasticity of demand* measures the percentage change in the quantity demanded of one good as a result of a given percentage change in the price of some other good. If the cross-price elasticity is positive, we say that the two goods are *substitutes*. If the elasticity is negative, the two goods are said to be *complements*.

K E Y T E R M S

price ceiling
short side of the market
black market
rent controls
price floor

excise tax
price elasticity of demand
inelastic demand
perfectly inelastic demand
elastic demand

perfectly (infinitely) elastic demand
unitary elastic demand
short-run elasticity
long-run elasticity

income elasticity of demand
economic necessity
economic luxury
cross-price elasticity of demand

R E V I E W Q U E S T I O N S

- What is the difference between the price elasticity of demand along a demand curve and the slope of that demand curve?
- Price elasticity of demand is defined at $\% \Delta Q / \% \Delta P$.
 - What formulas do economists use to calculate $\% \Delta Q$ and $\% \Delta P$?
 - Why do economists use these specific formulas?
 - Suppose that the price elasticity of demand for a good is -0.4 . Explain precisely what that means.
- For each of the following pairs of goods or services, indicate which good you would expect to have the *smaller* (in absolute value) price elasticity of demand. In each case, explain why.
 - Exxon gasoline; gasoline in general
 - Beauticians' services; plumbers' services
 - Automobiles; color photocopies
 - Coach-class airfare; business-class airfare
- Give some examples of goods for which demand would be almost perfectly *inelastic*. Then give some examples of goods with almost perfect *elastic* demands. In each case, justify your answers.
- What is the relationship between the price elasticity of demand for a good and total expenditure on that good? Explain how this relationship arises.
- Are short-run price elasticities of demand generally larger or smaller (in absolute value) than long-run elasticities? Why is this so?
- What factors determine the size of the price elasticity of demand for a good? Specifically, how does each factor influence elasticity?
- Which of the following goods are likely to be normal goods? Which are likely to be inferior goods? Defend your answers.
 - Canned spaghetti
 - Vacuum cleaners
 - Used books
 - Computer software
- How are the words *necessity* and *luxury* used differently in economics than in everyday speech?
- For each of the following pairs of goods, would you expect the cross-elasticity of demand to be positive or negative? Large (in absolute value) or small? Defend your answers.
 - Computer hardware and computer software
 - Antibiotics and over-the-counter decongestants
 - Gasoline and automobile repairs

P R O B L E M S A N D E X E R C I S E S

- The market for rice has the following supply and demand schedules:
- The demand for bottled water in a small town is as follows:

P (per ton)	Q_d (tons)	Q_s (tons)
\$10	100	0
\$20	80	30
\$30	60	40
\$40	50	50
\$50	40	60

P (per bottle)	Q_d (bottles per week)
\$1.00	500
\$1.50	400
\$2.00	300
\$2.50	200
\$3.00	100

To support rice producers, the government imposes a price floor of \$50 per ton.

- What quantity will be traded in the market? Why?
 - What price will prevail in the market after the price floor is imposed?
 - What other steps will the government have to take to enforce the floor price?
- Is this a straight-line demand curve? How do you know?
 - Calculate the price elasticity of demand for bottled water for a price rise from \$1.00 to \$1.50. Is demand elastic or inelastic for this price change?
 - Calculate the price elasticity of demand for a price rise from \$2.50 to \$3.00. Is demand elastic or inelastic for this price change?

- d. According to the chapter, demand should become less and less elastic as we move downward and rightward along a demand curve. Use your answers in *b.* and *c.* to confirm this relationship.
- e. Create another column for total expenditure on bottled water at each price.
- f. According to the chapter, a rise in price should *increase* total expenditure on bottled water when demand is inelastic, and *decrease* total expenditure when demand is elastic. Use your answers in *b.* and *c.* above, and the new total expenditure column you created, to confirm this.
3. Refer to Table 3 on p. 99 and answer the following questions:
- Is the demand for recreation more or less elastic than the demand for clothing?
 - If 10,000 two-liter bottles of Pepsi are currently being demanded in your community each month, and the price increases from \$0.90 to \$1.00 per bottle, what will happen to quantity demanded? Be specific.
 - By how much would the price of ground beef have to increase (in percentage terms) in order to reduce quantity demanded by 5 percent?
4. From the information in the following table, calculate the income elasticity of demand for this good if income increases from \$10,000 to \$20,000, and if income increases from \$40,000 to \$50,000. (All the quantities were measured at a price of \$10 per unit.)

Income	Quantity Demanded
\$10,000	50
\$20,000	60
\$30,000	70
\$40,000	80
\$50,000	90

C H A L L E N G E Q U E S T I O N S


- A subsidy is the opposite of a tax—it is a payment *from* the government rather than *to* the government. Suppose that the government institutes a subsidy of \$10 for every new computer installed in the United States. The money would be paid to the buyer of the computer. Use a supply and demand diagram to show the effect of this subsidy on the price and quantity of computers.
- As discussed in Chapters 3 and 4, price acts as an allocation mechanism, determining how the quantity produced is distributed among those who wish to consume the good. In the case of a price ceiling, the allocative mechanism is frustrated.

Consider the market for rental housing. If a rent ceiling is set below the market price:

 - Will there be a shortage or a surplus of rental housing?
 - Since price can no longer allocate rental housing, what other mechanism might emerge? What factors might be used to determine who gets rent-controlled apartments and who does not?

- Is this a normal or an inferior good? How can you tell?
 - Does the proportion of household income spent on this good increase or decrease as income increases?
 - Is this good considered an economic luxury, an economic necessity, or neither? Why?
5. Use the data in Table 7 on p. 109 to answer the following questions:
- If the price of entertainment increases by 2 percent, what will happen to the quantity of food demanded? Be specific.
 - If the price of electricity falls by 3 percent, what will happen to the quantity of natural gas demanded? Again, be specific.
 - If a shift in tastes increases the demand for poultry and drives up its price by 5 percent, what will happen to the quantity of ground beef demanded?
6. Three Guys Named Al, a moving company, is contemplating a price hike. Currently, they charge \$20 per hour, but Al thinks they could get \$30. Al disagrees, saying it will hurt the business. Al, the brains of the outfit, has calculated the price elasticity of demand for their moving services in the range from \$20 to \$30 and found it to be -0.5 .
- Should they do as Al suggests and raise the price? Why or why not?
 - Currently, Three Guys is the only moving company in town. Al reads in the paper that several new movers are planning to set up shop there within the next year. Twelve months from now, is the demand for Three Guys' services likely to be more elastic, less elastic, or the same? Why?

EXPERIENTIAL EXERCISES

1. Policy.com has a Web page devoted to tobacco-related issues. Review the articles at the site (<http://www.policy.com/issues/issue213.html>). Choose one and interpret the analysis using supply and demand curves and the concept of elasticity. For additional information, check *The Tobacco Wars* by Walter Adams and James Brock (Cincinnati, OH: South-Western College Publishing Co., 1999).
- 
2. Cross-price elasticities are important in the computer industry. Read the “Personal Technology” column in Thursday’s *Wall Street Journal* and find a story that describes a hardware or software product. Make a list of *other* products that you think would be substitutes (positive cross-price elasticity) for the product in the article. Arrange the items in your list from very close substitutes (very large cross-price elasticity) to more distant substitutes (smaller cross-price elasticity). For each item on the list, make your best guess about the numerical value of the elasticity. Using the guess, what would happen if the price of each item on your list rose by 10 percent? When you’ve finished, follow the same steps for a list of *complements* to the product in question. In this case, the cross-price elasticities will be negative.

CONSUMER CHOICE

CHAPTER

5

You are constantly making economic decisions. Some of them are rather trivial. (Should you buy the expensive coffee at Starbucks or make it more cheaply at home?) Others can have a profound impact on the way you live. The economic nature of all these decisions is rather obvious, since they all involve *spending*.

But in other cases, the economic nature of your decisions may be less obvious. Did you get up early today in order to get things done, or did you sleep in? Which leisure activities—movies, concerts, sports, hobbies—do you engage in, and how often do you decline an opportunity to have fun for lack of time? At this very moment, what have you decided *not* to do in order to make time to read this chapter? All of these are economic choices, too, because they require you to allocate a scarce resource—your *time*—among different alternatives.

To understand the economic choices that individuals make, we must know what they are trying to achieve (their goals) and the limitations they face in achieving them (their constraints). This should sound familiar: it is Key Step #2 of our four-step procedure. In this chapter, we focus on Key Step #2 as it pertains to individual households—the economic decision makers who buy goods and services and who make decisions about work and play.

But wait. How can we identify the goals and constraints of *consumers* when we are all so *different* from each other?

Indeed, we *are* different from one another . . . when it comes to *specific* goals and *specific* constraints. But at the highest level of generality, we are all very much alike. All of us, for example, would like to maximize our overall level of *satisfaction*. And all of us, as we attempt to satisfy our desires, come up against the same constraints: too little income or wealth to buy everything we might enjoy, and too little time to enjoy it all.

We'll start our analysis of individual choice with constraints, and then move on to goals. In most of the chapter, we will focus on choices about *spending*: how people decide what to buy. This is why the theory of individual decision making is often called “consumer theory.” Later, in the Using the Theory section, we'll see how the theory can be broadened to include decisions about allocating scarce *time* among different activities.

CHAPTER OUTLINE

The Budget Constraint

Changes in the Budget Line

The Consumer's Goal

Utility and Marginal Utility

Preferences

Rationality

Preferences and Marginal Utility

Consumer Decision Making

What Happens When Things Change?

Changes in Income

Changes in Price

The Individual's Demand Curve

Consumers in Markets

From Individual to Market

Demand

Challenges to Consumer Theory

Using the Theory: Improving Education

Appendix: Consumer Theory with Indifference Curves

The Indifference Map

The Marginal Rate of Substitution

Consumer Decision Making

Indifference Curves and the Individual Demand Curve

Identify Goals and Constraints



THE BUDGET CONSTRAINT

Virtually all individuals must face two facts of economic life: (1) they have to pay prices for the goods and services they buy, and (2) they have limited funds to spend. These two facts are summarized by the consumer's *budget constraint*:

A consumer's budget constraint identifies which combinations of goods and services the consumer can afford with a limited budget, at given prices.

Budget constraint The different combinations of goods a consumer can afford with a limited budget, at given prices.

Consider Max, a devoted fan of both movies and concerts, who has a total budget of \$150 to spend on both each month. For each movie, Max must pay a direct money cost of \$10 (the ticket price plus the cost of transportation), and for each concert, a direct money cost of \$30. If Max were to spend all of his \$150 budget on concerts at \$30 each, he could see at most five each month. If he were to spend it all on movies at \$10 each, he could see 15 of them.

But Max could also choose to spend *part* of his budget on concerts and *part* on movies. In this case, for each number of concerts, there is some *maximum* number of movies that he could see. For example, if he goes to one concert per month, it will cost him \$30 of his \$150 budget, leaving \$120 available for movies. Thus, if Max were to choose one concert, the *maximum* number of films he could choose would be $\$120/\$10 = 12$.

Figure 1 lists—for each number of concerts—the maximum number of movies that Max could see. Each combination of goods in the table is affordable for Max, since each will cost him exactly \$150. Combination *A*, at one extreme, represents no concerts and 15 movies. Combination *F*, the other extreme, represents 5 concerts and no movies. In each of the combinations between *A* and *F*, Max attends both concerts and movies.

The graph in Figure 1 plots the number of movies along the vertical axis and the number of concerts along the horizontal. Each of the points *A* through *F* corresponds to one of the combinations in the table. If we connect all of these points with a straight line, we have a graphical representation of Max's budget constraint, which we call Max's **budget line**.

Budget line The graphical representation of a budget constraint.

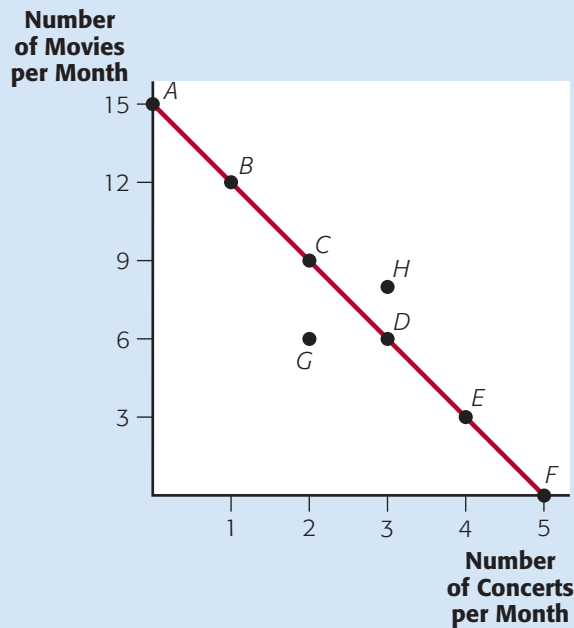
Note that any point below or to the left of the budget line is affordable. For example, 2 concerts and 6 movies—indicated by point *G*—would cost only $\$60 + \$60 = \$120$. Max could certainly afford this combination. On the other hand, he *cannot* afford any combination *above* and to the right of this line. Point *H*, representing 3 concerts and 8 movies, would cost $\$90 + \$80 = \$170$, which is beyond Max's budget. The budget line therefore serves as a *border* between those combinations that are affordable and those that are not.

Let's look at Max's budget line more closely. The *vertical intercept* is 15, the number of movies Max could see if he attended zero concerts. Starting at the vertical intercept (point *A*), notice that each time Max increases one unit along the horizontal axis (attends one more concert), he must decrease 3 units along the vertical axis (see three fewer movies). Thus, the slope of the budget line is equal to -3 . The slope tells us Max's *opportunity cost* of one more concert. That is, the opportunity cost of 1 more concert is 3 movies foregone.

Relative price The price of one good relative to the price of another.

There is an important relationship between the *prices* of two goods and the opportunity cost of having more of one or the other. The prices Max faces tell us how many dollars he must give up to get another unit of each good. If, however, we divide one money price by another money price, we get what is called a **relative price**—

THE BUDGET CONSTRAINT



Max's Consumption Possibilities with Income of \$150

	Concerts at \$30 each		Movies at \$10 each	
	Quantity	Total Expenditure on Concerts	Quantity	Total Expenditure on Movies
A	0	\$ 0	15	\$150
B	1	\$ 30	12	\$120
C	2	\$ 60	9	\$ 90
D	3	\$ 90	6	\$ 60
E	4	\$120	3	\$ 30
F	5	\$150	0	\$ 0

FIGURE 1

The budget line shows all combinations of concerts and movies Max could attend by spending \$150 each month. At point *A*, he could attend 15 movies, but no concerts. At *F*, he could attend 5 concerts but no movies. At points *B–E*, he attends both movies *and* concerts. The slope of the line ($-P_{\text{concert}}/P_{\text{movie}} = -3$) shows that the opportunity cost of another concert is 3 movies.

the price of one good *relative* to the other. Since $P_{\text{concert}} = \$30$ and $P_{\text{movie}} = \$10$, the *relative price of a concert* is the ratio $P_{\text{concert}}/P_{\text{movie}} = \$30/\$10 = 3$. Notice that 3 is the opportunity cost of another concert in terms of movies, and—except for the minus sign—it is also the slope of the budget line. That is, *the relative price of a concert, the opportunity cost of another concert, and the slope of the budget line* have the same absolute value. This is one example of a general relationship:

The slope of the budget line indicates the spending trade-off between one good and another—the amount of one good that must be sacrificed in order to buy more of another good. If P_y is the price of the good on the vertical axis and P_x is the price of the good on the horizontal axis, then the slope of the budget line is $-P_x/P_y$.



It's tempting to think that the slope of the budget line should be $-P_y/P_x$, where the price of the vertical-axis good, P_y , is in the numerator rather than in the denominator. But this is wrong. The budget line's slope is the change in *quantity* along the vertical axis divided by the change in *quantity* along the horizontal. As our example shows, when the slope is expressed in terms of *prices* rather than quantities, the formula is $-P_x/P_y$, with the price of the *horizontal-axis* good in the numerator.

CHANGES IN THE BUDGET LINE

To draw the budget line in Figure 1, we have assumed given prices for movies and concerts, and a given income that Max can spend on them. These “givens”—the prices of the goods and the consumer's income—are always *assumed*

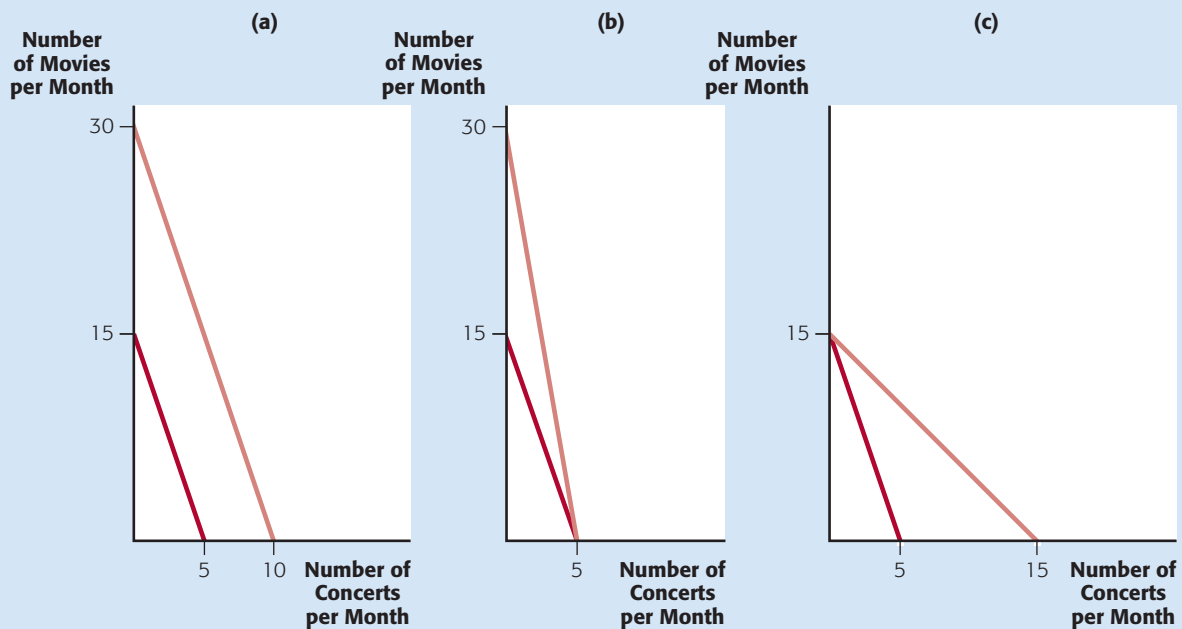
constant as we move along a budget line; if any one of them changes, the budget line will change as well. Let's see how.

Changes in Income. If Max's available income increases from \$150 to \$300 per month, then he can afford to see more movies, more concerts, or more of both, as shown by the change in his budget line in Figure 2(a). If Max were to devote *all* of his income to movies, he could now see 30 of them each month, instead of the 15 he was able to see before. Devoting his entire income to concerts would enable him to attend 10, rather than 5. Moreover, for any number of concerts, he will be able to see more movies than before. For example, before, when his budget was only \$150, choosing 2 concerts would allow Max to see only 9 movies. Now, with a budget of \$300, he can have 2 concerts and 24 movies.

Notice that the old and new budget lines in Figure 2(a) are parallel—they have the same slope of -3 . This is because we changed Max's income but *not* prices.

FIGURE 2

CHANGES IN THE BUDGET LINE



In panel (a), an increase in income leads to a rightward, parallel shift of the budget line. In panel (b), a decrease in the price of a movie causes the budget line to rotate upward; the horizontal intercept is unaffected. In panel (c), a decrease in the price of a concert leads to a rightward rotation of the budget line.

Since the ratio $P_{\text{concert}}/P_{\text{movie}}$ has not changed, the spending trade-off between movies and concerts remains the same. Thus,

An increase in income will shift the budget line upward (and rightward). A decrease in income will shift the budget line downward (and leftward). These shifts are parallel—changes in income do not affect the budget line's slope.

Changes in Price. Now let's go back to Max's original budget of \$150 and explore what happens to the budget line when a price changes. Suppose the price of a movie falls from \$10 to \$5. The graph in Figure 2(b) shows Max's old and new budget lines. When the price of a movie falls, the budget line rotates outward—the vertical intercept moves higher. The reason is this: When a movie costs \$10, Max could spend his entire \$150 on them and see 15; now that they cost \$5, he can see a maximum of 30. The horizontal intercept—representing how many concerts Max could see with his entire income—doesn't change at all, since there has been no change in the price of a concert. Notice that the new budget line is also *steeper* than the original one, with slope equal to $-P_{\text{concert}}/P_{\text{movie}} = -\$30/\$5 = -6$. Now, with movies costing \$5, the trade-off between movies and concerts is 6 to 1, instead of 3 to 1.

Panel (c) of Figure 2 illustrates another price change. This time, it's a fall in the price of a *concert* from \$30 to \$10. Once again, the budget line rotates, but now it is the horizontal (concerts) intercept that changes and the vertical (movies) intercept that remains fixed.

We could draw similar diagrams illustrating a *rise* in the price of a movie or a concert, but you should try to do this on your own. In each case, one of the budget line's intercepts will change, as well as its slope:

When the price of a good changes, the budget line rotates: Both its slope and one of its intercepts will change.

The budget constraint, as illustrated by the budget line, is one side of the story of consumer choice. It indicates the trade-off consumers *are able to* make between one good and another. But just as important is the trade-off that consumers *want to* make between one good and another, and this depends on consumers' *preferences*, the subject of the next section.

THE CONSUMER'S GOAL

Economists assume that *any* decision maker—a consumer, the manager of a business firm, or officials in a government agency—tries to make the *best* out of any situation. More specifically, we assume that consumers (the subject of this chapter) strive to maximize their **utility**—a quantitative measure of their well-being or satisfaction. Anything that makes the consumer better off is assumed to raise his utility. Anything that makes the consumer worse off will decrease his utility.

Are you troubled by this assumption? Many people are when they first encounter it in an economics course. One common objection is that it is unrealistic. It seems to imply that we are all engaged in a relentless, conscious pursuit of narrow goals—an implication contradicted by much of human behavior. As you read this paragraph, are you *consciously* trying to maximize your own well-being? Are you fully aware that reading this will improve your grade in economics and that, in turn,



The Bureau of Labor Statistics' Consumer Expenditure Survey will give you a snapshot picture of the consumption behavior of typical U.S. households (<http://stats.bls.gov/news.release/cesan.toc.htm>).



Identify Goals and Constraints

Utility Pleasure or satisfaction obtained from consuming goods and services.

will help you achieve other important goals? Perhaps. But more likely, you aren't thinking about any of this. You are reading this chapter right now because . . . well, because you *should*. In truth, we only rarely make decisions with conscious, hard calculations, and are more often guided by feelings that we may or may not be aware of. Why, then, do economists assume that people make decisions consciously and quantitatively when, in reality, they often don't?

This is an important question. Economists answer it this way: The ultimate purpose of building an economic model is to *understand and predict behavior*—the behavior of households, firms, government, and the overall economy. As long as people behave *as if* they are maximizing something, then we can build a good model by *assuming that they are*. Whether they *actually, consciously* maximize anything is an interesting philosophical question, but the answer doesn't affect the usefulness of the model.

Milton Friedman, Nobel-prize winning economist, put it this way:

Consider the problem of predicting the shots made by an expert billiard player. It seems not at all unreasonable that excellent predictions would be yielded by the hypothesis that the billiard player made his shots as if he knew the complicated mathematical formulas that would give the optimum directions of travel, could estimate accurately by eye the angles, etc., describing the location of the balls, could make lightning calculations from the formulas, and could then make the balls travel in the direction indicated by the formulas. Our confidence in this hypothesis is not based on the belief that billiard players, even expert ones, can or do go through the process described; it derives rather from the belief that, unless in some way or other they were capable of reaching essentially the same result, they would not in fact be expert billiard players.

Keep this in mind as we delve further into consumer theory. The consumer is assumed to maximize his or her *utility*. This does *not* mean that we really believe consumers consult some kind of utility meter every time they make a purchase or decide how to allocate their time. But it does mean that, for the most part, consumers behave *as if* they consult such a meter.

A second common objection to the assumption that people maximize utility is that it is narrow minded, focusing on people's selfishness, and ignoring their nobler motives. Indeed, in this chapter, we will assume that an individual's utility is assumed to increase as he gets more and more material things.

Still, utility maximization need not imply that people are selfish or that economists think they are. On the contrary, economists are very interested in cases where people take the interests of others into account. For example, much economic life takes place in the family, where people care a great deal about each other. Utility maximization would then be applied to the family as a whole. That is, we would assume that the *family*, rather than any one individual within it, is trying to make the best out of any situation.

Also remember that an economic model is always built for a specific purpose. Useful economic models have been built to explore charitable giving by individuals and corporations, volunteer activity, and ethical behavior such as honesty, fairness, and respect for fellow citizens. In these models, economics recognizes that people often care about their friends, their neighbors, their coworkers, and the broader society in which they live. Accordingly, these models assume that an individual's utility rises not only when he acquires more things for himself, but also when others are made better off.

But when we are exploring the more common questions of individual behavior in *markets*, there is little to gain by recognizing these nobler motives. After all, the same person who gives generously to charity will usually try to maximize his *private* gain when trading in the stock market or shopping for clothes. This is why, in this chapter, we will assume that each decision maker's utility depends on his *own* acquisitions.

UTILITY AND MARGINAL UTILITY

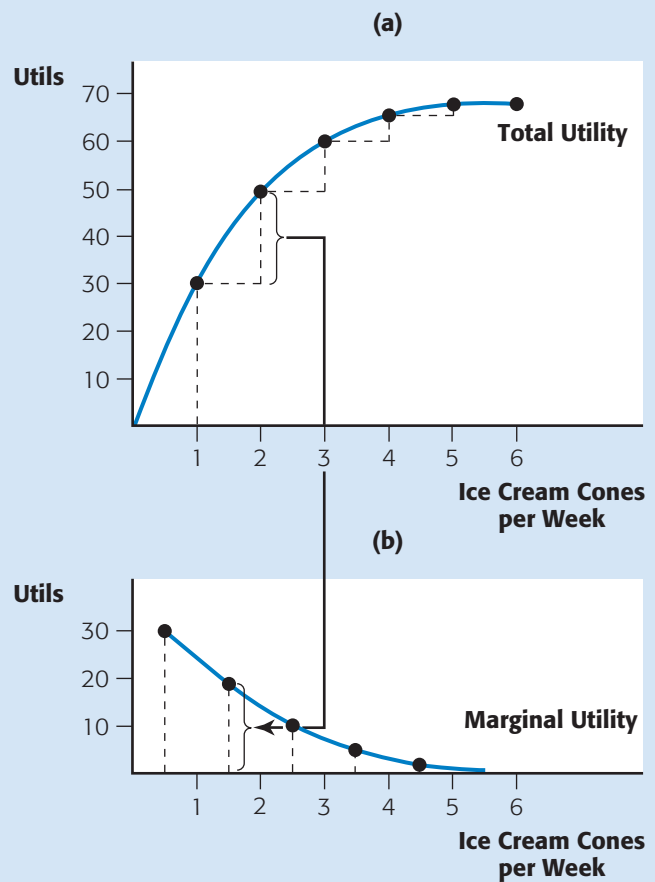
Figure 3 provides a graphical view of utility—in this case, the utility of a consumer named Lisa who likes ice cream cones. Look first at panel (a). On the horizontal axis, we'll measure the number of ice cream cones Lisa consumes each week. On the vertical axis, we'll measure the utility she derives from consuming each of them. If Lisa values ice cream cones, her utility will increase as she acquires more of them, as it does in the figure. There we see that when she has 1 cone, she enjoys total utility

TOTAL AND MARGINAL UTILITY

FIGURE 3

Lisa's Total and Marginal Utility from Consuming Ice Cream Cones

Number of Cones	Total Utility	Marginal Utility
0	0 utils	
1	30 utils	30 utils
2	50 utils	20 utils
3	60 utils	10 utils
4	65 utils	5 utils
5	68 utils	3 utils
6	68 utils	0 utils



Panel (a) shows Lisa's total utility from her consumption of ice cream cones. As her consumption of ice cream rises, so does her total utility. Panel (b) shows the corresponding marginal utility. *MU* falls as ice cream consumption rises, indicating that each additional ice cream cone per week provides less *additional* utility than the previous one did.

of 30 “utils,” but when she has 2 cones, her total utility grows to 50 utils, and so on. Throughout the figure, the total utility Lisa derives from consuming ice cream cones keeps rising as she gets to consume more and more of them.

But notice something interesting—and important: Although Lisa’s utility increases every time she acquires more ice cream, the *additional* utility she derives from each *successive* cone gets smaller and smaller as she gets more cones. We call the *change in utility* derived from consuming an *additional unit* of a good the *marginal utility* of that additional unit:

Marginal utility The change in total utility an individual obtains from consuming an additional unit of a good or service.

Marginal utility is the change in utility an individual enjoys from consuming an additional unit of a good.

What we’ve observed about Lisa’s utility can be restated this way: As she eats more and more ice cream cones in a given week, her *marginal utility* from another cone declines. In the nineteenth and early twentieth centuries, economists thought this pattern was typical of virtually *all* consumers consuming virtually any good or service, and they called it the **law of diminishing marginal utility**. The great economist Alfred Marshall (1842–1924) put it this way:

Law of diminishing marginal utility As consumption of a good or service increases, marginal utility decreases.

The marginal utility of a thing to anyone diminishes with every increase in the amount of it he already has.¹

According to the law of diminishing marginal utility, when you consume your first unit of some good, like an ice cream cone, you derive some amount of utility. When you get your second cone that week, you enjoy greater satisfaction than when you only had one, but the *extra* satisfaction you derive from the second is likely to be smaller than the satisfaction you derived from the first. Adding the third cone to your weekly consumption will no doubt increase your utility further, but again the *marginal utility* you derive from that third cone is likely to be less than the marginal utility you derived from the second. Figure 3 will again help us see what’s going on. The table summarizes the information in the total utility graph. The first two columns show, respectively, the quantity of cones Lisa consumes each week and the total utility she receives each week from consuming them. The third column is new. It shows the marginal utility she receives from each successive cone she consumes per week. As you can see in the table, Lisa’s total utility keeps increasing (marginal utility is always positive) until she consumes 5 cones per week, but the rate at which total utility increases gets smaller and smaller (her marginal utility diminishes) as her consumption increases.

Marginal utility is shown in panel (b) of Figure 3. Because marginal utility is the change in utility caused by a *change* in consumption from one level to another, we plot each marginal utility entry *between* the old and new consumption levels.

Notice the close relationship between the graph of total utility in panel (a) and the corresponding graph of marginal utility in panel (b). If you look closely at the two graphs, and you will see that for every one-unit increment in Lisa’s ice cream consumption her marginal utility is equal to the *change* in her total utility. The downward-sloping curve in panel (b) gives us a vivid illustration of the law of diminishing marginal utility.

¹ *Principles of Economics*, Book III, Ch. III, Appendix notes 1 & 2. Macmillan & Co., 1930.

One last thing about Figure 3: Because marginal utility diminishes for Lisa, by the time she has consumed a total of 5 cones per week, the marginal utility she derives from an additional cone has fallen all the way to zero. At this point, she is fully *satiated* with ice cream and gets no extra satisfaction or utility from eating any more of it in a typical week.

Once this satiation point is reached, even if ice cream were free, Lisa would turn it down (“Yechhh! Not more ice cream!!”).



The word *marginal* is one you will encounter again and again in your study of economics. Literally, a margin is an “edge,” or something *beyond*. In economics, marginal means “additional” or “incremental” and is used to describe what happens when a decision maker considers a small *change* from his current situation.

It is easy to confuse a *total* measure of something with its associated *marginal* measure because they are both measured in the same units. But they are not the same thing. The marginal always tells us the *change in the total* caused by *one more* of something. For example, both total utility and marginal utility are measured in utils. But marginal utility tells us the *change in total utility* when a consumer gets one more unit of a good.

PREFERENCES

In the previous section, we explored how a consumer’s well-being, or utility, changes as she consumes more and more of a *single* good. But ultimately, we want to understand how consumers make *choices* among *different combinations* of goods. As you’ll see, the concept of utility can help us here as well. More specifically, it helps us to characterize people’s *preferences*.

How can we possibly speak systematically about people’s preferences? After all, people are different. They like different things. American teens delight in having a Coke with dinner, while the very idea makes a French person shudder. What would satisfy a Buddhist monk would hardly satisfy the typical American.

And even among “typical Americans,” there is little consensus about tastes. Some read Jane Austen, while others pick John Grisham. Some like to spend their vacations traveling to distant lands, whereas others would prefer to stay home and sleep in every day. Even those who like Häagen-Dazs ice cream can’t agree on which is the best flavor—the company notices consistent, regional differences in consumption. In Los Angeles, chocolate chocolate chip is the clear favorite, while on most of the East Coast, it’s butter pecan (except in New York City, where coffee wins hands down).

In spite of such wide differences in preferences, we can find some important common denominators—things that seem to be true for a wide variety of people. In our theory of consumer choice, we will focus on these common denominators.

RATIONALITY

One common denominator—and a critical assumption behind consumer theory—is that people *have* preferences. More specifically, we assume that you can look at two alternatives and state either that you prefer one to the other or that you are entirely indifferent between the two—you value them equally.

Another common denominator is that preferences are *logically consistent*, or *transitive*. If, for example, you prefer a sports car to a jeep, and a jeep to a motorcycle, then we assume that you will also prefer a sports car to a motorcycle. When a consumer’s preferences are logically consistent in this manner, we say that she has **rational preferences**.

Rational preferences Preferences that satisfy two conditions: (1) Any two alternatives can be compared, and one is preferred or else the two are valued equally, and (2) the comparisons are logically consistent.

Notice that rationality is a matter of how you make your choices, and *not what choices you make*. You can be rational and like apples better than oranges, or oranges better than apples. You can be rational even if you like anchovies or brussels sprouts! What matters is that you make choices consistently, and most of us usually do. Imagine for a moment what it might be like if you didn't. How would you figure out what to order in a restaurant if you prefer the chef's salad to the Reuben sandwich and the Reuben to the hamburger, but prefer the hamburger to the chef's salad! Clearly, choosing consistently is an important part of just being able to choose.

PREFERENCES AND MARGINAL UTILITY

Another feature of preferences that virtually all of us share is this: We generally feel that *more is better*. Specifically, if we get more of some good or service, and nothing else is taken away from us, we will generally feel better off. Since marginal utility measures the change in utility from getting one more unit of a good, we can also state the “more is better” assumption this way: *Marginal utility is positive*.

This condition seems to be satisfied for the vast majority of goods we all consume. Of course, there are exceptions. If you hate eggplant, then the more of it you have, the worse off you are. In this case, the marginal utility of eggplant would be negative, violating the assumption. Similarly, a dieter who says, “Don't bring any ice cream into the house. I don't want to be tempted,” also violates the assumption. The model of consumer choice in this chapter is designed for preferences that satisfy the “more is better” condition, and it would have to be modified to take account of exceptions like these.

In addition to presuming that “more is better,” we'll make one other assumption about people's tastes: The more of a good someone consumes, the less *additional* satisfaction that person will get from consuming still more of it. Here, we are assuming that *marginal utility diminishes as more of a good is consumed*. This is what we assumed for Lisa and her ice cream, and it seems plausible that it would hold for most things that we value. Once again, there may be exceptions. If a fan takes special pride in owning every CD ever recorded by Garth Brooks, then each time she acquires another one, she comes closer to her goal, and her marginal utility might *rise* with each additional CD acquired. But—as with “more is better”—such exceptions are rare.

CONSUMER DECISION MAKING

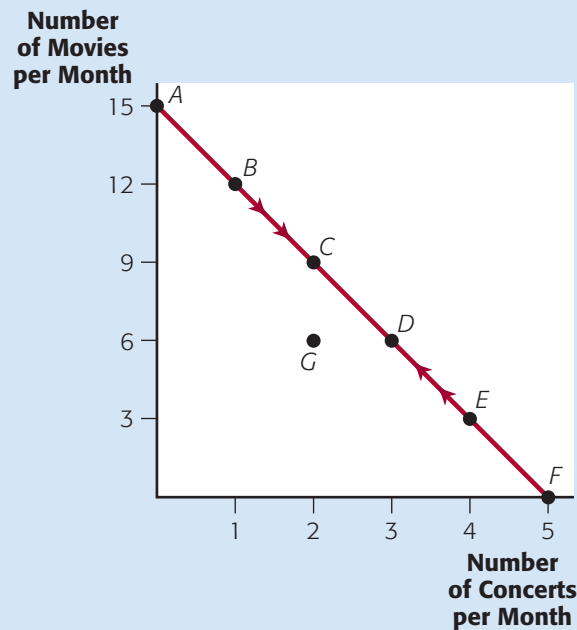
In order to understand demand, we need to bring the consumer's preferences and the consumer's constraints together. But are we really ready? After all, while you've learned quite a bit about the consumer's budget constraint, our characterization of consumer preferences has been rather minimal. We have made only three assumptions: (1) Consumers are rational, (2) the marginal utility of a good is positive, and (3) marginal utility declines as more of the good is consumed. With so little to go on, what can we hope to say about the *choices* a consumer will actually make? Surprisingly, we can say quite a bit.

Our first conclusion about consumer choice is very basic:

The consumer will always choose a point on the budget line, rather than a point below it.

CONSUMER DECISION MAKING

FIGURE 4



CONCERTS at \$30 each

MOVIES at \$10 each

(1) Point on Budget Line	(2) Number of Concerts per Month	(3) Marginal Utility from Last Concert	(4) Marginal Utility per Dollar Spent on Last Concert ($MU_{\text{concerts}} / P_{\text{concerts}}$)	(5) Number of Movies per Month	(6) Marginal Utility from Last Movie	(7) Marginal Utility per Dollar Spent on Last Movie ($MU_{\text{movies}} / P_{\text{movies}}$)
A	0	—	—	15	50	5
B	1	1,500	50	12	100	10
C	2	1,200	40	9	150	15
D	3	600	20	6	200	20
E	4	390	13	3	350	35
F	5	300	10	0	—	—

The budget line shows the maximum number of movies Max could attend for each number of concerts he attends. He would never choose an interior point like G because there are affordable points—on the line—that make him better off. Max will choose a point on the budget line. More specifically, he will choose the point at which the marginal utilities per dollar spent on movies and concerts are equal. From the table, this occurs at point D.

To see why this is so, look at Figure 4. There you'll see Max's budget line, reproduced from Figure 1, where the price of concerts is \$30, the price of movies is \$10, and his monthly budget is \$150. Max would never choose point G, representing 2 concerts and 6 movies, since there are affordable points—on the budget line—that we know make him better off. For example, point C has the same number of concerts (2), but more movies (9). "More is better" tells us that Max will prefer C to

G , so we know G won't be chosen. For the same reason, Max must prefer point D , with 3 concerts and 6 movies, to point G . Indeed, if we look at any point below the budget line, we can always find at least one point *on* the budget line that is preferred, as long as more is better.

Knowing what Max will not do—knowing he *will not* choose a point inside his budget line—is helpful. It tells us that we can narrow our search for the point he *will* choose to just the ones along the budget line AF . But how can Max find the one point along the budget line that gives him a higher utility than all the others?

To answer this question, we'll introduce a concept we'll be coming back to again and again in this text: **marginal decision making**.

Marginal decision making To understand and predict the behavior of individual decision makers, we focus on the incremental or marginal effects of their actions.

To understand and predict the behavior of individual decision makers, we focus on the incremental or marginal effects of their actions.

Marginal decision making can be compared to the children's game in which one child is blindfolded and must find a hidden object. As he moves around, the others tell him only "warmer" or "colder" to indicate whether he is getting closer or farther away from the object. Eventually the child will find the object with only these hints to direct him. In consumer theory, we can think of maximum utility as the hidden object the consumer is looking for, and we imagine him deciding whether some change in his collection of goods makes him better off or worse off—"warmer" or "colder." If he continually makes changes that make him better off, until no such changes are left, then he will discover the combination that makes him as well off as possible.

Marginal decision making is a central concept in economics in general and consumer theory in particular. Before we put it to use, however, a small warning: Taken literally, consumer theory will seem hopelessly unrealistic. "Surely," you may think, "people don't actually *use* concepts like budget lines or marginal utility when they make decisions." And you would be absolutely correct. After all, you've been making economic decisions all your life without even *knowing* about these concepts.

But keep in mind that consumer theory, like many theories in economics, is an "as-if" theory. Economists do not claim that the model of consumer choice describes the psychological mechanics consumers actually use when they make decisions. Rather, they claim that consumers generally choose their goods and services *as if* they follow the model. This is why our highly structured way of looking at decision making—while not a realistic description of *how* people make choices—has proven so useful in explaining the nature of those choices.

With this perspective in mind, let's apply marginal decision making to Max and his choice between movies and concerts. To do this, we need hypothetical information about Max's preferences, which is provided in the table in Figure 4.

Each row of the table corresponds to a different point on Max's budget line. For example, the row labeled C corresponds to point C on the budget line. The second entry in each row tells us the number of concerts that Max attends each month, and the third entry tells us the marginal utility he gets from consuming *the last* concert. For example, at point C , Max attends two concerts, and the second one gives him an additional 1,200 utils beyond the first. Notice that as we move *down* along the budget line, from point A to B to C and so on, the number of concerts increases, and the marginal utility numbers in the table get smaller, consistent with the law of diminishing marginal utility.

The fourth entry in each row shows something new: the *marginal utility per dollar* spent on concerts, obtained by dividing the marginal utility of the last concert by the price of a concert ($MU_{\text{concerts}}/P_{\text{concerts}}$). This tells us the gain in utility Max

gets for each dollar he spends on the last concert. For example, at point C, Max gains 1,200 utils from his second concert during the month, so his marginal utility per dollar spent on that concert is $1,200 \text{ utils}/\$30 = 40 \text{ utils per dollar}$. Marginal utility per dollar, like marginal utility itself, declines as more concerts are consumed.

The last three entries in each row give us similar information for movies: the number of movies attended, the marginal utility derived from the last movie, and the marginal utility per dollar spent on the last movie ($MU_{\text{movies}}/P_{\text{movies}}$). As we travel *up* this column, Max attends more movies, and both marginal utility and marginal utility per dollar decline—once again, consistent with the law of diminishing marginal utility.

To understand how Max can find the best point on his budget line—the one that gives him the highest utility—suppose that he is initially at point B: 1 concert and 12 movies. Is he maximizing his utility? Let's see. Comparing the fourth and seventh entries in row B of the table, we see that Max's marginal utility per dollar spent on concerts is 50 utils, while his marginal utility per dollar spent on movies is only 10 utils. Since he gains more additional utility from each dollar spent on concerts than from each dollar spent on movies, he will have a net gain in utility if he shifts some of his dollars from movies to concerts. To do this, he must travel farther down his budget line.

Next suppose that, after shifting his spending from movies to concerts, Max arrives at point C on his budget line. What should he do then? At point C, Max's *MU* per dollar spent on concerts is 40 utils, while his *MU* per dollar spent on movies is 15 utils. Once again, he would gain utility by shifting from movies to concerts, traveling down his budget line once again.

Now suppose that Max arrives at point D. At this point, the *MU* per dollar spent on both movies and concerts is the same: 20 utils. There is no further gain from shifting spending from movies to concerts. At point D, Max has exploited all opportunities to make himself better off by moving down the budget line. He has maximized his utility.

But wait . . . what if Max had started at a point on his budget line *below* point D, with too many movies and too few concerts? Would he still end up at the same place? Yes, he would. Suppose Max finds himself at point E, with 4 concerts and 3 movies. Here, marginal utilities per dollar are 13 utils for concerts and 35 utils for movies. Now, Max could make himself better off by shifting spending away from concerts and toward movies. He will travel *up* the budget line, once again arriving at point D, where no further move will improve his well-being.

As you can see, whether Max begins at a point on his budget line above point D or below it, marginal decision making will always bring him back to point D. What is so special about point D? It is the only point on the budget line where *marginal utility per dollar* is the same for both goods. When this condition holds, there is nothing to gain by shifting spending in either direction.

What is true for Max and his choice between movies and concerts is true for *any* consumer and *any* two goods. We can generalize our result this way: For any two goods x and y , with prices P_x and P_y , whenever $MU_x/P_x > MU_y/P_y$, a consumer is made better off shifting spending away from y and toward x . When $MU_y/P_y > MU_x/P_x$, a consumer is made better off by shifting spending away from x and toward y . This leads us to an important conclusion:

A utility-maximizing consumer will choose the point on the budget line where marginal utility per dollar is the same for both goods ($MU_x/P_x = MU_y/P_y$). At that point, there is no further gain from reallocating expenditures in either direction.



In finding the utility-maximizing combination of goods for a consumer, why do we use marginal utility *per dollar* instead of just marginal utility?

Shouldn't the consumer always shift spending wherever *marginal utility* is greater? The answer is no. The following thought experiment will help you see why. Imagine that you like to ski and you like going out for dinner.

Further, given your current combination of skiing and dining out, your marginal utility for one more skiing trip is 2,000 utils, and your marginal utility for an additional dinner is 1,000 utils. Should you shift your spending from dining out to skiing? It might seem so, since skiing has the higher marginal utility.

But what if skiing costs \$200 per trip, while a dinner out costs only \$20? Then, while it's true that another skiing trip will give you twice as much utility as another dinner out, it's also true that *skiing costs ten times as much*. You would have to sacrifice *ten* restaurant meals for one skiing trip, and that would make you *worse* off. Instead, you should shift your spending in the other direction: from skiing to dining out. Money spent on additional skiing trips will give you $1,000 \text{ utils}/\$200 = 5$ utils per dollar, while money spent on additional dinners will give you $1,000 \text{ utils}/\$20 = 50$ utils per dollar. Dining out clearly gives you "more bang for the buck" than skiing. The lesson of this example: When trying to find the utility-maximizing combination of goods, compare marginal utilities *per dollar*, not marginal utilities alone.

We can generalize even further. Suppose there are more than two goods an individual can buy. For example, we could imagine that Max wants to divide his entertainment budget among movies, concerts, plays, football games, and what have you. Or we can think of a consumer who must allocate her entire income among thousands of different goods and services each month: different types of food, clothing, entertainment, transportation, and so on. Does our description of the optimal choice for the consumer still hold? Indeed, it does. No matter how many goods there are to choose from, when the consumer

is doing as well as possible, it must be true that $MU_x/P_x = MU_y/P_y$ for any pair of goods x and y . If this condition is *not* satisfied, the consumer will be better off consuming more of one and less of the other good in the pair.²

WHAT HAPPENS WHEN THINGS CHANGE?

If every one of our decisions had to be made only once, life would be much easier. But that's not how life is. Just when you think you've figured out what to do, things change. In a market economy, as you've learned, prices can change for any number of reasons. (See Chapter 3.) A consumer's income can change as well. He may lose a job or find a new one; she may get a raise or a cut in pay. Changes in our incomes or the prices we face cause us to rethink our spending decisions: What maximized utility before the change is unlikely to maximize it afterward. The result is a change in our behavior.

CHANGES IN INCOME

Figure 5 illustrates how an increase in income might affect Max's choice between movies and concerts. As before, we assume that movies cost \$10 each, that concerts cost \$30 each, and that these prices will remain constant. Initially, Max has \$150 in income to spend on the two goods, so his budget line is the line from point A to point F . As we've already seen, under these conditions, Max would choose point D (3 concerts and 6 movies) to maximize utility.

If Max's income increases to \$300, his budget line will shift upward and outward in the figure. How will he respond? As always, he will search along his budget

² There is one exception to this statement: Sometimes the optimal choice is to buy *none* of some good. For example, if $MU_y/P_y > MU_x/P_x$, no matter how small a quantity of good x a person consumes, it will always pay to reduce consumption of good x further, until its quantity is zero. Economists call this a "corner solution," because—when there are only two goods being considered—the individual will locate at one of the endpoints of the budget line in a corner of the diagram.

EFFECTS OF AN INCREASE IN INCOME

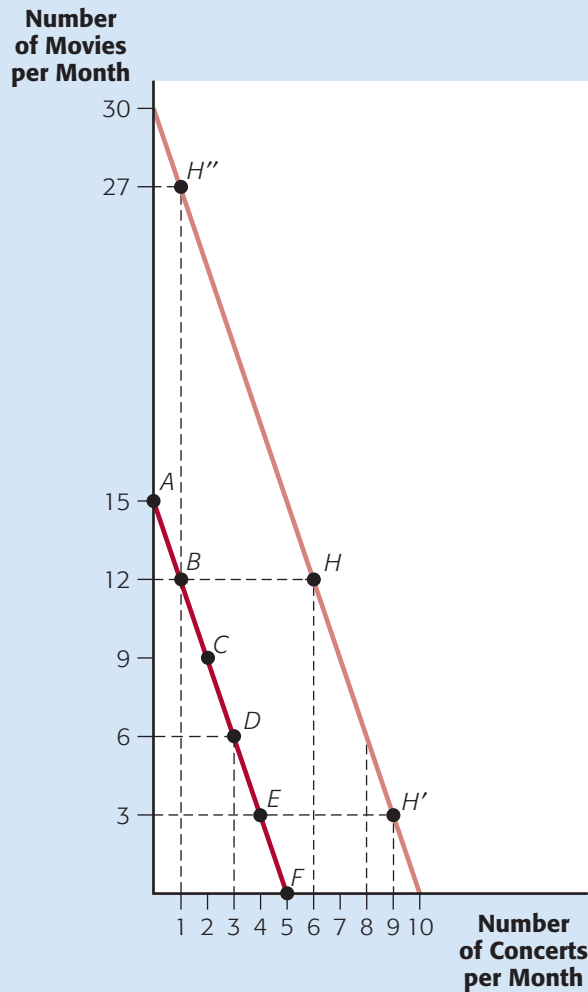


FIGURE 5

A doubling of Max's income causes a parallel, rightward shift of his budget line. More combinations of movies and concerts are now available to him. He will choose the point on the new budget line at which marginal utilities per dollar are equal for the two goods.

line until he finds the point where the marginal utility per dollar spent on both goods is the same. Without more information—such as that provided in the table in Figure 4—we can't be certain which point will satisfy this condition. But we can discuss some of the possibilities.

Figure 5 illustrates three alternative possibilities. If Max's best combination ends up being point *H*, he would attend 12 movies and 6 concerts. If we compare his initial choice (point *D*) with this new choice (point *H*), we see that the rise in income has caused him to consume more of *both* goods. As you learned in Chapter 3, when an increase in income causes a consumer to buy *more* of something, we call that thing a *normal good*. If, for Max, point *H* happens to be where the marginal utilities per dollar for the two goods are equal, then, for him, both movies and concerts are normal goods.

Alternatively, Max's marginal utilities per dollar might be equal at a point like *H'*, with 9 concerts and 3 movies. In this case, the increase in income would cause Max's consumption of concerts to increase (from 3 to 9), but his consumption of



It's tempting to think that *inferior* goods are of lower quality than *normal* goods. But economists don't define normal or inferior based on the intrinsic properties of a good, but rather by the choices people make when their incomes increase. For example, Max may think that both movies and concerts are high-quality goods. When his income is low, he may see movies on most weekends because, being cheaper, they enable him to have some entertainment every weekend. But if his income increases, he can afford to switch from movies to concerts on some of his weekends. If Max makes this choice—and attends fewer movies—then his *behavior* tells us that movies are inferior for him. If instead he chose to see more movies and fewer concerts when his income increased, then concerts would be the inferior good.

movies to *fall* (from 6 to 3). If so, movies would be an *inferior good* for Max—one for which demand decreases when income increases—while concerts would be a *normal good*.

Finally, let's consider another possible outcome for Max: point H' . At this point, he attends more movies and fewer concerts compared to point D . If point H' is where

Max's marginal utilities per dollar are equal after the increase in income, then *concerts* would be the inferior good, and movies would be normal.

CHANGES IN PRICE

In Chapter 3, you were introduced to the *law of demand*, which holds that a rise in the price of a good reduces the quantity demanded, and a fall in price increases quantity demanded. In this section, we use the tools of consumer theory to analyze what is *behind* the law of demand, to see *why* consumers behave as they do when a price changes. In the process, you will learn why exceptions to the law of demand are so rare.

Let's explore what happens to Max when the price of a concert decreases from \$30 to \$10, while his income remains at \$150 and the price of a movie remains \$10. The drop in the price of concerts rotates Max's budget line rightward, pivoting around its vertical intercept, as illustrated in the upper panel of Figure 6. What will Max do after his budget line rotates in this way? Again, he will select the combination of movies and concerts on his budget line that makes him as well off as possible. This will be the combination at which the marginal utility per dollar spent on both goods is the same. In the figure, we assume that this occurs at point J on the new budget line, where Max consumes 4 concerts and 11 movies.

If the price of a concert drops once again, to \$5, the budget line rotates rightward again. In the figure, Max will now choose point K , attending 6 concerts and 12 movies.

THE INDIVIDUAL'S DEMAND CURVE

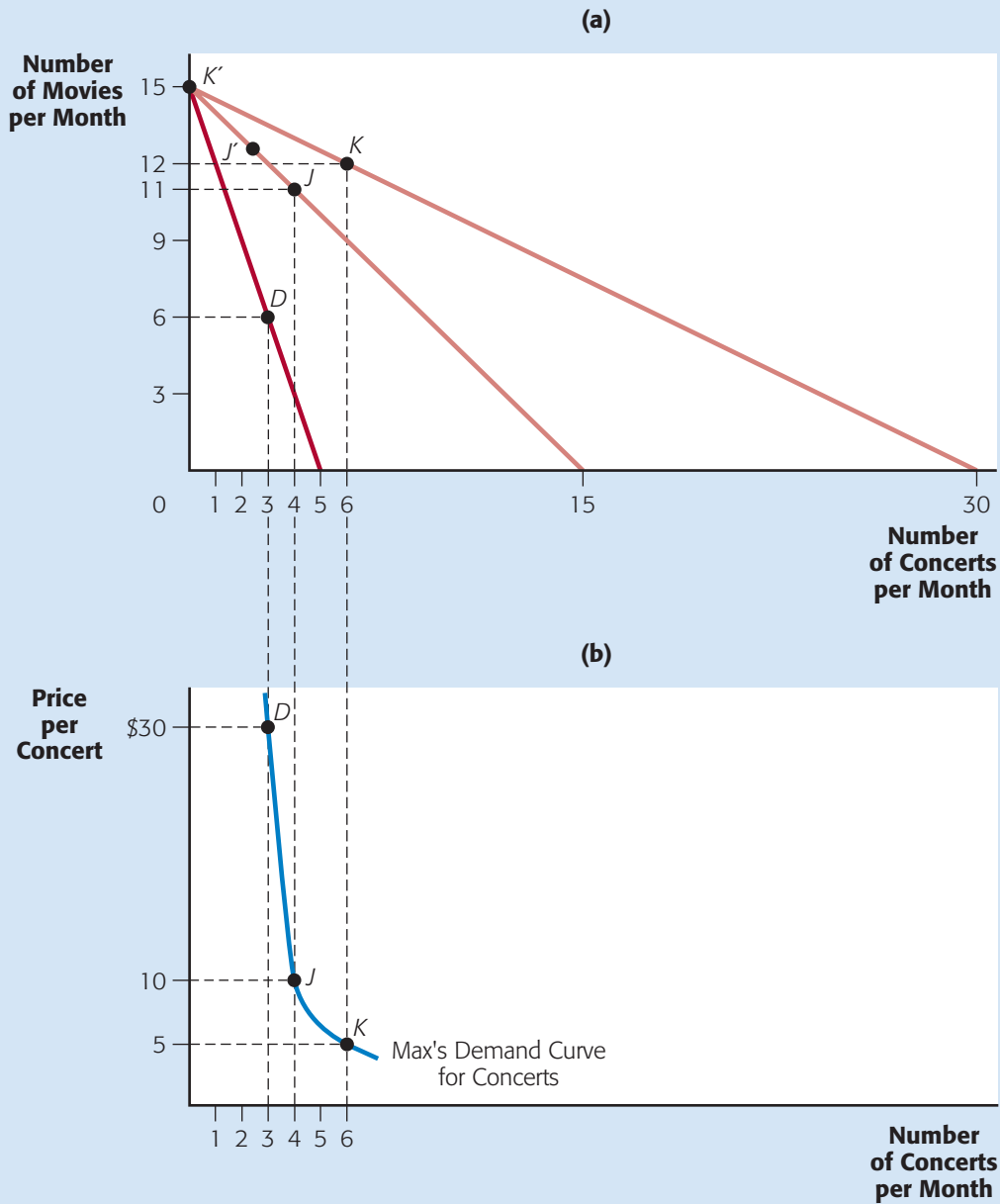
You've just seen that each time the price of concerts changes, so does the quantity of concerts Max will want to see. The lower panel of Figure 6 highlights this relationship by plotting the quantity of concerts demanded on the horizontal axis and the price of concerts on the vertical axis. For example, in both the upper and lower panels, point D tells us that when the price of concerts is \$30, Max will see three of them. When we connect points like D , J , and K in the lower panel, we get Max's **individual demand curve**, which shows *the quantity of a good he demands at each different price*. Notice that Max's demand curve for concerts slopes downward—a fall in the price of concerts increases the quantity demanded—showing that Max's responses to price changes obey the law of demand.

But if Max's preferences had been different, could his response to a price change have violated the law of demand? In particular, could he have chosen points such as J' and K' instead of J and K in panel (a) of Figure 6? If he did, a fall in the price of

Individual demand curve A curve showing the quantity of a good or service demanded by a particular individual at each different price.

DERIVING THE DEMAND CURVE

FIGURE 6



In panel (a), a decrease in the price of concerts causes Max's budget lines to rotate outward. At \$30 per concert, he maximizes utility at point D in both panels and attends 3 concerts. If the price falls to \$10 per concert, he increases his consumption to 4 concerts per month, at point J . At a price of \$5 each, he attends 6 concerts, shown at point K . Max's demand curve in panel (b) is obtained by connecting points such as D , J , and K .

concerts would have led him to want *fewer* of them, and his demand curve (which you are invited to draw for yourself) would have sloped *upward*. Is that possible?

The answer is yes . . . and no. Yes, it is theoretically possible, but no, it does not seem to happen in practice. To understand why, we must look deeper into the effects of a price change on quantity demanded. In doing so, we'll gain more insight into the process of consumer decision making.

The Substitution Effect. When the price of a good changes, we can identify two separate effects on quantity demanded. As you will see, these two effects sometimes work together and sometimes work in opposite directions.

Suppose the price of a good falls. Then it becomes less expensive *relative to* other goods whose prices have not fallen. Some of these other goods are *substitutes* for the now cheaper good—they are different goods, but they are used to satisfy the same general desire. (For example, Coke and Pepsi are very close substitutes for each other, since they both satisfy the same desire for a carbonated cola drink with a little caffeine.) When *one* of the ways of satisfying a desire becomes relatively cheaper, consumers will purchase more of it, and purchase less of the substitute good.

In Max's case, concerts and movies, while different, both satisfy his desire to be entertained. When the price of concerts falls, so does its relative price (relative to movies). Max can now get more entertainment from his budget by substituting concerts in place of movies, so he will demand more concerts.

This impact of a price decrease is called a **substitution effect**—the consumer substitutes *toward* the good whose price has decreased, and away from other goods whose prices have remained unchanged.

Substitution effect As the price of a good falls, the consumer substitutes that good in place of other goods whose prices have not changed.

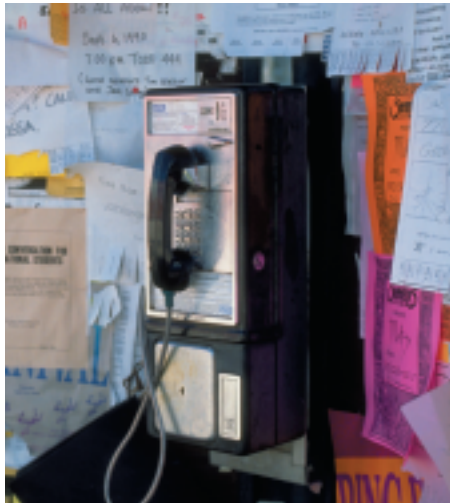
The substitution effect of a price change arises from a change in the relative price of a good, and it always moves quantity demanded in the opposite direction to the price change. When price decreases, the substitution effect works to increase quantity demanded; when price increases, the substitution effect works to decrease quantity demanded.

The substitution effect is a powerful force in the marketplace. For example, while the price of cellular phone calls has fallen in recent years, the price of pay phone calls has remained more or less the same. This fall in the relative price of cell phone calls has caused consumers to substitute toward them and away from using regular pay phones. As a result, many private providers of pay phones are having financial difficulty.

The substitution effect is also important from a theoretical perspective: It is the main factor responsible for the law of demand. Indeed, if the substitution effect were the *only* effect of a price change, the law of demand would be more than a law; it would be a logical necessity. But as we are about to see, a price change has another effect as well.

The Income Effect. In Figure 6, when the price of concerts decreases from \$30 to \$10, Max's budget line rotates rightward. Max now has a wider range of options than before: He can consume more concerts, more movies, or *more of both*. The price decline of *one* good has increased Max's total purchasing power over *both* goods.

A price cut gives the consumer a gift, which is rather like an increase in *income*. Indeed, in an important sense, it *is* an increase in *available* income: Point *D* (3 concerts and 6 movies) originally cost Max \$150, but after the decrease in the price of concerts, the same combination would cost him just $(6 \times \$10) + (3 \times \$10) = \$90$,



Cheaper cell phone calls, and the substitution effect, may soon drive pay phones out of the market.

leaving him with \$60 in *available income* to spend on more movies or concerts or both. This leads to the second effect of a change in price:

*The **income effect** of a price change is the impact on quantity demanded that arises from a change in purchasing power over both goods. A drop in price increases purchasing power, while a rise in price decreases purchasing power.*

Income effect As the price of a good decreases, the consumer's purchasing power increases, causing a change in quantity demanded for the good.

How will a change in purchasing power influence the quantity of a good demanded? That depends. Recall that an increase in income will increase the demand for normal goods and decrease the demand for inferior goods. The same is true for the *income effect* of a price cut: It can work to either *increase* or *decrease* the quantity of a good demanded, depending on whether the good is normal or inferior. For example, if concerts are a normal good for Max, then the income effect of a price cut will lead him to consume more of them; if concerts are inferior, the income effect will lead him to consume fewer.

Combining Substitution and Income Effects. Now let's look again at the impact of a price change, considering the substitution and income effects together. A change in the price of a good changes both the relative price of the good (the substitution effect) and the overall purchasing power of the consumer (the income effect). The ultimate impact of the price change on quantity demanded will depend on *both* of these effects. In most cases, these two effects work together to push quantity demanded in the same direction, but they can occasionally oppose each other. To help clarify this, we'll consider the total impact of a price change on different types of goods.

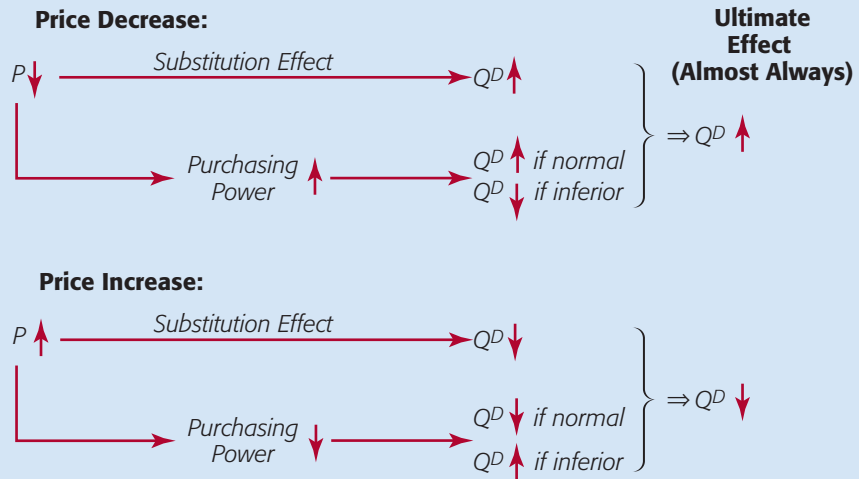
Normal Goods. Normal goods are the easier category to consider. When the price of a normal good falls, the substitution effect *increases* quantity demanded. The price drop will also increase the consumer's purchasing power, and—for a normal good—*increase* quantity demanded even further. The opposite occurs when price increases: The substitution effect decreases quantity demanded, and the decline in purchasing power further decreases it. Figure 7 summarizes how the substitution and income effects combine to make the price and quantity of a normal good move in opposite directions:

For normal goods, the substitution and income effects work together, causing quantity demanded to move in the opposite direction of the price. Normal goods, therefore, must always obey the law of demand.

Inferior Goods. Now let's see how a price change affects the demand for *inferior* goods. As an example, consider ground beef. For many people, ground beef is an inferior good: A rise in income would decrease demand for it, since it would make steak—a preferable alternative—more affordable. If the price of ground beef falls, the substitution effect would work, as always, to *increase* quantity demanded. The price cut will also, as always, increase the consumer's purchasing power. But if ground beef is inferior, the rise in purchasing power will *decrease* quantity demanded. Thus, we have two opposing effects: the substitution effect, increasing quantity demanded, and the income effect, decreasing quantity demanded. In theory, either of these effects could dominate the other, so the quantity demanded could move in either direction. In practice, however, the substitution effect almost always dominates for inferior goods.

FIGURE 7

INCOME AND SUBSTITUTION EFFECTS



Why does the substitution effect almost always dominate? Because we consume such a wide variety of goods and services that a price cut in any one of them changes our purchasing power by only a small amount. For example, suppose you have an income of \$20,000 per year, and you spend \$500 per year on ground beef. If the price of ground beef falls by, say, 20 percent, this would save you \$100—like a gift of \$100 in income. But \$100 is only $\frac{1}{2}$ percent of your income. Thus, a 20 percent fall in the price of ground beef would cause only a $\frac{1}{2}$ percent rise in your purchasing power. Even if ground beef is, for you, an inferior good, we would expect only a tiny decrease in your quantity demanded when your purchasing power changes by such a small amount. Thus, the income effect should be very small. On the other hand, the *substitution* effect should be rather large: With ground beef now 20 percent cheaper, you will likely substitute away from other purchases (such as steak) and buy more ground beef.

For inferior goods, the substitution and income effects of a price change work against each other. The substitution effect moves quantity demanded in the opposite direction of the price, while the income effect moves it in the same direction as the price. But since the substitution effect virtually always dominates, consumption of inferior goods—like normal goods—will virtually always obey the law of demand.

CONSUMERS IN MARKETS

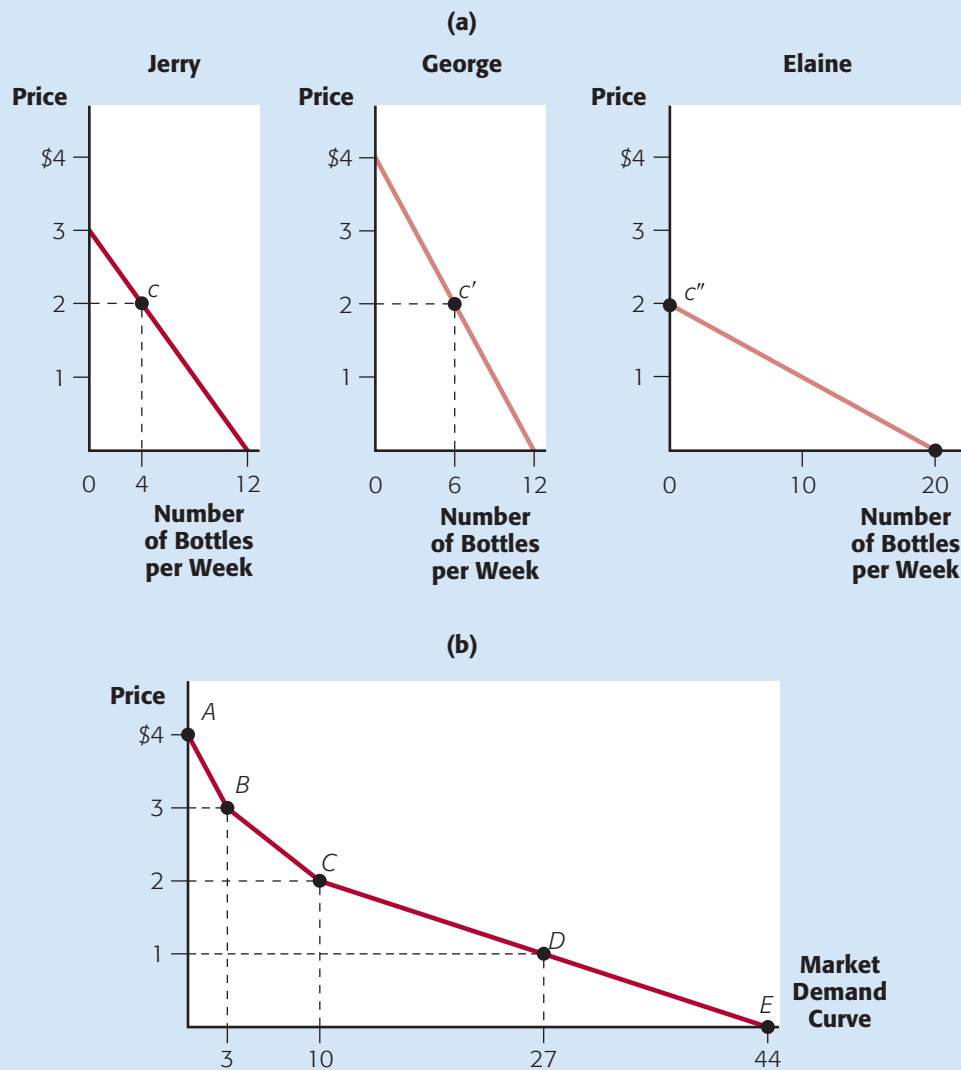
So far, we've looked only at the behavior of an individual consumer. But one of the goals of consumer theory is to explain how large *groups* of individuals react to and are affected by changes in their economic environment. For this purpose, we need the *market demand curve*. In Chapter 3, you learned what a market demand curve is and how it can be used to help determine equilibrium price and quantity in a market. In this section, we revisit the market demand curve to learn where it comes from.

FROM INDIVIDUAL TO MARKET DEMAND

The market demand curve tells us the quantity of a good demanded by all consumers in a market, so it makes sense that we can derive it by adding up the individual demand curves of every consumer in that market. Figure 8 illustrates how this can be done in a small local market for bottled water, where, for simplicity, we assume that there are only three consumers—Jerry, George, and Elaine. The first three diagrams show the individual demand curves. If the market price were, say, \$2 per bottle, Jerry would buy 4 bottles each week (point c), George would buy 6 (point c'), and Elaine would buy zero (point c''). Thus, the market quantity demanded at a price of \$2 would be $4 + 6 + 0 = 10$, which is point C on the market demand curve. To obtain

FROM INDIVIDUAL TO MARKET DEMAND

FIGURE 8



The individual demand curves show how much bottled water will be demanded by Jerry, George, and Elaine at different prices. As the price falls, each demands more. The market demand curve in panel (b) is obtained by adding up the total quantity demanded by all market participants at different prices.

the entire market demand curve, we repeat this procedure at each different price, adding up the quantities demanded by each individual to obtain the total quantity demanded in the market. (Verify on your own that points *A*, *B*, *D*, and *E* have been obtained in the same way.) In effect, we obtain the market demand curve by summing horizontally across each of the individual demand curves:

The market demand curve is found by horizontally summing the individual demand curves of every consumer in the market.

Notice that as long as each individual's demand curve is downward sloping (and this will virtually always be the case), then the market demand curve will also be downward sloping. More directly, if a rise in price makes each consumer buy fewer units, then it will reduce the quantity bought by *all* consumers as well. Indeed, the market demand curve can still obey the law of demand even when *some* individuals violate it. Thus, although we are already quite confident about the law of demand at the individual level, we can be even *more* confident at the market level. This is why we always draw market demand curves with a downward slope.



To get an insight into the economic forces driving a fad, read "Pokemon Economics" at http://www.dismal.com/todays_econ/te_120299.stm

CHALLENGES TO CONSUMER THEORY

In some circumstances, our model of consumer choice will not work well, at least not without some modification. One problem is *uncertainty*. In our model, the consumer knows with certainty the outcome of any choice—so many movies and concerts—and knows with certainty how much income is available for spending. But in many real-world situations, you make your choice and you take your chances. When you buy a car, it might be a lemon; when you pay for some types of surgery, there is a substantial risk that it will be unsuccessful; and when you buy a house, you cannot be sure of its condition or how much you will like the neighborhood. Income, too, is often uncertain. Employees risk being laid off, and self-employed lawyers, doctors, and small-business owners might have a good year or a bad year. When uncertainty is an important aspect of consumer choice, economists use other, more complex models. But even these models are based on the one you have learned in this chapter.

Another problem is *imperfect information*. In our model, consumers are assumed to *know* exactly what goods they are buying and the prices at which they can buy them. But in the real world, we must sometimes spend time and money to get this information. Prices can be different in different stores and on different days, depending on whether there is a sale, so we might have to make phone calls or shop around. To be sure of the quality of our purchases, we may have to subscribe to *Consumer Reports* magazine or spend time inspecting goods or getting advice from others. Over the past few decades, economists have been intensely interested in imperfect information and its consequences for decision-making behavior.

A third problem is that people can spend more than their incomes in any given year, by borrowing funds or spending out of savings. Or they may spend less than their incomes because they choose to save or pay back debts. Economic models have been built to deal with all of these complications. In these models, consumers make choices for this year and for future years at the same time and are constrained by their total income in *all* years.

Finally, there are cases where individuals do not, in fact, choose rationally—as we have defined the term. For example, people will sometimes *judge quality by price*. Diamonds, designer dresses, men's suits, doctor's services, and even automobiles are sometimes perceived as being better if their prices are higher. This means

that the consumer cannot compare any two bundles of goods by themselves; he must first know their prices. And when prices change, so will the consumer's preferences—violating our description of rational preferences. In recent years, economists have teamed up with psychologists to study violations of rational preferences.

In sum, there are a variety of cases where the theory of the consumer, as presented in this chapter, would not work well. Economists have developed more complex models to deal with some of these cases, and research continues on the others. But we should not exaggerate their importance. If you think about your own economic decisions, you will find that, in most cases, your choices *are* rational, and the simple theory of consumer decision making presented in this chapter describes them quite accurately.

IMPROVING EDUCATION

So far in this chapter, we've considered the problem of a consumer trying to maximize utility by selecting the best combination of goods and services. But consumer theory can be extended to consider almost *any* decision between two alternatives. Economists use the model of consumer theory to understand how people choose between work and leisure, between spending now and investing for the future, and even between honest work and criminal activities. In this section, we apply the insights of consumer theory to another issue: improving the quality of education.³

Billions of dollars have been spent over the past few decades trying to improve the quality of education in our schools, colleges, and universities. In 1999 alone, the U.S. Department of Education spent about \$900 million in such efforts. Much of this money is spent on research to assess new educational techniques. For example, suppose it is thought that computer-assisted instruction might help students learn better or more quickly. A typical research project to test this hypothesis would be a *controlled experiment* in which one group of students would be taught with the computer-assisted instruction and the other group would be taught without it. Then students in both groups would be tested. If the first group scores significantly higher, computer-assisted instruction will be deemed successful; if not, it will be deemed unsuccessful. To the disappointment of education researchers, most promising new techniques are found to be unsuccessful: Students seem to score about the same, no matter which techniques are tried.

Economists find these studies highly suspect, since the experimenters treat students as passive responders to stimuli. Presented with a stimulus (the new technique), students are assumed to give a simple response (scoring higher on the exam). Where in this model, economists ask, are students treated as *decision makers* like the rest of us? In particular, where is the recognition that students must make *choices* about allocating their scarce time?

Let's apply our model of consumer choice to a student's time allocation problem. To keep things simple, we'll assume a bleak world in which there are only two activities: studying economics and studying French. Instead of costing money, each of these activities costs *time*, and there is only so much time available. And instead of buying quantities of two goods, students "buy" points on their exams with hours spent studying.

Using the THEORY

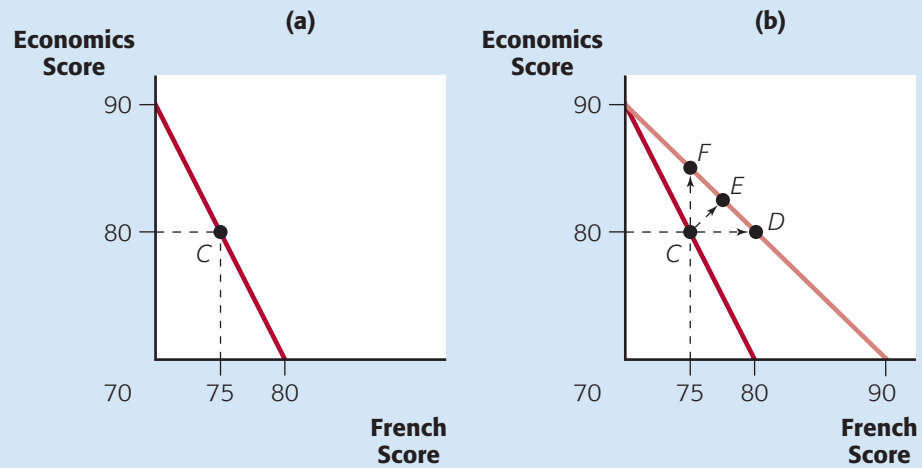


³ This section is based on ideas originally published in Richard B. McKenzie and Gordon Tullock, *The New World of Economics*, 3d ed. (Burr Ridge, IL: Irwin, 1981).

FIGURE 9

Panel (a) shows combinations of French and economics test scores that can be obtained for a given amount of study time. The slope of -2 indicates that each additional point in French requires a sacrifice of 2 points in economics. The student chooses point C. Panel (b) shows that computer-assisted French instruction causes the budget line to rotate outward; French points are now less expensive. The student might move to point D, attaining a higher French score. Or she might choose F , using all of the time freed up in French to study economics. Or she might choose an intermediate point such as E.

TIME ALLOCATION



Panel (a) of Figure 9 shows how we can represent the time allocation problem graphically. The economics test score is measured on the vertical axis and the French score on the horizontal axis. The straight line in the figure is the student's budget line, showing the trade-off between economics and French scores. Our student can achieve any combination of scores on this budget line with her scarce time.

A few things are worth noting about the budget line in the figure. First, the more study time you devote to a subject, the better you will do on the test. But that means *less* study time for the other subject and a lower test score there. Thus, the opportunity cost of scoring better in French is scoring lower in economics, and vice versa. This is why the budget line has a negative slope: The higher the score in French, the lower the score in economics. As our student moves downward along the budget line, she is shifting hours away from studying economics and toward studying French.

Second, notice that the vertical and horizontal axes both start at 70 rather than 0. This is to keep our example from becoming too depressing. If our student devotes *all* her study time to economics and none to French, she would score 90 in economics but still be able to score 70 (rather than zero) in French, just by attending class and paying attention. If she devotes all her time to French, she would score 80 in French and 70 in economics. (*Warning:* Do not try to use this example to convince your economics instructor you deserve at least a 70 on your next exam.)

Finally, the budget line in the figure is drawn as a straight line with a slope of -2 . Therefore, in this example, each additional point in French requires our student to sacrifice two points in economics, regardless of where she is on her budget line. This assumption just helps make the analysis more concrete; none of our conclusions would be different if we assumed a different slope for the budget line, or even a curved budget line, where the trade-off would change as we moved along it. But let's take a moment to understand what our example implies.

As you've learned, the slope of any budget line is $-P_x/P_y$, where x is the good measured on the horizontal axis and y is the good measured on the vertical axis. In our example, $-P_x/P_y$ translates into $-P_{\text{French point}}/P_{\text{econ point}}$. But what is the "price" of a test point in French or economics? Unlike the case of Max, who had to allocate

his scarce *fun*ds between concerts and movies, our student must allocate her scarce *time* between the two “goods” she desires: test points in French and test points in economics. The *price* of a test point is therefore not a money price, but rather a *time price*: the number of study hours needed to achieve an additional point. For example, if it takes an additional two hours of studying to achieve another point in French, then the price per point in French is two hours. In our example, we assume that the price remains constant no matter how many hours are spent studying a subject. That is, it takes an additional two hours of study time to increase the French score from 70 to 71, from 71 to 72, and so on. Moreover, in the figure, we assume that the price per point in economics is one-half the price per point in French. The slope of the budget line is therefore $-P_{\text{French point}}/P_{\text{econ point}} = -2$.

Now let’s turn our attention to student decision making. Our student derives utility from both her economics score and her French score—the greater either score, the greater is her utility. But among all those combinations of scores on her budget line, which will give her the highest total utility? We can answer this question using the same technique we used for Max and his decision between concerts and movies, but with one important difference: Instead of looking for the combination of two goods such that marginal utilities *per dollar* are equal, we look for the combination of test points in the two subjects such that marginal utilities *per hour* are equal. After all, it is hours that must be spent on additional test points, not dollars. In general:

In allocating time between two activities that provide utility, an individual will select the combination of activities such that the marginal utility per hour of one activity is equal to the marginal utility per hour of the other activity.

Suppose this condition is satisfied for our student at point C, where she scores 80 in economics and 75 in French. This is where the marginal utility per hour in French is equal to the marginal utility per hour in economics.

Now, let’s introduce a new computer-assisted technique in the French class, one that is, in fact, remarkably effective: It enables students to learn more French with the same study time or to study less and learn the same amount. This is a *decrease* in the price of French points—it now takes fewer hours to earn a point in French—so the budget line will rotate outward, as shown in panel (b) of Figure 9. On the new budget line, if our student devotes all of her time to French, she can score higher than before—90 instead of 80—so the horizontal intercept moves rightward. But since nothing has changed in her economics course, the vertical intercept remains unaffected. Notice, too, that the budget line’s slope has changed—to -1 . Now, the opportunity cost of an additional point in French is one point in economics rather than two.

After the new technique is introduced in the French course, our *decision-making* student will locate at a point on her new budget line, the point where marginal utilities per hour are equal in the two courses. Where that point is, of course, will depend on her preferences, and panel (b) illustrates some alternative possibilities. At point D, her performance in French would improve, but her economics performance would remain the same. This seems to be the kind of result education researchers have in mind when they design their experiments: If a successful technique is introduced in the French course, we should be able to measure the impact with a French test.

Point F illustrates a different choice: *Only* the economics performance improves, while the French score remains unchanged. Here, even though the technique in French is successful (it does, indeed, shift the budget line), none of its success shows up in higher French scores.

But wait: How can a new technique in the French course improve performance in economics but not at all in French? The answer is found by breaking down the impact of the new technique into our familiar income and substitution effects. You can see that the new technique lowers the time cost of getting additional points in French. The substitution effect (French points are relatively cheaper) will tend to improve her score in French, as she substitutes her time away from studying economics and toward studying French. But there is also an “income” effect: The “purchasing power” of her time has increased, since now she could use her fixed allotment of study time to “buy” higher scores in *both* courses. If performance in French is a “normal good,” this increase in “purchasing power” will work to increase her French score, but if it is an “inferior good,” it could work to *decrease* her French score. Point *F* could come about because French performance is *such* an inferior good that the negative income effect exactly cancels out the positive substitution effect. In this case, the education researchers will incorrectly judge the new technique a complete failure—it does not affect French scores at all.

Could this actually happen? Perhaps. It is easy to imagine a student deciding that 75 in French is good enough and using any time savings from better French instruction to improve her performance in some other course. More commonly, we expect a student to choose a point such as *E*, somewhere between points *D* and *F*, with performance improving in *both* courses. But even in this case, the higher French score measures just a *part* of the impact of the technique; the remaining effect is seen in a higher economics score.

This leads us to a general conclusion: When we recognize that students make *choices*, we expect only *some* of the impact of a better technique to show up in the course in which it is used. In the real world, college students typically take several courses at once and have other competing interests for their time as well (cultural events, parties, movies, telephone calls, exercising, and so on). Any time saved due to better teaching in a single course might well be “spent” on *all* of these alternatives, with only a little devoted to performing better in that single course. Thus, we cannot fully measure the impact of a new technique by looking at the score in one course alone. This suggests why educational research is conducted as it is: A more accurate assessment would require a thorough accounting for all of a student’s time, which is both expensive and difficult to achieve. Nevertheless, we remain justified in treating this research with some skepticism.

S U M M A R Y

Consumers face two simple facts of life: They have to pay for the goods and services they buy, and they have limited incomes to spend. These facts are summarized in the consumer’s *budget constraint*. Given their preferences, consumers decide which goods to consume by choosing the combination along their budget constraint that yields the greatest *utility*, or satisfaction.

According to the *law of diminishing marginal utility*, the marginal, or additional, utility derived from a good declines as more of it is consumed. A utility-maximizing consumer will choose the combination of goods along his or her budget constraint at which the marginal utility per dollar spent is the same for all goods. This is an example of the *principle of marginal decision making*.

An increase in income shifts the budget constraint outward. The consumer responds by choosing more of all normal goods and less of inferior goods. A change in the price of a good causes the budget constraint to rotate. The consumer responds by purchasing more of a good whose price has fallen and less of a good whose price has risen. By tracing out a consumer’s reaction to a series of prices, we can generate a downward-sloping *demand curve* for a good. The downward slope reflects the interaction of the *substitution effect* and the *income effect*. For a normal good, both effects contribute to the downward slope of the demand curve. For an inferior good, we can have confidence that the substitution effect dominates the income effect, so—once again—the demand curve will slope downward.

KEY TERMS

budget constraint
budget line
relative price
utility

marginal utility
law of diminishing marginal utility

rational preferences
marginal decision making
individual demand curve

substitution effect
income effect

REVIEW QUESTIONS

1. What variables are assumed constant along a budget line?
2. What kinds of changes will shift or rotate the budget line?
3. Explain the relationship between a total quantity and a marginal quantity.
4. State and explain the law of diminishing marginal utility. Can you think of a good or service you consume that is not subject to this law? Could marginal utility be negative? Give an example.
5. Economists usually assume that consumer preferences are logically consistent. What does that mean? What are some other assumptions economists make about preferences?
6. Discuss the following statement: "Economists' assumption of consumer rationality is too strong. For example, anyone who smokes cigarettes is clearly being irrational."
7. What condition will be satisfied when a consumer has chosen the combination of goods that maximizes utility subject to a budget constraint?
8. What are income and substitution effects? How are they related to the law of demand?
9. "The demand curve for an inferior good is upward sloping." True or false? Explain.
10. How is a market demand curve derived?

PROBLEMS AND EXERCISES

1. Parvez, a pharmacology student, has allocated \$120 per month to spend on paperback novels and used CDs. Novels cost \$8 each; CDs cost \$6 each. Draw his budget line. What would happen to that budget line if the price of a CD increased to \$10?
2. Parvez, our consumer from the previous question, is spending \$120 monthly on paperback novels and used CDs. For novels, $MU/P = 5$; for CDs, $MU/P = 4$. Is he maximizing his utility? If not, should he consume (1) more novels and fewer CDs or (2) more CDs and fewer novels?
3. Anita consumes both pizza and Pepsi. The following tables show the amount of utility she obtains from different amounts of these two goods:

Pizza		Pepsi	
Quantity	Utility	Quantity	Utility
4 slices	115	5 cans	63
5 slices	135	6 cans	75
6 slices	154	7 cans	86
7 slices	171	8 cans	96

- Suppose Pepsi costs \$0.50 per can, pizza costs \$1 per slice, and Anita has \$9 to spend on food and drink. What combination of pizza and Pepsi will maximize her utility?
4. Oprah is trying to decide how to allocate a 15-minute segment between two guests on an upcoming show. Pauly Shore's antics start to wear thin fast; her marginal utility from his appearance is given by $MU = 500 - 20T$, where T is the number of minutes Pauly is on. (So, the first minute Pauly is on gives Oprah 480 utils; the second minute, 460; and so on.) Tony Randall, however, is always a solid guest. Oprah can count on him for a constant 200 utils every minute he is in front of the camera. Pauly demands \$200 per minute for his appearance; Tony is happy with \$100 a minute. To maximize her utility, how much time should Oprah give each guest?
 5. Three people have the following individual demand schedules for Count Chocula cereal that show how many boxes each would purchase monthly at different prices:

Price	Person 1	Person 2	Person 3
\$5.00	0	1	2
\$4.50	0	2	3
\$4.00	0	3	4
\$3.50	1	3	5

- What is the market demand schedule for this cereal? (Assume that these three people are the only buyers.) Draw the market demand curve.
 - Why might the three people have different demand schedules?
6. What would happen to the market demand curve for polyester suits—an inferior good—if consumers' incomes rose?
7. Larsen E. Pulp, head of Pulp Fiction Publishing Co., just got some bad news: The price of paper, the company's most important input, has increased.
- On a supply/demand diagram, show what will happen to the price of Pulp's output (novels).
 - Explain the resulting substitution and income effects for a typical Pulp customer. For each effect, will the customer's quantity demanded increase or decrease? Be sure to state any assumptions you are making.

CHALLENGE QUESTION

- The Smiths are a low-income family with \$10,000 available annually to spend on food and shelter. Food costs \$2 per unit, and shelter costs \$1 per square foot per year. The Smiths are currently dividing the \$10,000 equally between food and shelter.
 - Draw their budget constraint on a diagram with food on the vertical axis and shelter on the horizontal axis. Label their current consumption choice. How much do they spend on food? On shelter?
 - Suppose the price of shelter rises to \$2 per square foot. Draw the new budget line. Can the Smiths continue to consume the same amounts of food and shelter as previously?
 - In response to the increased price of shelter, the government makes available a special income supplement. The Smiths receive a cash grant of \$5,000 that must be spent on food and shelter. Draw their new budget line and compare it to the line you derived in part a. *Could* the Smiths consume the same combination of food and shelter as in part a?
 - With the cash grant and with shelter priced at \$2 per square foot, *will* the family consume the same combination as in part a? Why, or why not?
- When an economy is experiencing inflation, the prices of most goods and services are rising but at different rates. Imagine a simpler inflationary situation in which *all* prices—and all wages and incomes—are rising at the same rate, say 5 percent per year. What would happen to consumer choices in such a situation? (*Hint*: Think about what would happen to the budget line.)

EXPERIENTIAL EXERCISES

- Together with some other students in your class, determine your individual and group demands for gasoline. Make up a chart listing the following prices per gallon: \$0.75, \$1.50, \$2.25, \$3.00, \$3.75, \$4.50. Ask each student—and yourself—how many gallons *per month* they would purchase at each possible price. Then do a, b, and c below.
 - Plot each student's demand curve. Check to see whether each curve is consistent with the law of demand.
 - Derive the "market" demand curve by adding up the quantities demanded by *all* students at each possible price.
 - What do you think will happen to that market demand curve after your class graduates and their incomes increase?
- When you consume something, you pay a money price, but there is also a "time price" involved. It takes valuable time to decide what you wish to buy, to compare items and prices, and to actually use or consume the good. Use Infotrac or the "Work and Family" column in the Wednesday *Wall Street Journal* to find an example of a new good, service, or government policy that you think will *reduce* the time price of some product. How do you think it will affect the demand for the product? Will any related products be affected?

APPENDIX

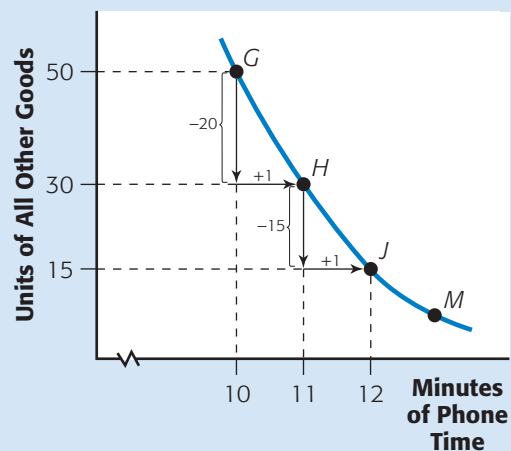
CONSUMER THEORY WITH INDIFFERENCE CURVES

One of the drawbacks in the theory of consumer choice presented in the body of this chapter—based on marginal utility—is that you could never “see” any of the important information about the consumer’s preferences. But there is another way to characterize preferences that is much more visual: using *indifference curves*. This appendix assumes that you have read the section on the budget constraint in the chapter. However, other than in one footnote, it does not assume any familiarity with marginal utility theory.

Consider Kate, whose sister is spending her junior year abroad in Moscow. Kate likes to talk to her sister on the phone, but, like every consumer, she faces a limited budget. To make our example more realistic, we’ll recognize that Kate buys a number of different goods—not just two—and that her budget constraint requires that all of her purchases together must fall within her budget. How can we use a two-dimensional diagram to indicate purchases of more than two goods? By using an economist’s trick: On one axis, we measure long-distance phone calls, and on the other we measure units of “all other goods” together. In Figure A.1, the horizontal axis measures the minutes of phone time each week, and the vertical axis measures units of *all other goods combined*. Point G, for example, represents a combination where Kate speaks to her sister for 10 minutes each week and buys 50 units of all other goods.

What will we assume about Kate’s preferences? We discussed two of our assumptions in the body of the chapter: that her choices are logically consistent (rational) and that more is better for every good. Now we introduce one more feature that seems common among people’s tastes: With many types of goods and services, people tend to prefer variety over extremes in what they consume, other things being equal. To be more precise, suppose you were indifferent between having 10 new release videos and having 10 hit CDs. That is, you will find either of these two combinations of goods equally satisfying. According to the preference for diversity in consumption, if offered a combination containing, instead, 5 of the videos along with 5 of the new CDs, you

FIGURE A.1
AN INDIFFERENCE CURVE



This curve shows all combinations of phone time and other goods that make Kate equally happy. At point G, when Kate is spending little time on the phone, the curve is relatively steep. She is willing to trade a lot of “all other goods” for one more minute of phone time. At M, the curve is flatter, indicating that she is willing to trade fewer “other goods” for an additional minute of phone time.

would prefer that alternative with variety to either of the two extreme options that concentrate your consumption on just one good. Of course, there are exceptions to this rule, too: Someone who collects both stamps and coins might prefer a complete collection of either one to a half-complete collection of each. Such instances seem special and rare, however, so economists generally assume at least some preference for diversity in a person’s consumption.

Now let’s begin characterizing Kate’s preferences by picking a point at random, such as point G in the

figure. Next, we ask Kate to tell us how much we can reduce her consumption of all other goods if we give her one more minute of conversation with her sister, so that she will be no better and no worse off after the change. Suppose she tells us, “I’d trade 20 units of all other goods for one more minute of conversation with my sister and feel no better and no worse off for the change.” Then Kate must be *indifferent between* point *G* on the one hand and point *H* on the other, since point *H* gives her one more minute of phone time and 20 units less of all other goods than point *G*.

Next, we ask Kate to imagine herself at point *H*, and we ask her the same question. If she answers, “I’d trade off 15 units of other goods for another minute of conversation,” then she must be indifferent between point *H* and point *J*, since *J* gives her one more minute of phone time and 15 units less of other goods than point *H*. Now we know that Kate is indifferent between point *J* and point *H* and between point *H* and point *G*. So long as she is rational, she must be entirely indifferent among all three points—*G*, *H*, and *J*. (Remember, rational choice requires consistent choice.) By continuing in this way, we can trace out a set of points that—as far as Kate is concerned—are equally satisfying and so give her the same total utility from her consumption. When we connect these points with a line, we obtain one of Kate’s indifference curves.

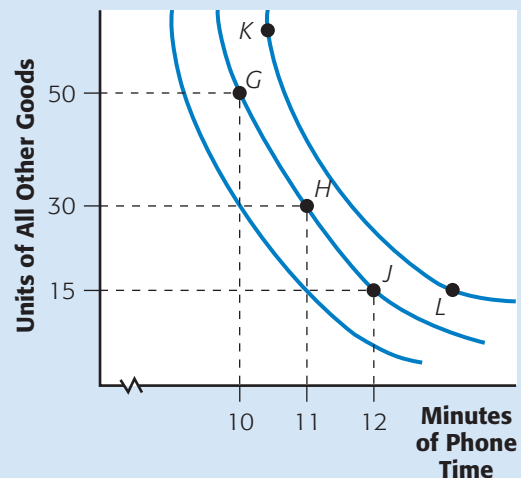
An indifference curve represents all combinations of two categories of goods that make the consumer equally well off.

Notice two things about the indifference curve in Figure A.1. First, it slopes downward. Second, it bows away from the origin, becoming flatter as we move southeasterly along it. Each of these results from the assumptions we’ve made about preferences. The indifference curve slopes downward because every time we give Kate one more minute of phone time, we make her better off (more is better). In order to find another point on her original indifference curve, we must make her worse off by the same amount, *taking away* spending on other goods. Thus, each time we move rightward along an indifference curve, we must also move downward.

The slope of an indifference curve—the change along the vertical axis divided by the change along the horizontal axis as we move along it—tells us the rate at which Kate could trade all other goods for phone calls

and still remain indifferent. But what about the *curvature* of her indifference curve? Why has it been drawn negatively sloped *and* bowed away from the origin? Interestingly, this kind of shape must result whenever preferences satisfy the “principle of diversity.” At points such as *G*—high on her indifference curve—Kate consumes a lot of “all other goods” and relatively few phone calls compared to points lower down, such as *M*. When Kate has a preference for variety in her consumption, she will be quite willing to trade off a lot of her other goods for one more minute of phone calls, and this is reflected by the relatively steep slope of the indifference curve at *G*. Similarly, if Kate is spending a lot of time on the phone and consumes relatively little other goods, as is the case at point *M*, she will be very reluctant to trade off any more of her now relatively scarce other goods for more time on the phone. This is reflected in the relatively flat slope of the indifference curve at *M*. Thus, we can expect that whenever preferences show some taste for variety in consumption, an indifference curve will become flatter as we move along it downward and rightward.

FIGURE A.2
AN INDIFFERENCE MAP



These three indifference curves are part of Kate’s indifference map. She prefers higher curves to lower ones.

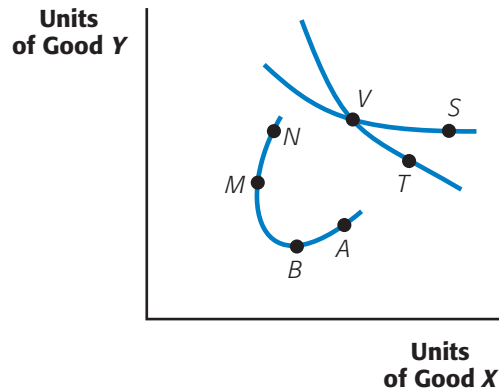
THE INDIFFERENCE MAP

To trace out the indifference curve in Figure A.1, we began at a specific point—point *G*. In Figure A.2 we've reproduced the indifference curve through *G*, *H*, and *J*. But now consider the new point *K*, which involves more phone time and more of other goods than point *G*. We know that point *K* is preferred to point *G* ("more is better"), so it is not on the indifference curve that goes through *G*. However, we can use the same procedure we used earlier to find a *new* indifference curve, connecting all points indifferent to point *K*. Indeed, we can repeat this procedure for any initial starting point we might choose, tracing out dozens, hundreds, or even thousands of Kate's indifference curves—as many as we'd like.

The result would be an *indifference map*—a set of indifference curves that describe Kate's preferences, like the three curves in Figure A.2. Although we cannot say how much satisfaction Kate experiences on any particular indifference curve, we do know that she would always prefer any point on a higher indifference curve to any point on a lower one. For example, consider the points *G* and *L*. *L* involves more phone time but less of other goods than *G*. How can we know if Kate prefers *L* to *G*, or *G* to *L*? Kate's indifference map tells us that she *must* prefer *L* to *G*. Why? We know that she prefers *K* to *G*, since *K* has more of both phone time and other goods. We also know that Kate is indifferent between *L* and *K*, since they are on the same indifference curve. Since she is indifferent between *L* and *K* but prefers *K* to *G*, then she must also prefer *L* to *G*.



There are two common mistakes students make when drawing indifference curves. One is to allow the ends of the curve to "curl up," like the curve through point *B* in the following figure, so that the curve slopes upward at the ends. This violates our assumption of "more is better." To see why, notice that point *A* has more of both goods than point *B*. So as long as "more is better," *A* must be preferred to *B*. But then *A* and *B* are not indifferent, so they cannot lie on the same indifference curve. For the same reason, points *M* and *N* cannot lie on the same indifference curve. Remember that indifference curves cannot slope upward.



The second mistake is to allow two indifference curves to cross. For example, look at the two indifference curves passing through point *V*. *T* and *V* are on the same indifference curve, so the consumer must be indifferent between them. But *V* and *S* are also on the same indifference curve, so the consumer is indifferent between them, too. Since rationality requires the consumer's preferences to be consistent, the consumer must then also be indifferent between *T* and *S*, but this is impossible because *S* has more of both goods than *T*, a violation of "more is better." Remember that indifference curves cannot cross.

The same technique could be used to show that

any point on a higher indifference curve is preferred to any point on a lower one.

Thus, Kate's indifference map tells us how she ranks all alternatives imaginable. This is why we say that an indifference map gives us a complete characterization of someone's preferences: It allows us to look at any two points and—just by seeing which indifference curves they are on—immediately know which, if either, is preferred.

THE MARGINAL RATE OF SUBSTITUTION

The slope of the indifference curve along any one of its segments tells us the rate at which a consumer is willing to trade off one good for another and still remain indifferent. The *absolute value* of this slope is called the *marginal rate of substitution*, or *MRS*. The *MRS* plays an important role in the indifference curve approach, so let's define it. When the quantity of good *y* is measured on the vertical axis and the quantity of good *x* is measured on the horizontal axis,

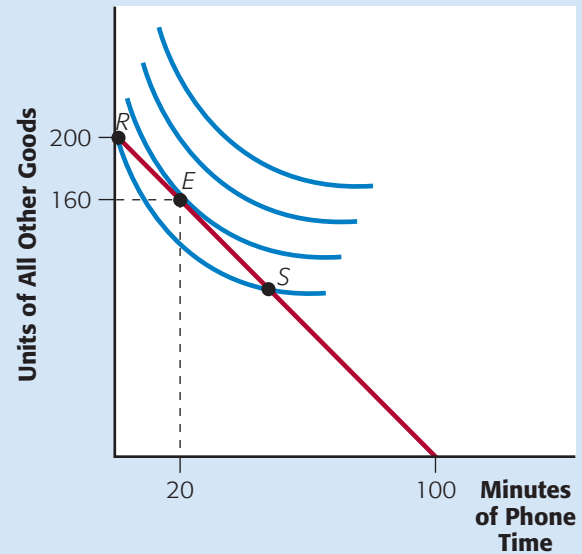
*the marginal rate of substitution of good y for good x ($MRS_{y,x}$) along any segment of an indifference curve is the (absolute value of) the indifference curve's slope along that segment. The *MRS* tells us the decrease in the quantity of good y needed to accompany a one-unit increase in good x, in order to keep the consumer indifferent to the change.*

Although the *MRS* has a technical definition, its meaning is quite simple. Look back at Figure A.1. If Kate were currently consuming at point *G*, she could tell us the following: "If you gave me one more phone call (good *x*) and took away 20 units of all other goods (good *y*), then I'd be just as well off as I am now." Then Kate would be telling us that her *MRS* is 20. As you can see in the figure, this is also equal to the absolute value of the indifference curve's slope along the segment *GH*. (The decrease along the vertical axis is 20, and the increase along the horizontal axis is 1, so the absolute value of the slope is $20/1 = 20$.)

CONSUMER DECISION MAKING

Now we can combine everything you've learned about budget lines in the chapter, and what you've learned about indifference maps and the marginal rate of substitution in this appendix, to determine the combination of goods Kate should choose. Figure A.3 adds Kate's budget line to her indifference map. In drawing the budget line, we suppose that she has a weekly budget of \$200 to spend on phone calls and all other goods, and that long-distance phone rates are \$2 per minute during the time that Kate likes to call her sister. We also assume that the price of a unit of all other goods is \$1.

FIGURE A.3
CONSUMER DECISION MAKING



Kate's most preferred combination of phone calls and "all other goods" is at point *E*. It is a point on the highest indifference curve attainable, given her budget and the prices of the two goods. At a point like *R*, her *MRS* exceeds the slope of her budget line, so she would be better off increasing her phone time and moving to a higher indifference curve. At *S*, her *MRS* is less than the slope of the line, so she would be better off cutting back on phone time.

If Kate devotes all of her income to phone calls and none to other goods, she could have 100 minutes of phone time per month. This is the horizontal intercept of her budget line in the figure. On the other hand, she could choose zero phone time and 200 units of all other goods—the vertical intercept. Since the price of a minute of phone time (good *x*) is \$2, and the price per unit of all other goods (good *y*) is \$1, the slope of Kate's budget line is $-P_x/P_y = -\$2/\$1 = -2$. Each additional minute of phone time requires Kate to give up \$2 in other goods.

Kate's optimal combination of phone calls and other goods will satisfy two criteria: (1) It will be a point on her budget line, and (2) it will lie on the highest indifference curve possible. Kate can find this point

by traveling down her budget line from point R. As she does so, she will pass through a variety of indifference curves. (To see this clearly, pencil in some indifference curves *between* the ones drawn in the figure.) At first, each indifference curve is higher than the one before until she reaches the highest curve possible. This occurs at point E, where she buys 160 units of other goods and 20 minutes of long-distance calls per week. Any further moves down the budget line will put her on lower indifference curves, so these moves would make her worse off. Point E is her optimal choice, then.

Notice two things about point E. First, it occurs where the indifference curve and the budget line touch but don't cross. As you can see in the diagram, when an indifference curve actually crosses the budget line, we can always find some other point on the budget line that lies on a higher indifference curve.

Second, at point E, the slope of the indifference curve is the same as the slope of the budget line. Does this make sense? It should—when you think about it this way: The slope of the indifference curve is minus the marginal rate of substitution between two goods. The MRS tells us the rate at which the consumer could trade one good for the other and remain indifferent. The slope of the budget line, by contrast, tells us the rate at which the consumer *is actually able* to trade one good for the other. If there is any difference between the rate at which a consumer could trade one good for the other with indifference and the rate at which she is *able* to trade, she can always make herself better off by moving to another point on the budget line. For example, suppose Kate were at point R, where her indifference curve is steeper (slope = -10) than her budget line (slope = -2). Since her MRS at point R is 10, she could give up 10 units of other goods for one more minute of phone time and remain indifferent. But Kate's budget line tells us that she is *able* to trade just 2 units of other goods for another minute of phone time. If trading away 10 units of other goods for another minute would leave her indifferent, but she *actually has* to give up only 2 units, then she must be better off by making the trade. We conclude that *when Kate's indifference curve is steeper than her budget line, she should spend more on phone calls and less on other goods.*

Using similar reasoning, convince yourself that Kate should make the opposite move—spending less on phone calls and more on other goods—if her indifference curve is flatter than her budget line, as it is at point S. Only when the indifference curve and the budget line

have the same slope—when they touch but do not cross—is Kate as well off as possible. More generally,

the optimal combination of goods for a consumer is that combination on the budget line at which the indifference curve has the same slope as the budget line.

Now remember that (the absolute value of) the slope of an indifference curve at any point is equal to the marginal rate of substitution between the two goods ($MRS_{y,x}$), while (the absolute value of) the slope of the budget line is equal to P_x/P_y . Using these two facts, our conclusion about the consumer's optimal choice can be expressed this way:

The optimal combination of two goods x and y is that combination on the budget line for which
 $MRS_{y,x} = P_x/P_y$.

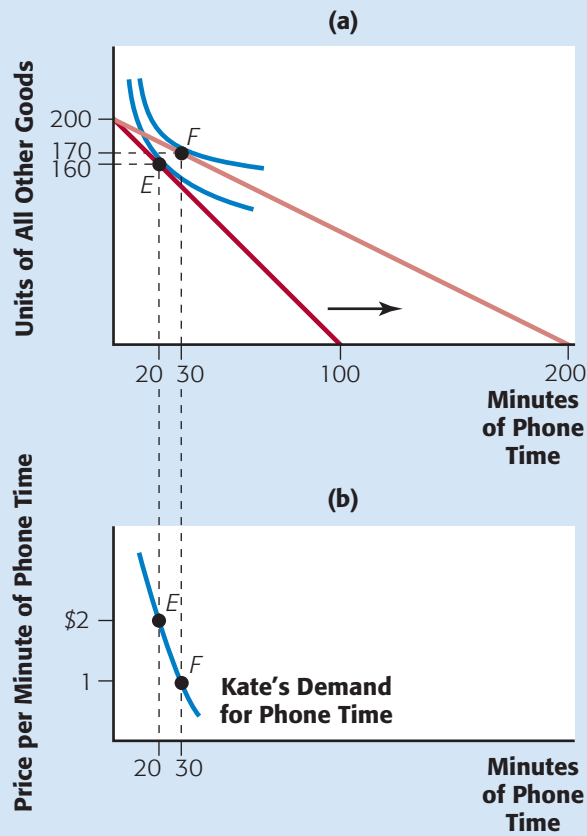
If this condition is not met, there will be a difference between the rate at which a consumer could trade good y for good x and *remain indifferent*, and the rate at which she *could actually* make the trade. This will always give the consumer an opportunity to make herself better off.⁴

⁴ The body of this chapter covers the marginal utility approach to consumer theory. You might be wondering whether the indifference curve approach leads to a different consumer choice. The answer is no: Both approaches lead to the same optimal combination of goods. Here is the proof:

First, note that $MRS_{y,x} = MU_x/MU_y$ for any change along an indifference curve. Why? The $MRS_{y,x}$ tells us the decrease in good y per unit change in good x that keeps the consumer indifferent. When good y decreases, utility falls by $MU_y \times \Delta y$ (the change in utility per unit change in y times the change in y). When x increases, utility increases by $MU_x \times \Delta x$. Now remember that as we move along an indifference curve, total utility must remain unchanged to keep the consumer indifferent. Thus, as we move along an indifference curve, it must be true that $MU_y \times \Delta y = MU_x \times \Delta x$, or $\Delta y/\Delta x = MU_x/MU_y$. The left-hand side is the change in y per unit change in x that keeps the consumer indifferent, or $MRS_{y,x}$, so we have $MRS_{y,x} = MU_x/MU_y$.

Now, in the marginal utility approach, the consumer chooses a combination of goods such that $MU_x/P_x = MU_y/P_y$, or, rearranging, $MU_y/MU_x = P_x/P_y$. In the indifference curve approach, the consumer chooses the combination such that $MRS_{y,x} = P_x/P_y$. Since $MU_y/MU_x = MRS_{y,x}$, using the marginal utility approach or the MRS approach will give us the same optimal combination of goods.

FIGURE A.4
DERIVING THE DEMAND CURVE



At \$2 per minute of phone time, Kate chooses point E in panel (a) and spends 20 minutes on the phone. If the price falls to \$1 per minute, her budget line rotates outward; she moves to point F , consuming 30 minutes of phone time. The demand curve in panel (b) is obtained by connecting price–quantity combinations like E and F .

INDIFFERENCE CURVES AND THE INDIVIDUAL DEMAND CURVE

We can also use indifference curves to derive Kate's demand curve for long-distance phone calls. In panel (a) of Figure A.4, we show what happens when long-distance rates fall from \$2 per minute to \$1 per minute. First, the horizontal intercept of Kate's budget line will move rightward, from 100 minutes to 200 minutes. Second, Kate will travel down her new budget line until she reaches point F , which places her on the highest indifference curve possible. At this point, she buys 170 units of other goods and speaks to her sister for 30 minutes each week. Panel (b) shows Kate's demand curve for long-distance phone time, based on the information in panel (a). At \$2 per minute, she buys 20 minutes of phone time, and at \$1 per minute, she buys 30 minutes. Notice that Kate's demand curve for long-distance phone calls satisfies the law of demand: A drop in long-distance rates increases the quantity of phone time demanded.

PRODUCTION AND COST

CHAPTER

6

In the early 1990s, the Russian Federation began a remarkable transformation from centrally planned socialism to market capitalism. In some areas, progress has been remarkable. The Russian government has transformed its legal and financial systems, privatized virtually all state-owned factories and stores, and granted substantial autonomy to regional and city governments. Indeed, in the early 1990s, Russia seemed poised for a period of remarkable economic growth and rising living standards.

But it hasn't worked out that way. Since 1989, output per person—and the average living standard—has *fallen* by about 30 percent. What happened?

The entire answer to that question is controversial, and complex. But there is widespread agreement about at least *one* part of the answer: *There is something peculiar about many Russian firms.* After you read this chapter, you'll understand just what that peculiarity is.

In this chapter, we begin our study of the business firms that produce and sell goods and services. The first section addresses some very general, but important, questions. What are business firms? What advantages do business firms enjoy over other ways of organizing production? Why do so many of us work as employees of firms? Then, in the remainder of the chapter, we turn our attention to the nature of production and cost. You will see that there are many different ways of measuring costs, each telling us something different about the firm.

THE NATURE OF THE FIRM

A business firm is an organization, owned and operated by private individuals, that specializes in production.

Your first image when you hear the word *production* may be a busy, noisy factory where goods are assembled, piece by piece, and then carted off to a warehouse for eventual sale to the public. Large manufacturers may come to mind—General Motors, Boeing, or even Ben & Jerry's. All of these companies produce things, but the word *production* encompasses more than just manufacturing.

CHAPTER OUTLINE

The Nature of the Firm

Types of Business Firms
Why Employees?
The Limits to the Firm

Thinking About Production

The Short Run and the Long Run

Production in the Short Run

Marginal Returns to Labor

Thinking About Costs

The Irrelevance of Sunk Costs
Explicit Versus Implicit Costs

Costs in the Short Run

Measuring Short-Run Costs
Explaining the Shape of the
Marginal Cost Curve
The Relationship Between Average and Marginal Costs

Production and Cost in the Long Run

The Relationship Between Long-Run and Short-Run Cost
Explaining the Shape of the
LRATC Curve

Using the Theory: Cost Curves and Economic Reform in Russia

Business firm A firm, owned and operated by private individuals, that specializes in production.

Production is the process of combining inputs to make outputs.

Some outputs are, indeed, physical *goods*, like automobiles, aircraft, or ice cream. But outputs can also be *services*. Indeed, many of America's largest corporations produce services. Think of Citicorp (banking services), American Airlines (transportation services), Bell Atlantic (telecommunications services), and Wal-Mart (retailing services).

Figure 1 illustrates the relationships between the firm and those it deals with. Notice that we have put the firm's management in the center of the diagram. It is the managers who must decide what the firm will do, both day-to-day and over a longer time horizon. When we refer to the firm as a *decision maker*, we mean the manager or managers who actually make the decisions.

As you can see in the figure, the firm must deal with a variety of individuals and organizations. It sells its output to *customers*—which can be households, government agencies, or other firms—and receives *revenue* from them in return. For example, Ford Motor Company sells its automobiles to households, to other firms (such as rental car companies), and to government agencies (such as local police departments). Ford earns revenue from all of these customers.

Where does the revenue go? Much of it goes to *input suppliers*. Ford must pay for labor, machinery, steel, rubber, electricity, factory buildings and the land underneath them, and much, much more. The total of all of these payments makes up the firm's *costs* of production.

When costs are deducted from revenue, what remains is the firm's **profit**:

$$\text{Profit} = \text{Revenue} - \text{Costs.}$$

Figure 1 shows that the firm's profit (after taxes) accrues to the *owners* who provided the firm's initial financing.

Finally, every firm must deal with the government. On the one hand, it pays taxes to the government, and must obey government laws and regulations. On the other hand, firms receive valuable services from the government. These include the use of public capital, like roads and bridges, as well as the presence of a legal and financial system that help the economy run smoothly.

TYPES OF BUSINESS FIRMS

There are about 20 million business firms in the United States, and each of them falls into one of three legal categories, based on the rules and conditions of ownership. In a **sole proprietorship**, a single individual starts the firm, owns it, and is entitled to all of the profit after taxes. In Figure 2, you can see that most business firms are sole proprietorships. This is not surprising, since they are the easiest form of business to start. In many cases, the owner just begins doing business. For tax purposes, the firm's profit is simply treated as part of the owner's personal income and is subject to the personal income tax.

In a **partnership**, responsibilities are shared among several co-owners. One clear advantage of a partnership is that each owner can take time off, leaving the others responsible for the firm. In addition, partners can often share many inputs—such as secretaries, advertising, and reception areas—reducing the costs for each partner. Of course, the profits must be shared with the co-owners as well. Partnerships are common among professionals, such as doctors, lawyers, and architects.

Although sole proprietorships and partnerships are easy to create, they share two problems that ultimately make many owners decide against them. The first is

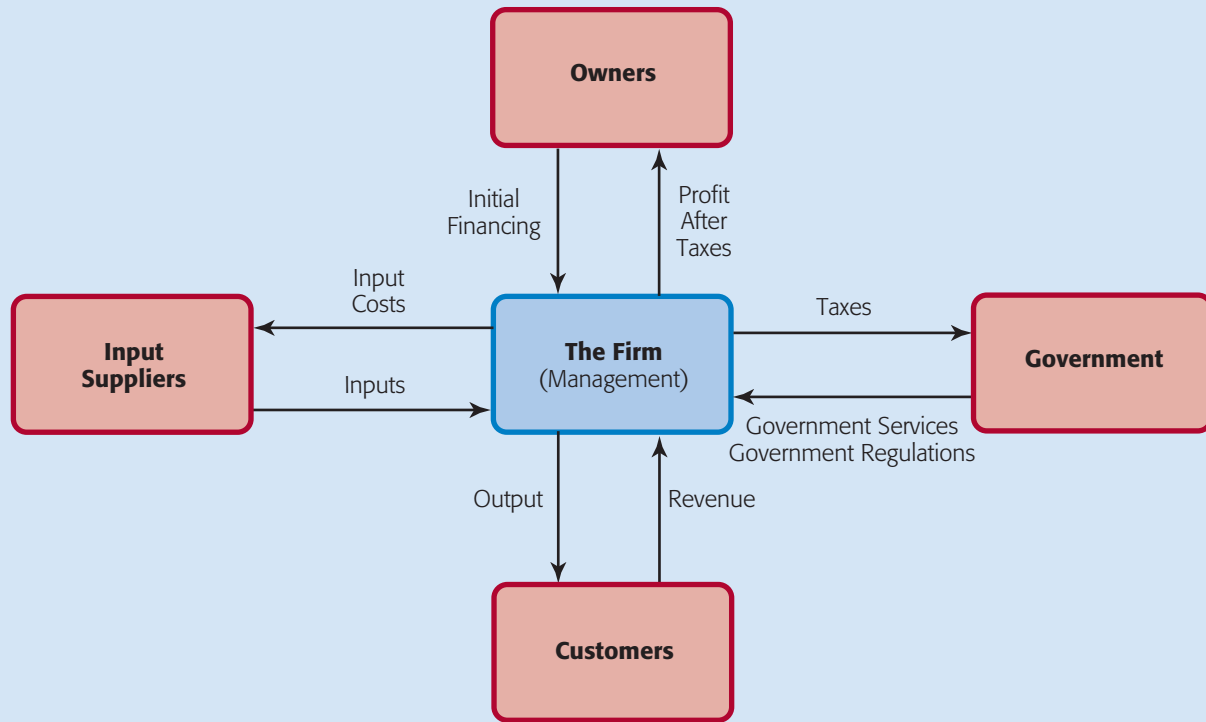
Profit Total revenue minus total cost.

Sole proprietorship A firm owned by a single individual.

Partnership A firm owned and usually operated by several individuals who share in the profits and bear personal responsibility for any losses.

THE FIRM AND ITS ENVIRONMENT

FIGURE 1



unlimited liability: In either of these types of businesses, each owner is held personally responsible for the obligations of the firm. If the business runs up debts and closes down, or is successfully sued for a large sum of money, the owners will usually have to honor these obligations out of their own pockets.

The second problem is the difficulty of raising money to expand the business. In a sole proprietorship or partnership, owners must think very carefully before bringing in new partners—especially strangers—because each partner bears full responsibility for the poor judgment of any one of them. Thus, when owners need additional funds, they must usually use their own money or borrow from a bank. In either case, the current owners bear all the risk if the business fails.

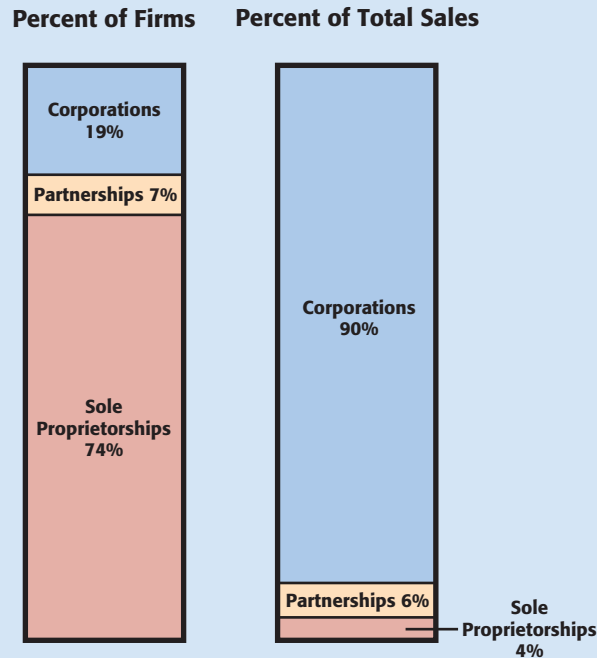
These drawbacks lead many business firms to choose the third type of organization: a **corporation**. In this type of firm, ownership is divided among those who buy shares of *stock*. Each share of stock entitles its owner to a vote for the board of directors, which in turn hires the corporation's top managers. And each share of stock entitles its owner to a share of the corporation's profit—some of which is paid out as *dividends*. The corporate form of organization makes it easier to raise additional funds: The corporation simply sells additional shares of stock, thereby bringing in new owners. People are less hesitant to become co-owners of a corporation because of its other chief advantage: *limited liability*. The owners (stockholders) of a corporation can lose only what they have paid for the stock they own; they will never have to reach into their own pockets to honor the firm's obligations.

Why, then, doesn't every firm choose the corporate form? Because a corporation, in addition to its many advantages, also has its additional costs. To set up a corporation, government documents must be filed, and lawyers and accountants are

Corporation A firm owned by those who buy shares of stock and whose liability is limited to the amount of their investment in the firm.

FIGURE 2

FORMS OF BUSINESS ORGANIZATION



usually hired to help with the job. And once you incorporate, you are subject to a variety of laws and regulations that apply only to corporations. Finally, owners of corporations suffer *double taxation*. First, the corporation pays taxes on its total profit. Then, households must pay income taxes on the portion of profit they receive as dividends. Each dollar of profits is thus taxed twice: once as corporate profits and again as household income. Still, for the largest firms, the advantages of incorporating outweigh the disadvantages. Although only a minority—about 20 percent—of businesses choose to be corporations, they tend to be large firms, producing about 90 percent of our national output (see Figure 2).

WHY EMPLOYEES?

Most firms have *employees*—people who work for the firm and receive a wage or salary, but are not themselves owners. Indeed, most of us will spend the greater part of our lives working as employees of firms owned by other people. We are so accustomed to this arrangement that we rarely think about it. But life didn't have to be this way. There is no law to prevent each of us from operating our own one-person firms as independent contractors. Indeed, there would be many advantages to this sort of arrangement: We could each determine our own hours, we could set our own work rules, and no one could fire us, no matter what we did. So why don't more of us do it?

To understand why so many people work as employees, consider the alternative: each of us working as independent contractors. In such an economy, we would each specialize in a craft or profession and trade with each other, but *work* only for ourselves. If you wanted to buy a futon frame, you would go to an independent furniture maker. She, in turn, would buy her saw from an independent saw maker, her

lumber from a lumber cutter, and so on throughout the economy. In this way, we would each operate on our own. And we would each concentrate on an activity in which we had a comparative advantage, learn to do it well, and buy the materials we needed from other individuals. As a result, we would all enjoy a greater standard of living than would be possible if each of us were entirely self-sufficient. But we would not be enjoying the highest standard of living possible.

The Advantages of Employment. Suppose that, in this economy of independent contractors, someone got a brilliant idea: to set up a new organization, a *firm with employees*, to produce futon frames. In this firm, hundreds or even thousands of employees would promise to show up for work every day in exchange for an agreed-upon wage or salary. Would this kind of production have major advantages over production by independent contractors? Absolutely.

Gains from Specialization. One advantage of production by firms with employees is the possibility of further gains from specialization. When many people work within a single organization, assembly line methods, in which each worker specializes in *one aspect* of production, become feasible. Whereas the independent contractor must design the futon frame, make it, deal with customers, and advertise her services, at the furniture factory each of these tasks would be performed by different individuals who would work full time at their activity. This increases the gains from specialization.

Lower Transaction Costs. Another advantage for a firm with employees is lower **transaction costs**—a term economists use for the hassles of doing business. It takes time to find reliable suppliers of cloth, wood, and tools, and time to negotiate deals with each of them. In a world of independent contractors, where business relationships would be more temporary and flexible, each of us would spend a great deal of time searching for high-quality, reliable suppliers and negotiating contracts with them. As a result, transaction costs would be high.

In a firm with employees, however, many supplies and services can be produced *inside* the organization, by *employees*. The firm's owners negotiate just *one* contract with each person—an employment contract—specifying the responsibilities and obligations of both sides. As long as employee turnover isn't too great, the firm can enjoy significant savings on transaction costs.

Reduced Risk. Finally, the large firm with employees offers opportunities for everyone involved to reduce *risk*. When workers join firms and agree to work for a stable wage or salary, they receive a kind of insurance that protects them against fluctuations in their incomes. The protection is not complete—there is always the possibility of being laid off when times are bad. But it is understood by both firms and workers that those who remain on the job will continue to receive their regular wage or salary, regardless of business conditions. Many people—preferring not to gamble with the source of their livelihood—place a high value on this feature of employment contracts, a feature not available to the independent contractor.

But how can firms provide this kind of protection to employees? Doesn't offering stable wages, even when business is bad, increase the variability of the firm's profits? Won't this increase the risk faced by the firm's owners?

Perhaps. But large firms create opportunities for owners to reduce their risk, too, through **diversification**. To *diversify* is to spread the source of your income among several different alternatives, as suggested by the saying "Don't put all your

Transaction costs The time costs and other costs required to carry out market exchanges.

Diversification The process of reducing risk by spreading sources of income among different alternatives.

eggs in one basket.” With large firms, two kinds of diversification are possible. First, the firm itself can produce several different product lines, so that if one is selling poorly, another may be selling well. This is diversification *within* the firm.

Second, owners need not limit themselves to ownership of just one firm; instead, they can spread their investment, buying shares in a *portfolio* of firms. The portfolio can be carefully chosen so that when some firms are doing poorly, others are likely to be doing well. This is diversification *among* firms, and it allows the income of each owner to be more stable than the profits at any one firm.

You can see that a large firm with employees offers several advantages over independent contractors. These advantages help it attract customers, workers, and potential owners. Here’s why:

- The greater gains from specialization and the saving on transaction costs allow the firm with employees to produce a given amount of output using fewer resources than would a collection of independent contractors.
- Since the firm can produce its output using fewer resources, it can charge lower prices—attracting *customers* away from independent contractors.
- Since the firm saves on resources, it can afford to pay a higher wage rate to its workers than they could earn as independent contractors. The firm can also provide its workers with valuable insurance against income fluctuations. These advantages induce many independent contractors to become *employees*.
- Opportunities for diversification within and among firms help reduce the risk to potential *owners*, enticing them to organize firms.

Since modern firms with employees have such an edge in winning customers, attracting workers, and enticing potential owners, it is not surprising that they produce so much of our output.

THE LIMITS TO THE FIRM

From all of this, you might be tempted to conclude that bigger is always better—the larger the firm, the greater will be the cost savings. But if that were true, there would be just one enormous firm in the economy, and we’d all be working for it! In fact, there are limits to the gains from specialization, the savings on transaction costs, and opportunities for diversification. Bigger is *not always* better.

Why? Because as firms expand in size, they begin to encounter difficulties. For one thing, larger firms have more layers of management than small firms. Major corporations like IBM, General Motors, and Bell Atlantic each have several hundred high-level managers, and thousands more at lower levels. In a firm with so many managers, communication and decision making become more complex and time consuming. Indeed, for much of the 1980s, IBM was criticized by its stockholders for failing to keep up with rapid changes in the market for small computers. According to its critics, IBM had grown so large that decision making had become sluggish. In the 1990s, Compaq—which had grown into a huge firm with a large bureaucracy—was unable to keep up with clever production and marketing strategies developed by smaller, more nimble competitors like Dell and Gateway.

Large firms also have difficulty *monitoring* their workers, to prevent shirking or sloppy work. These problems increase costs at the firm, counteracting the cost advantages of bigness described earlier. Eventually, as a firm continues to grow, a point is reached at which the advantages of further growth are outweighed by the disadvantages. This explains why firms do not grow indefinitely larger.

In some types of production, the disadvantages of bigness set in right away, and independent contractors will have the advantage. Plumbing, shoe repair, gar-

dening, and psychotherapy are almost always provided by independent contractors rather than large firms. These are jobs where it is best to have a single professional or craftsperson perform a *variety* of tasks; further specialization *within* the craft would create losses rather than gains. (Imagine the disadvantages of a plumbing firm in which each worker specializes: One finds your problem, another removes your old pipes, another locates a replacement pipe, another writes the bill, and so on.)

THINKING ABOUT PRODUCTION

When you think of production, it is quite natural to think of *outputs*—the things firms *make*—and *inputs*—the things firms *use* to make outputs. Inputs include resources (labor, capital, and land), as well as raw materials and other goods and services provided by other firms. For example, to produce this book, South-Western College Publishing Company used a variety of inputs: *labor* (including that provided by the authors, editors, artists, printers, and company managers), *human capital* (the knowledge and skills possessed by each of the preceding workers); *physical capital* (including computers, delivery trucks, and a company headquarters building in Cincinnati); and *land* (under the headquarters). The company also used many inputs that were produced by *other* firms, including raw materials such as paper and ink, as well as the services of trucking companies, telephone companies, and Internet access providers.

The way in which these inputs may be combined to produce output is the firm's **technology**. We leave it to engineers and scientists to spell out a firm's technology and to discover ways to improve it. When thinking about the firm, economists consider technology as a given, a *constraint* on the firm's production. This constraint is spelled out by the firm's *production function*:

For each different combination of inputs, the production function tells us the maximum quantity of output a firm can produce over some period of time.

The idea behind a production function is illustrated in Figure 3. Quantities of each input are plugged into the box representing the production function, and the maximum quantity of goods or services produced pops out. The production function itself—the box—is a mathematical function relating inputs and outputs.

When a firm uses many different inputs, production functions can be quite complicated. This is true even of small firms. For example, the production function for a video store would tell us how many videos it could rent per day with different combinations of floor space, shelving, sales clerks, cash registers, videos in stock, lighting, air conditioning, and so on.

Technology A method by which inputs are combined to produce a good or service.



Identify Goals and Constraints

Production function A function that indicates the maximum amount of output a firm can produce over some period of time from each combination of inputs.

THE FIRM'S PRODUCTION FUNCTION



FIGURE 3

In this chapter, to keep things simple, we'll spell out the production function for a mythical firm that uses only two inputs: capital and labor. Our firm is Spotless Car Wash, whose output is a service: the number of cars washed. The firm's capital is the number of automated car-washing lines, and its labor is the number of full-time workers who drive the cars onto the line, drive them out, towel them down at the end, and deal with customers.¹

THE SHORT RUN AND THE LONG RUN

When a firm alters its level of production, its input requirements will change. Some inputs, such as labor, can be adjusted relatively quickly. Other inputs—for example, capital equipment—may be more difficult to change. Why? Leases or rental agreements may commit the firm to keep paying for equipment over some period of time, whether the equipment is used or not. Or there may be practical difficulties in adjusting capital, like a long lead time needed to acquire new equipment or sell off existing equipment. These considerations make it useful to categorize firms' decisions into one of two sorts: *long-run decisions* and *short-run decisions*. The **long run** is a time horizon long enough for a firm to vary *all* of its inputs.

Long run A time horizon long enough for a firm to vary all of its inputs.

The long run will be different for different firms. For a surgeon who would need several months to obtain a new surgical laser, to find a buyer for the one he has, or to find a larger or a smaller office, the long run is several months or more. At Spotless Car Wash, it might take a year to acquire and install an additional automated line or to sell the ones it already has. For Spotless, then, the long run would be any period longer than a year.

When a firm makes long-run decisions, it makes choices about *all* of its inputs. But firms must also make decisions over shorter time horizons, during which some of its inputs *cannot* be adjusted.

The short run is a time horizon over which at least one of the firm's inputs cannot be varied.

Short run A time horizon during which at least one of the firm's inputs cannot be varied.

For Spotless Car Wash, the short run would be any period *less* than a year, the period during which it is stuck with a certain number of automated lines.

You can think of the short run and long run as two different lenses that a firm's manager must look through to make decisions. The short-run lens makes at least one of the inputs appear to be fixed, but the long-run lens makes all inputs appear variable. To guide the firm over the next several years, the manager must use the long-run lens; to determine what the firm should do next week, the short-run lens is best.

PRODUCTION IN THE SHORT RUN

In this section, we'll be describing important features of production in the short run. Remember that in the short run, at least one of the firm's inputs cannot be varied. As a result, the firm will have two types of inputs: fixed and variable.

¹ Of course, a car wash would use other inputs besides just capital and labor: water, washrags, soap, electricity, and so on. But the costs of these inputs would be minor when compared to the costs of labor and capital. To keep our example simple, we will ignore these other inputs entirely.

Fixed inputs are those whose quantity remains constant, regardless of how much output is produced. Variable inputs are those whose quantity changes as the level of output changes.

When firms make short-run decisions, there is nothing they can do about their fixed inputs: They are stuck with whatever quantity they have. They can, however, make choices about their variable inputs. Indeed, we see examples of such short-run decisions all the time. Boeing might decide *this month* to cut its production of aircraft by 5 percent and lay off thousands of workers, even though it cannot change its factory buildings or capital equipment for another year or more. For Boeing, labor is variable, while its factory and equipment are fixed. Levi Strauss might decide to increase production of blue jeans over the next quarter by obtaining additional workers, cotton cloth, and sewing machines, yet continue to make do with the same factories because there isn't time to expand them or acquire new ones. Here, workers, cloth, and sewing machines are all variable, while only the factory building is fixed.

Spotless Car Wash uses only two inputs to produce its output—labor and capital. Its only variable input is labor, and its only fixed input is capital. The three columns in Table 1 describe Spotless's production function in the short run. Column 1 shows the quantity of the fixed input, capital (K); column 2 the quantity of the variable input, labor (L). Note that in the short run, Spotless is stuck with one unit of capital—one automated line—but it can take on as many or as few workers as it wishes. Column 3 shows the firm's *total product* (Q).

Total product is the maximum quantity of output that can be produced from a given combination of inputs.

For example, the table shows us that with one automated line but no labor, total product is zero. With one line and six workers, output is 185 cars washed per day.

Figure 4 shows Spotless's *total product curve*. The horizontal axis represents the number of workers, while the vertical axis measures total product. (The amount of capital—which is held fixed at one automated line—is not shown on the graph.) Notice that each time the firm hires another worker, output increases, so the total product curve slopes upward. The vertical arrows in the figure show precisely *how much* output increases with each one-unit rise in employment. We call this rise in output the *marginal product of labor*.

Fixed input An input whose quantity remains constant, regardless of how much output is produced.



Identify Goals and Constraints

Variable input An input whose usage changes as the level of output changes.

Total product The maximum quantity of output that can be produced from a given combination of inputs.

TABLE 1

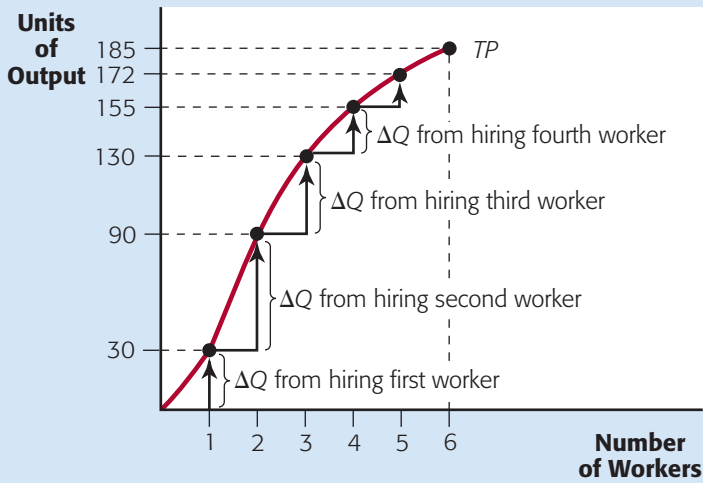
SHORT-RUN PRODUCTION
AT SPOTLESS CAR WASH

Quantity of Capital	Quantity of Labor	Total Product (Cars Washed per Day)
1	0	0
1	1	30
1	2	90
1	3	130
1	4	155
1	5	172
1	6	185

FIGURE 4

The total product (TP) curve shows the total amount of output that can be produced using various numbers of workers. The marginal product of labor (MPL) curve is the change in total product when another worker is hired. The MPL for each change in employment is indicated by the length of the vertical arrows.

TOTAL AND MARGINAL PRODUCT



Marginal product of labor The additional output produced when one more worker is hired.

The marginal product of labor (MPL) is the additional output produced when one more worker is hired. Mathematically, the marginal product of labor is the change in total product (ΔQ) divided by the change in the number of workers hired (ΔL): $MPL = \Delta Q / \Delta L$.

For example, if employment rises from 2 to 3 workers, total product rises from 90 to 130, so the marginal product of labor for *that* change in employment is $130 - 90 = 40$ units of output.

MARGINAL RETURNS TO LABOR

Look at the vertical arrows in Figure 4, which measure the marginal product of labor, and you may notice something interesting. As more and more workers are hired, the MPL first increases (the vertical arrows get longer) and then decreases (the arrows get shorter). This pattern is believed to be typical at many types of firms, so it's worth exploring.

Increasing marginal returns to labor The marginal product of labor increases as more labor is hired.

Increasing Returns to Labor. When the marginal product of labor increases as employment rises, we say there are **increasing marginal returns to labor**. Each time a worker is hired, total output rises by more than it did when the previous worker was hired. Why does this happen? One reason is that additional workers may allow production to become more specialized. Another reason is that at very low levels of employment, there may not be enough workers to properly operate the available capital. In either case, the additional worker not only produces some additional output as an individual, but also makes all other workers more productive.

At Spotless Car Wash, increasing returns to labor are observed up to the hiring of the second worker. Why? While one worker could operate the car wash alone, he or she would have to do everything: drive the cars on and off the line, towel them down, and deal with customers. Much of this worker's time would be spent switching from one task to another. The result, as we see in Table 1, is that one worker can wash only 30 cars each day. Add a second worker, though, and now

specialization is possible. One worker can collect money and drive the cars onto the line, and the other can drive them off and towel them down. Thus, with two workers, output rises all the way to 90 car washes per day; the second worker adds more to production (60 car washes) than the first (30 car washes) by making *both* workers more productive.

Diminishing Returns to Labor. When the marginal product of labor is decreasing, we say there are **diminishing marginal returns to labor**: Output rises when another worker is added, but the rise is smaller and smaller with each successive worker. Why does this happen? For one thing, as we keep adding workers, additional gains from specialization will be harder and harder to come by. Moreover, each worker will have less and less of the fixed inputs with which to work.

This last point is worth stressing. It applies not just to labor but to any variable input. In all kinds of production, if we keep increasing the quantity of any one input, while holding the others fixed, diminishing marginal returns will eventually set in. If a farmer keeps adding additional pounds of fertilizer to a fixed amount of land, the yield may continually increase, but eventually the *size* of the increase—the marginal product of fertilizer—will begin to come down. If a small bakery continues to acquire additional ovens without hiring any workers or enlarging its floor space, eventually the additional output of bread—the marginal product of ovens—will decline. This tendency is so pervasive and widespread that it has the force of a law, and economists have given that law a name:

The law of diminishing (marginal) returns states that as we continue to add more of any one input (holding the other inputs constant), its marginal product will eventually decline.

The law of diminishing returns is a physical law, not an economic one. It is based on the nature of production—on the physical relationship between inputs and outputs with a given technology. At Spotless, diminishing returns set in after two workers have been hired. Beyond this point, the firm is crowding more and more workers into a car wash with just one automated line. Output continues to increase—since there is usually *something* an additional worker can do to move the cars through the line more quickly—but the increase is less dramatic each time.

This section has been concerned with *production*—the *physical* relationship between inputs and outputs. But a more critical concern for a firm is: What will it *cost* to produce any level of output? Cost is measured in dollars and cents, not in physical units of inputs or outputs. But as you are about to see, what you've learned about production will help you understand the behavior of costs.

THINKING ABOUT COSTS

Talk to people who own or manage businesses, and it won't be long before the word *cost* comes up. People in business worry about measuring costs, controlling costs, and—most of all—reducing costs. This is not surprising: Owners want their firms to earn the highest possible profit, and costs must be subtracted from a firm's revenue to determine its profit. We will postpone a thorough discussion of profit until the next chapter. Here, we focus on just the costs of production: how economists think about costs, how costs are measured, and how they change as the firm adjusts its level of output.

Diminishing marginal returns to labor The marginal product of labor decreases as more labor is hired.

Law of diminishing marginal returns As more and more of any input is added to a fixed amount of other inputs, its marginal product will eventually decline.



Dwight Lee's "Opportunity cost and hidden invention" (<http://www.fee.org/freemen/99/9904/lee.html>) is an interesting debunking of the myth that corporations try to suppress inventions that make their products obsolete.

Let's begin by revisiting a familiar notion. In Chapter 2 you learned that economists always think of cost as *opportunity cost*—what we must give up in order to do something. This concept applies to the firm as well:

A firm's total cost of production is the opportunity cost of the owners—everything they must give up in order to produce output.

This notion—that the cost of production is its opportunity cost—is at the core of economists' thinking about costs. It can help us understand a common mistake people make when thinking about the costs of a decision.

THE IRRELEVANCE OF SUNK COSTS

Suppose you bought a used car for \$5,000 last year. During the year, you've paid \$3,000 for various repairs, and now the car is acting up again. A trustworthy mechanic tells you that the car needs a major overhaul, which will cost you \$7,000. On the other hand, he knows someone selling the same model of car—with no defects—for \$6,000. What should you do?

Some people faced with this decision might be tempted to repair the car they own. They might reason that a \$7,000 repair job is worth it, to prevent the loss of a car that has already cost them \$8,000 (\$5,000 to purchase it plus \$3,000 in repairs). But this would be faulty reasoning. The \$8,000 already paid is an example of a *sunk cost*:

A sunk cost is a cost that was paid in the past and will not change regardless of your present decision. Sunk costs should be ignored when making current decisions.

Sunk cost A cost that was incurred in the past and does not change in response to a present decision.

Why ignore sunk costs? Because they are not part of the *opportunity cost* of the action you are considering. Opportunity cost, remember, is what you must give up when you choose some action. But sunk costs have *already* been given up, so they are not part of the cost of making your choice. In the case of your car, the \$8,000 you have paid is gone, whether you buy another car or have yours repaired. The only costs that are relevant are those that *depend* on your decision and will change with it. Since you would have to pay \$7,000 to repair your car, but only \$6,000 to buy an equivalent one, you are better off giving up on your car and buying the replacement.

In many personal decisions, sunk costs are lurking in the background, tempting the decision maker to miscalculate and make a poor choice. For example, if you have completed two years of medical school and then discover you'd rather be a lawyer than a doctor, you might be tempted to stay in medical school because you have already spent so much money and time on it. But those costs you have already paid are sunk and irrelevant to your decision. The only costs that matter now are those that will *change* with your decision: the costs of your *remaining* years of medical school on the one hand, or completing three years of law school on the other.

Sunk costs should be ignored in business decisions as well. For example, South-Western Publishing Company has paid a number of costs to put this book in your hands. One of these costs was management's salaries. Suppose South-Western sells out the entire first printing and is considering whether to print another 20,000 copies. Should it consider the costs of management salaries? Absolutely not. These salaries are sunk costs and have no relevance to the decision. The only costs that matter are those that will *change* if the second printing is ordered: the costs of printing, binding, and shipping the books. Business firms, like other decision makers,



The money you've already spent on your car is a *sunk cost*, and should not influence your current decisions about repairs.

should ignore sunk costs when making choices. Only costs that are *not* sunk should enter the decision-making process.

EXPLICIT VERSUS IMPLICIT COSTS

The concept of opportunity cost also helps us classify costs into two types. Table 2 lists several different costs an owner might have to bear. On the left-hand side are the firm's **explicit costs**—instances where the firm actually pays out money for its inputs. These payments include *wages* and *salaries* for its workers, *rent* for its use of buildings and property, *interest* on any loans that were taken out to buy equipment, and payments for raw materials. Payments such as these are clearly part of the owners' opportunity cost, since the owners could have used the funds paid out to buy other valuable things.

But money payments are not the only opportunity costs to the firm. On the right-hand side of the table are some other possible costs an owner might bear. These we call **implicit costs** because, although they are indeed costs to the firm, *no money actually changes hands*. Let's consider them one at a time.

Suppose you own a restaurant, and you also happen to own the building and the land underneath. You don't have to pay any rent, so under "rent paid out," your explicit cost would be zero. Does this mean that the building and the land are free? To an accountant—who focuses on actual money payments—the answer is yes. But to an economist—who thinks of opportunity cost—the answer is *absolutely not*. By choosing to use your land and building for your restaurant, you are sacrificing the opportunity to rent them to someone else. This *foregone rent* is an implicit cost, and it is as much a cost of production as the rent you would pay if someone else owned the building. In both cases, something is given up to produce your output.

Now suppose that instead of borrowing the money to start up your restaurant—to buy ovens, dishes, tables, chairs, and an initial inventory of food—you used your *own* money. You therefore have no debts, and no interest to pay on them, so your interest on loans is zero. But there is still a cost to be considered: You *could* have put your money in a bank account, lent it to someone else, or invested it elsewhere. In any of these cases, you would have earned *investment income* on your money. Economists measure the opportunity cost of funds you invest in a business as the income you *could* have earned on these funds by investing them elsewhere. This *foregone investment income* is an implicit cost of doing business.

Finally, suppose you decide to manage your restaurant yourself. Have you escaped the costs of hiring a manager? Not really, because you are still bearing an opportunity cost: You could have done something else with your time. We measure the value of your time as the income you *could* have earned by devoting your labor to your next-best income-earning activity. This *foregone labor income*—the wage or salary you could be earning elsewhere—is an implicit cost of your business, and therefore part of its opportunity cost.

Explicit costs Money actually paid out for the use of inputs.

Implicit costs The cost of inputs for which there is no direct money payment.

TABLE 2	
Explicit Costs	Implicit Costs
Rent paid out	Opportunity cost of:
Interest on loans	Owner's land (rent foregone)
Managers' salaries	Owner's money (investment income foregone)
Hourly workers' wages	Owner's time (labor income foregone)
Cost of raw materials	

A FIRM'S COSTS

COSTS IN THE SHORT RUN

Remember that, in the short run, one or more of the firm's inputs is fixed. No matter how much output is produced, the quantity of these fixed inputs remains the same. Other inputs, by contrast, can be varied as output changes. Because the firm has these two different types of inputs in the short run, it will also face two different types of costs.

Fixed costs Costs of fixed inputs.

The costs of a firm's fixed inputs are called, not surprisingly, **fixed costs**. Like the fixed inputs themselves, fixed costs remain the same no matter what the level of output. In most businesses, we can treat rent and interest—whether explicit or implicit—as fixed costs, since producing more or less output in the short run will not cause any of these costs to change. Managers typically refer to these costs as their *overhead costs*, or simply, overhead.

Variable costs Costs of variable inputs.

The costs of obtaining the firm's variable inputs are its **variable costs**. These costs, like the usage of variable inputs themselves, will rise as output increases. In most businesses, we treat the wages of hourly employees and the costs of raw materials as variable costs, since quantities of both labor and raw materials can usually be adjusted rather rapidly.

MEASURING SHORT-RUN COSTS

In Table 3, we return to our mythical firm—Spotless Car Wash—and ask: What happens to *costs* as output changes in the short run. The first three columns of the table give the relationship between inputs and outputs—the production function—just as in Table 1, which was discussed earlier. But there is one slight difference: In Table 3, we've reversed the order of the columns, putting total output first. We are changing our perspective slightly: Now we want to observe how a change in the quantity of *output* causes the firm's *inputs*—and therefore its *costs*—to change.

In addition to Spotless's production function, we need to know one more thing before we can analyze its costs: what it must *pay* for its inputs. In Table 3, the price

TABLE 3

SHORT-RUN COSTS FOR SPOTLESS CAR WASH

(1) Output (per Day)	(2) Capital	(3) Labor	(4) TFC	(5) TVC	(6) TC	(7) MC	(8) AFC	(9) AVC	(10) ATC
0	1	0	\$75	\$ 0	\$ 75		—	—	—
30	1	1	\$75	\$ 60	\$135	\$2.00	\$2.50	\$2.00	\$4.50
90	1	2	\$75	\$120	\$195	\$1.00	\$0.83	\$1.33	\$2.17
130	1	3	\$75	\$180	\$255	\$1.50	\$0.58	\$1.38	\$1.96
155	1	4	\$75	\$240	\$315	\$2.40	\$0.48	\$1.55	\$2.03
172	1	5	\$75	\$300	\$375	\$3.53	\$0.44	\$1.74	\$2.18
185	1	6	\$75	\$360	\$435	\$4.62	\$0.41	\$1.95	\$2.35

of labor is set at \$60 per worker per day, and the price of each automated car-washing line at \$75 per day.

How do Spotless's short-run costs change as its output changes? Get ready, because there are a surprising number of different ways to answer that question, as illustrated in the remaining columns of Table 3.

Total Costs. Columns 4, 5, and 6 in the table show three different types of total costs. In column 4, we have Spotless's **total fixed cost (TFC)**—the cost of all inputs that are fixed in the short run. Like the quantity of fixed inputs themselves, fixed costs remain the same no matter what the level of output. For Spotless Car Wash, the daily cost of renting or owning one automated line is \$75, so total fixed cost is \$75. Running down the column, you can see that this cost—because it is fixed—remains the same no matter how many cars are washed each day.

Column 5 shows **total variable cost (TVC)**—the cost of all variable inputs. For Spotless, labor is the only variable input. As output increases, more labor will be needed, so *TVC* will rise. For example, to wash 90 cars each day requires 2 workers, and each worker must be paid \$60 per day, so *TVC* will be $2 \times \$60 = \120 . But to wash 130 cars requires 3 workers, so *TVC* will rise to $3 \times \$60 = \180 .

Finally, column 6 shows **total cost (TC)**—the sum of all fixed and variable costs:

$$TC = TFC + TVC.$$

For example, at 90 units of output, $TFC = \$75$ and $TVC = \$120$, so $TC = \$75 + \$120 = \$195$. Because total variable cost rises with output, total cost rises as well.

Now look at Figure 5, where we've graphed all three total cost curves for Spotless Car Wash. Both the *TC* and *TVC* curves slope upward—since these costs increase along with output. Notice that there are *two* ways in which *TFC* is represented in the graph. One is the *TFC* curve, which is a horizontal line, since *TFC* has the same value at any level of output. The other is the *vertical distance* between the rising *TVC* and *TC* curves, since *TFC* is always the *difference* between *TVC* and *TC*. In the graph, this vertical distance must remain the same, at \$75, no matter what the level of output.

Total fixed cost The cost of all inputs that are fixed in the short run.

Total variable cost The cost of all variable inputs used in producing a particular level of output.

Total cost The costs of all inputs—fixed and variable.

THE FIRM'S TOTAL COST CURVES

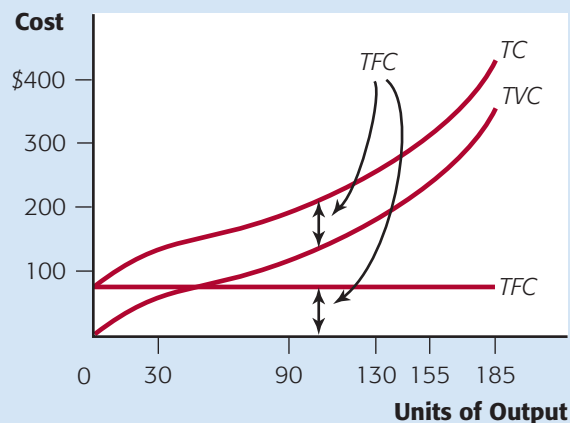


FIGURE 5

At any level of output, total cost (*TC*) is the sum of total fixed cost (*TFC*) and total variable cost (*TVC*).

Average Costs. While total costs are important, sometimes it is more useful to track a firm's costs *per unit* of output, which we call its *average cost*. There are three different types of average cost, each obtained from one of the total cost concepts just discussed.

Average fixed cost Total fixed cost divided by the quantity of output produced.

The firm's **average fixed cost (AFC)** is its total fixed cost divided by the quantity (Q) of output:

$$E_{x,y} = \frac{\%Q_x^D}{\%\Delta P_y}$$

No matter what kind of production or what kind of firm, *AFC* will always fall as output rises. Why? Because *TFC* remains constant, so a rise in Q *must* cause the ratio TFC/Q to fall. Business managers often refer to this decline in *AFC* as “spreading their overhead” over more output. For example, a restaurant has overhead costs for its buildings, furniture, and cooking equipment. The more meals it serves, the lower will be its overhead cost per meal. Does *AFC* fall with output at Spotless Car Wash? Look at Table 3 again. When output is 30 units, *AFC* is $\$75/30 = \2.50 . But at 90 units of output, *AFC* drops to $\$75/90 = \0.83 . And *AFC* keeps declining as we continue down the column. The more output produced, the lower is fixed cost per unit of output.

Average variable cost Total variable cost divided by the quantity of output produced.

Average variable cost (AVC)—in column 9 of Table 3—is the cost of the variable inputs per unit of output:

$$E_{x,y} = \frac{\%Q_x^D}{\%\Delta P_y}$$

For example, at 30 units of output, $TVC = \$60$, so $AVC = TVC/Q = \$60/30 = \2.00 .

What happens to *AVC* as output rises? Based on mathematics alone, we can't be sure. On the one hand, a rise in Q raises the denominator of the fraction TVC/Q . On the other hand, *TVC* increases, so the numerator rises as well. Thus, it's possible for *AVC* to either rise or fall, depending on whether *TVC* or Q rises by a greater percentage. But if you run your finger down the *AVC* column in Table 3, you'll see a pattern: The *AVC* numbers first decrease and then increase. Economists believe that this pattern of decreasing and then increasing average variable cost is typical at many firms. When plotted in Figure 6, this pattern causes the *AVC* curve to have a U shape. We'll discuss the reason for this characteristic U shape a bit later.

Average total cost Total cost divided by the quantity of output produced.

Average total cost (ATC)—shown in column 10—is the total cost per unit of output:

$$E_{x,y} = \frac{\%Q_x^D}{\%\Delta P_y}$$

For example, at 90 units of output, $TC = \$195$, so $ATC = TC/Q = \$195/90 = \2.17 . As output rises, *ATC*, like *AVC*, can either rise or fall, since both the numerator and denominator of the fraction TC/Q rises. But we usually expect *ATC*, like *AVC*, to first decrease and then increase, so the *ATC* curve will also be U-shaped. However—as you can see in Figure 6—it is not identical to the *AVC* curve. At each level of output, the vertical distance between the two curves is average *fixed cost* (*AFC*). Since *AFC* declines as output increases, the *ATC* curve and the *AVC* curve must get closer and closer together as we move rightward.

Marginal Cost. The total and average costs we've considered so far tell us about the firm's cost at a particular *level* of output. For many purposes, however, we are more interested in how cost *changes* when output *changes*. This information is provided by another cost concept:

AVERAGE AND MARGINAL COSTS

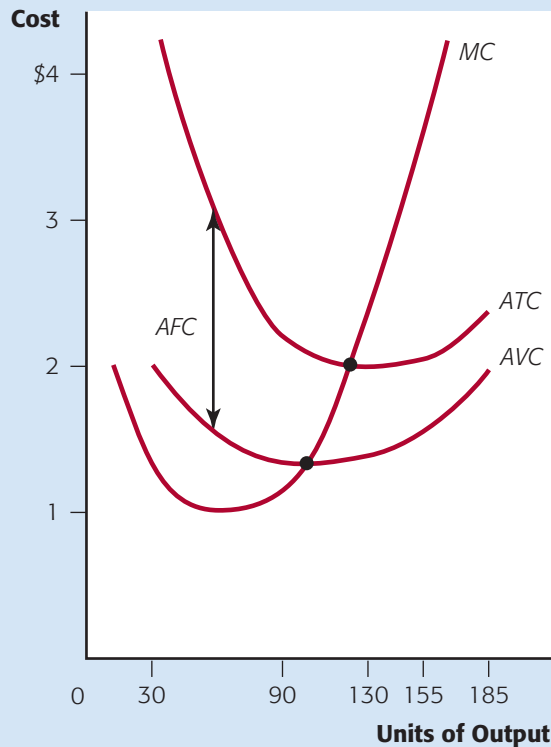


FIGURE 6

Average variable cost (AVC) and average total cost (ATC) are U-shaped, first decreasing and then increasing. Average fixed cost (AFC)—the vertical distance between ATC and AVC—becomes smaller as output increases.

The marginal cost (MC) curve is also U-shaped, reflecting first increasing and then diminishing marginal returns to labor. MC passes through the minimum points of both the AVC and ATC curves.

Marginal cost (MC) is the increase in total cost from producing one more unit of output. Mathematically, MC is calculated by dividing the change in total cost (ΔTC) by the change in output (ΔQ):

$$MC = \frac{\Delta TC}{\Delta Q}$$

Marginal cost The increase in total cost from producing one more unit of output.

For Spotless Car Wash, marginal cost is entered in column 7 of Table 3 and graphed in Figure 6. Since marginal cost tells us what happens to total cost when output *changes*, the entries in the table are placed *between* one output level and another. For example, when output rises from 90 to 130, total cost rises from \$195 to \$255. For this change in output, we have $\Delta TC = \$255 - \$195 = \$60$, while $\Delta Q = 40$, so $MC = \$60/40 = \1.50 . This entry is listed *between* the output levels 90 and 130 in the table and plotted *between* them in Figure 6.

EXPLAINING THE SHAPE OF THE MARGINAL COST CURVE

As you can see in Table 3 (and also in Figure 6), MC first declines and then rises. Why is this? Here, we can use what we learned earlier about marginal returns to labor. At low levels of employment and output, there are increasing marginal returns to labor: $MPL = \Delta Q/\Delta L$ is rising. That is, each worker hired adds more to production than the worker before. But that means that *fewer additional workers are needed to produce an additional unit of output*. Now, since additional labor is the

firm's cost of increasing production, the cost of an additional unit of output (MC) must be falling. Thus, as long as MPL is rising, MC must be falling.

For Spotless, since MPL rises when employment increases from zero to one and then one to two workers, MC must fall as the firm's output rises from zero to 30 units (produced by one worker) and then from 30 to 90 units (produced by two workers).

At higher levels of output, we have the opposite situation: Diminishing marginal returns set in and the marginal product of labor ($\Delta Q/\Delta L$) falls. Therefore, additional units of output require *more* and *more* additional labor. As a result, each additional unit of output costs more and more to produce. Thus, as long as MPL is falling, MC must be rising.

For Spotless, diminishing marginal returns to labor occur for all workers beyond the second, so MC rises for all output levels beyond 90 (the amount produced by two workers).

To sum up:

When the marginal product of labor (MPL) rises, marginal cost (MC) falls. When MPL falls, MC rises. Since MPL ordinarily rises and then falls, MC will do the opposite—it will fall and then rise. Thus, the MC curve is U-shaped.

THE RELATIONSHIP BETWEEN AVERAGE AND MARGINAL COSTS

Although marginal cost and average cost are not the same, there is an important relationship between them. Look again at Figure 6 and notice that all three curves— MC , AVC , and ATC —first fall and then rise, but not all at the same time. The MC curve bottoms out before either the AVC or ATC curve. Further, the MC curve intersects each of the average curves *at their lowest points*. These graphical features of Figure 6 are no accident; indeed, they follow from the laws of mathematics. To understand this, let's consider a related example with which you are probably more familiar.

An Example: Average and Marginal Test Scores. Suppose you take five tests in your economics course during the term, with the results listed in Table 4. To your immense pleasure, you score 100 on your first test. Your total score—the total number of points you have received thus far during the term—is 100. Your marginal score—the *change* in your total caused by the most recent test—will also be 100, since your total rose from 0 to 100. Your average score so far is 100 as well.

TABLE 4

AVERAGE AND MARGINAL TEST SCORES	Number of Tests Taken	Total Score	Marginal Score	Average Score
	0	0		—
	1	100	100	100
	2	150	50	75
	3	210	60	70
	4	280	70	70
	5	360	80	72

Now suppose that, for the second test, you forget to study actively. Instead, you just read the text while simultaneously watching music videos and eavesdropping on your roommate's phone conversations. As a result, you get a 50. Your marginal score is 50. Since this score is lower than your previous average of 100, the second test will pull your average down. Indeed, whenever you score lower than your previous average, you will always decrease the average. In the table, we see that your average after the second test falls to 75.

Now you start to worry, so you turn off the TV while studying, and your performance improves a bit: You get a 60. Does the improvement in your score—from 50 to 60—increase your *average* score? Absolutely not. Your average will decrease once again, because your *marginal* score of 60 is *still* lower than your previous average of 75. As we know, when you score lower than your average, it pulls the average down—even if you're improving. In the table, we see that your average now falls to 70.

For your fourth exam, you study a bit harder and score a 70. This time, since your score is precisely *equal* to your previous average, the average remains unchanged at 70.

Finally, on your fifth and last test, your score improves once again, this time to 80. This time, you've scored *higher* than your previous average, pulling your average up from 70 to 72.

This example may be easy to understand because you are used to figuring out your average score in a course as you take additional exams. But the relationship between marginal and average spelled out here is universal—it is the same for grade point averages, batting averages—and costs.

Average and Marginal Cost. Now let's apply our previous discussion to a firm's cost curves. Whenever marginal cost is below average cost, we know that the cost of producing *one more* unit of output is *less* than the average cost of all units produced so far. Therefore, producing one more unit will bring the average down. That is, when marginal cost is below average cost, average cost will come down. This applies to both average *variable* cost and average *total* cost.

For example, when Spotless is producing 30 units of output, its *ATC* is \$4.50 and its *AVC* is \$2.00 (See Table 3 on p. 166). But if it increases output from 30 to 90 units, the marginal cost of these *additional* units is just \$1.00. Since *MC* is less than both *ATC* and *AVC* for this change, it pulls both averages down. Graphically, when the *MC* curve lies below one of the average curves (*ATC* or *AVC*), that average curve will slope downward.

Now consider a change in output from 90 units to 130 units. Marginal cost for this change is \$1.50. But the *AVC* at 90 units is \$1.33. Since *MC* is greater than *AVC*, this change in output will pull *AVC* up. Accordingly, the *AVC* curve begins to slope upward. However, *ATC* at 90 units is \$2.17. Since *MC* is still *less* than *ATC*, the *ATC* curve will continue to slope downward.

Finally, consider the change from 130 to 155 units. For this change in output, *MC* is \$2.40, which is greater than the previous values of both *AVC* (\$1.38) and *ATC* (\$1.96). If the firm makes this move, both *AVC* and *ATC* will rise.

Now, let's put together what we know about marginal cost and what we know about the relationship between marginal and average cost. Remember that marginal cost drops rapidly when the firm begins increasing output from low levels of production, due to increasing marginal returns to labor. Thus, *MC* will initially drop *below* *AVC* and *ATC*, pulling these averages down. But if the firm keeps increasing its output, diminishing returns to labor will set in. *MC* will keep on rising, until it *exceeds* *AVC* and *ATC*. Once this happens, further increases in output will *raise* both *AVC* and *ATC*.

When we state this argument in terms of the curves graphed in Figure 6, we can finally understand why the *AVC* and *ATC* curves are U-shaped.

At low levels of output, the MC curve lies below the AVC and ATC curves, so these curves will slope downward. At higher levels of output, the MC curve will rise above the AVC and ATC curves, so these curves will slope upward. Thus, as output increases, the average curves will first slope downward and then slope upward. That is, they will have a U shape.

There is one more important observation to make before we leave the short run. We've just seen that whenever the *MC* curve lies *below* the *ATC* curve, *ATC* is falling. But when the *MC* curve crosses the *ATC* curve and rises *above* it, *ATC* will be rising. As a result, the *MC* curve must intersect the *ATC* curve at its *minimum* point, as it does in Figure 6. And the same is true of the *AVC* curve.

The MC curve will intersect the minimum points of the AVC and ATC curves.

TIME TO TAKE A BREAK

By now, your mind may be swimming with concepts and terms: total, average, and marginal cost curves; fixed and variable costs; explicit and implicit costs. . . . We are covering a lot of ground here and still have a bit more to cover: production and cost in the *long run*.

As difficult as it may seem to keep these concepts straight, they will become increasingly easy to handle as you use them in the chapters to come. But it's best not to overload your brain with too much new material at one time. So if this is your first trip through this chapter, now is a good time for a break. Then, when you're fresh, come back and review the material you've read so far. When the terms and concepts start to feel familiar, you are ready to move on to the long run.

PRODUCTION AND COST IN THE LONG RUN

Most of the business firms you have contact with—such as your supermarket, the stores where you buy new clothes, your telephone company, and your internet service provider—plan to be around for quite some time. They have a long-term planning horizon, as well as a short-term one. But so far, we've considered the behavior of costs only in the short run.

In the long run, costs behave differently, because the firm can adjust *all* of its inputs in any way it wants:

In the long run, there are no fixed inputs or fixed costs; all inputs and all costs are variable. The firm must decide what combination of inputs to use in producing any level of output.

How will the firm choose? Its goal is to earn the highest possible profit, and to do this, it must follow the *least cost rule*:

To produce any given level of output, the firm will choose the input mix with the lowest cost.



<http://>

Examples of economies of scale can be found at <http://bized.ac.uk/stafsup/options/notes/econ204.htm>. For an interesting discussion of some physical and biological sources of economies of scale, read Robert Pool's "Why nature loves economies of scale" at <http://www.newscientist.com/ns/970412/scales.html>

Identify Goals and Constraints



Let's apply the least cost rule to Spotless Car Wash. Suppose we want to know the cost of washing 185 cars per day. In the short run, of course, Spotless does not have to worry about how it would produce this level of output: It is stuck with one automated line, and the only way to wash 185 cars is to hire six workers (see Table 3 on p. 166). Total cost in the short run will be $6 \times \$60 + \$75 = \$435$.



When you read the *least cost rule* of production, you might begin to think that the firm's goal is to have the *least possible cost*. But this is not true. To convince yourself, just realize that the least possible cost would be zero, and in the long run this could be achieved by not using any inputs and producing nothing!

The least cost rule says that any *given* level of output should be produced at the lowest possible cost. The firm's goal is to maximize *profit*, and the least cost rule helps it do that. For example, if the firm is considering producing 10 units of output, and there are two ways to produce that number of units—one costing \$6,000 and the other costing \$5,000—the firm should always choose the latter because it is cheaper. If it chose the former, and it ended up producing 10 units of output, it would not be earning the highest possible profit. But notice that \$5,000 is not the "lowest possible cost" for the firm; *it is the lowest possible cost for producing 10 units*.

In the long run, however, Spotless can vary the number of automated lines as well as the number of workers. Its *long-run* production function will tell us all the different combinations of *both* inputs that can be used to produce any output level. Suppose four different input combinations can be used to wash 185 cars per day. These are listed in Table 5. Combination A uses the least capital and the most labor—no automated lines at all and nine workers washing the cars by hand. Combination D uses the most capital and the least labor—three automated lines with only three workers. Since each automated line costs \$75 per day and each worker costs \$60 per day, it is easy to calculate the cost of each production method. Spotless will choose the one with the lowest cost—combination C, with two automated lines and 4 workers, for a total cost of \$390 per day.

Retracing our steps, we have found that if Spotless wants to wash 185 cars per day, it will examine the different methods of doing so and select the one with the least cost. Once it has determined the cheapest production method, the other, more expensive methods can be ignored.

Table 6 shows the results of going through this procedure for several different levels of output. The second column, **long-run total cost (LRTC)**, tells us the cost of producing each quantity of output *when the least-cost input mix is chosen*. For each output level, different production methods are examined, the cheapest one is chosen, and the others are ignored. Notice that the LRTC of zero units of output is \$0. This will always be true for any firm. In the long run, all inputs can be adjusted as the firm wishes, and the cheapest way to produce zero output is to use *no* inputs at all. (For comparison, what is the *short-run* total cost of producing zero units? Why can it never be \$0?)

The third column in Table 6 gives the **long-run average total cost (LRATC)**, the cost per unit of output in the long run:

$$E_{x,y} = \frac{\%Q_x^D}{\%\Delta P_y}$$

Long-run total cost The cost of producing each quantity of output when the least-cost input mix is chosen in the long run.

Long-run average total cost The cost per unit of output in the long run, when all inputs are variable.

TABLE 5

FOUR WAYS TO WASH 185 CARS PER DAY

Method	Quantity of Capital	Quantity of Labor	Cost
A	0	9	\$540
B	1	6	\$435
C	2	4	\$390
D	3	3	\$405

TABLE 6

LONG-RUN COSTS FOR SPOTLESS CAR WASH

Output	LRTC	LRATC
0	\$ 0	—
30	\$ 100	\$3.33
90	\$ 195	\$2.17
130	\$ 255	\$1.96
155	\$ 315	\$2.03
172	\$ 360	\$2.09
185	\$ 390	\$2.11
200	\$ 450	\$2.25
250	\$ 650	\$2.60
300	\$1,200	\$4.00

Long-run average total cost is similar to average total cost, which was defined earlier. Both are obtained by dividing total cost by the level of output. There is one important difference, however: To calculate *ATC*, we used total cost (*TC*), which pertains to the short run, in the numerator. In calculating *LRATC*, we use *long-run* total cost (*LRTC*) in the numerator. Thus, *LRATC* tells us the cost per unit when the firm can vary *all* of its inputs and always chooses the cheapest input mix possible. *ATC*, however, tells us the cost per unit when the firm is stuck with some collection of fixed inputs.

THE RELATIONSHIP BETWEEN LONG-RUN AND SHORT-RUN COSTS

If you compare Table 6 (long run) with Table 3 (short run), you will see something important: For some output levels, *LRTC* is smaller than *TC*. For example, Spotless can wash 185 cars for an *LRTC* of \$390. But earlier, we saw that in the short run, the *TC* of washing these same 185 cars was \$435. To understand the reason for this difference, look back at Table 5, which lists the four different ways of washing 185 cars per day. In the short run, the firm is stuck with just one automated line, so its only option is method *B*. In the long run, however, the firm can adjust *all* of its inputs, so it can choose any of the four methods of production, including method *C*, which is cheapest. In many cases, the freedom to choose among different production methods enables the firm to select a cheaper input mix in the long run than it can in the short run. Thus, in the long run, the firm may be able to save money.

But not always. At some output levels, the freedom to adjust all inputs doesn't save the firm a dime. To wash 130 cars, for example, the long-run cost—the cost when using the cheapest input mix—is the same as the short-run total cost ($LRTC = TC = \$255$). For this output level, it must be that the *short-run* input mix is also the least-cost input mix. Thus, if Spotless wants to wash 130 cars per day, it would choose in the long run the same production method it is already using in the short run. At this output level, the firm could not save money by adjusting its capital in the long run. (There are other output levels listed in the tables for which $LRTC = TC$. Can you find them?)

What we have found for Spotless Car Wash is true for all firms:

Long-run total cost of producing a given level of output can be less than or equal to, but never greater than, short-run total cost ($LRTC \leq TC$).

We can also state this relationship in terms of *average costs*. That is, we can divide both sides of the inequality by Q and obtain $LRTC/Q \leq TC/Q$. Using our definitions, this translates to $LRATC \leq ATC$.

Long-run average cost of producing a given level of output can be less than or equal to, but never greater than, short-run average total cost ($LRATC \leq ATC$).

Average Cost and Plant Size. Often, economists refer to the collection of fixed inputs at the firm's disposal as its **plant**. For example, the plant of a computer manufacturer such as Compaq would consist of its factory building and the assembly lines inside it. The plant of the Hertz car-rental company would include all of its automobiles and rental offices. For Spotless Car Wash, we've assumed that the plant is simply the company's capital equipment—the automated lines for washing cars. If Spotless were to add to its capital, then each time it acquired another automated line, it would have a different—and larger—plant. Viewed in this way, we can distinguish between the long run and the short run as follows: *In the long run, the firm can change the size of its plant; in the short run, it is stuck with its current plant.*

Now think about the ATC curve, which tells us the firm's average total cost in the short run. This curve is always drawn for a specific plant. That is, the ATC curve tells us how average cost behaves in the short run, *when the firm uses a plant of a given size*. If the firm had a different size plant, it would be moving along a different ATC curve. In fact, there is a different ATC curve for each different plant the firm could have. In the long run, then, the firm can choose on which ATC curve it wants to operate. And, as we know, to produce any level of output, it will always choose that ATC curve—among all of the ATC curves available—that enables it to produce at lowest possible average total cost. This insight tells us how we can graph the firm's $LRATC$ curve.

Graphing the $LRATC$ Curve. Look at Figure 7, which shows several different ATC curves for Spotless Car Wash. There is a lot going on in this figure, so let's take it one step at a time. First, find the curve labeled ATC_1 . This is our familiar ATC curve—the same one shown in Figure 6—which we used to find Spotless's average total cost in the short run, when it was stuck with one automated line. The other ATC curves refer to *different* plants that Spotless *might* have had instead. For example, the curve labeled ATC_0 shows how average total cost would behave if Spotless had a plant with *zero* automated lines; ATC_2 shows average total cost with *two* automated lines, and so on. Since, in the long run, the firm can choose which size plant to operate, it can also choose on which of these ATC curves it wants to operate. And, as we know, in the long run, it will always choose the plant with the lowest possible average total cost.

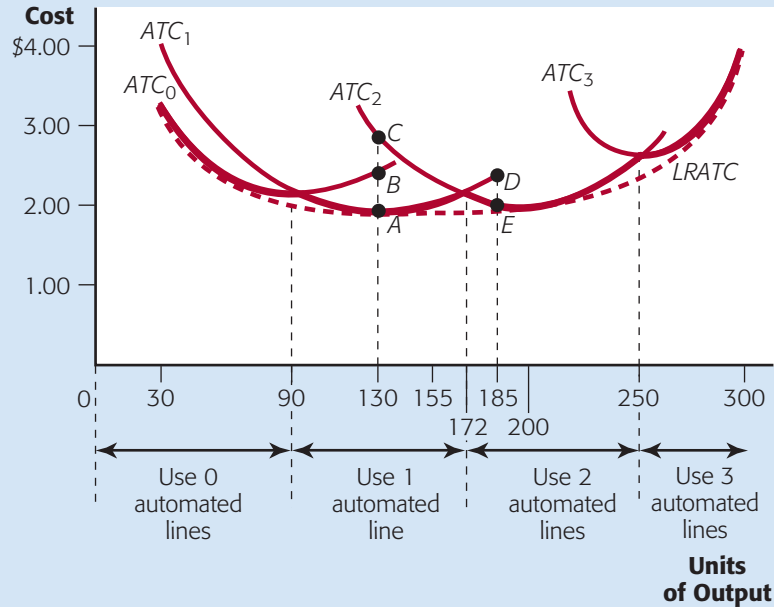
Let's take a specific example. Suppose that Spotless thinks that it might wash 130 cars per day. In the long run, what size plant should it choose? Scanning the different ATC curves in Figure 7, we see that the lowest possible per-unit cost—\$1.96 per car—is at point A along ATC_1 . The best plant for washing 130 cars per day, therefore, will have just one automated line. For this output level, Spotless would never choose a plant with zero lines, since it would then have to operate on ATC_0 at point B . Since point B is higher than point A , we know that point B represents a larger per-unit cost. Nor would the firm choose a plant with two lines—operating on ATC_2 at point C —for this would mean a still larger per-unit cost. Of all the possibilities, only point A along ATC_1 enables Spotless to achieve the lowest per-unit cost for washing 130 cars. Thus, to produce 130 units of output in the long run, Spotless would choose to operate at point A on ATC_1 . Thus, point A is the $LRATC$ of 130 units.

Plant The collection of fixed inputs at a firm's disposal.

FIGURE 7

Average-total cost curves ATC_0 , ATC_1 , ATC_2 , and ATC_3 show average costs when the firm has zero, one, two, and three production lines, respectively. The $LRATC$ curve combines portions of all the firm's ATC curves. The firm will choose the lowest-cost ATC curve for each level of output.

LONG-RUN AVERAGE TOTAL COST



Now, suppose instead that Spotless wanted to produce 185 units of output in the long run. A plant with one automated line is no longer the best choice. Instead, the firm would choose a plant with *two* automated lines. How do we know? For an output of 185, the firm could choose point D on ATC_1 , or point E on ATC_2 . Since point E is lower, it is the better choice. At this point, average total cost would be \$2.11, so this would be the $LRATC$ of 185 units.

Continuing in this way, we could find the $LRATC$ for *every* output level Spotless might produce. To produce any given level of output, the firm will always operate on the *lowest* ATC curve available. As output increases, it will move along an ATC curve until another, lower ATC curve becomes available—one with lower costs. At that point, the firm will increase its plant size, so it can move to the lower ATC curve. For example, as Spotless increases its output level from 90 to 172 units of output, it will continue to use a plant with one automated line and move along ATC_1 . But if it wants to produce *more* than 172 units in the long run, it will increase its plant to *two* automated lines and begin moving along ATC_2 .

Thus, we can trace out Spotless's $LRATC$ curve by combining just the lowest portions of all the ATC curves from which the firm can choose. In Figure 7, this is the thick, scallop-shaped curve.

A firm's $LRATC$ curve combines portions of each ATC curve available to the firm in the long run. For each output level, the firm will always choose to operate on the ATC curve with the lowest possible cost.

Figure 7 also gives us a view of the different options facing the firm in the short run and the long run. Once Spotless builds a plant with one automated line, its options in the short run are limited—it can only move along ATC_1 . If it wants to increase its output from 130 to 185 units, it must move from point A to point D . But in the long run, it can move along its $LRATC$ curve—from point A to point E —by changing the size of its plant.

More generally,

in the short run, a firm can only move along its current ATC curve. In the long run, however, it can move from one ATC curve to another by varying the size of its plant. As it does so, it will also be moving along its LRATC curve.

EXPLAINING THE SHAPE OF THE LRATC CURVE

In Figure 7, the LRATC curve has a scalloped look because the firm can only choose among four different plants. But many firms—especially large ones—can choose among hundreds or even thousands of different plant sizes. Each plant would be represented by a different ATC curve, so there would be hundreds of ATC curves crowded into the figure. As a result, the scallops would disappear, and the LRATC curve would appear as a smooth curve, like the dashed line in Figure 7.

In Figure 8, which reproduces this smoothed-out LRATC curve, you can see that the curve is U-shaped—much like the AVC and ATC curves you learned about earlier. That is, as output increases, long-run average costs first decline, then remain constant, and finally rise. Although there is no law or rule of logic that requires an LRATC curve to have all three of these phases, in many industries this seems to be the case. Let’s see why, by considering each of the three phases in turn.

Economies of Scale. When an increase in output causes LRATC to decrease, we say that the firm is enjoying **economies of scale**: the more output produced, the lower the cost per unit.

On a purely mathematical level, economies of scale mean that long-run total cost is rising by a smaller proportion than output. For example, if a doubling of

Economies of scale Long-run average total cost decreases as output increases.

THE SHAPE OF LRATC

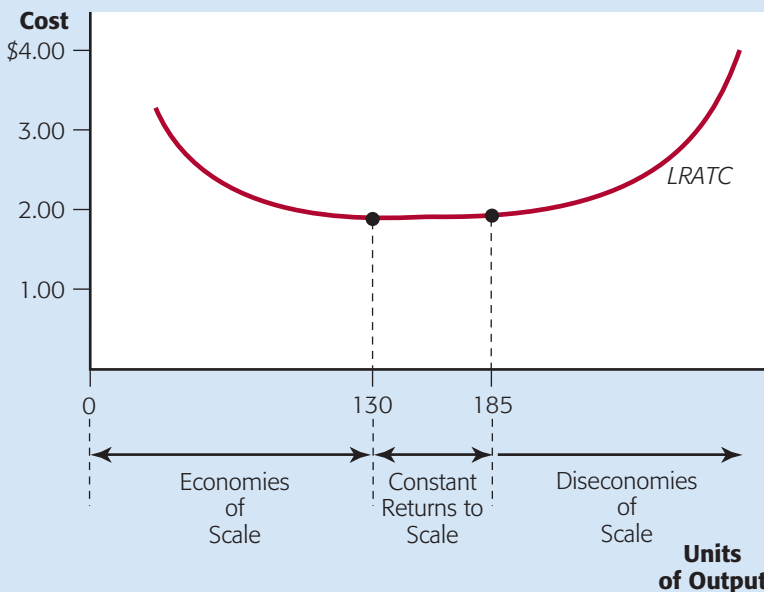


FIGURE 8

If long-run total cost rises proportionately less than output, production reflects economies of scale, and LRATC slopes downward. If cost rises proportionately more than output, there are diseconomies of scale, and LRATC slopes upward. Between those regions, cost and output rise proportionately, yielding constant returns to scale.

output (Q) can be accomplished with less than a doubling of costs, then the ratio $LRTC/Q = LRATC$ will decline, and—voilà!—economies of scale.

When long-run total cost rises proportionately less than output, production is characterized by economies of scale, and the LRATC curve slopes downward.

So much for the mathematics. But in the real world, *why* should total costs ever increase by a smaller proportion than output? Why should a firm experience economies of scale?

Gains from Specialization. One reason for economies of scale is gains from specialization. At very low levels of output, workers may have to perform a greater variety of tasks, slowing them down and making them less productive. But as output increases and workers are added, more possibilities for specialization are created. At Spotless, an increase in output and employment might permit one worker to specialize in taking cash from customers, a second to drive the cars onto the line, a third to towel them down, a fourth to work on advertising, and so on. Since each worker is more productive, output will increase by a greater proportion than costs.

The greatest opportunities for increased specialization occur when a firm is producing at a relatively low level of output, with a relatively small plant and small workforce. Thus, economies of scale are more likely to occur at lower levels of output.

More Efficient Use of Lumpy Inputs. Another explanation for economies of scale involves the “lumpy” nature of many types of plant and equipment. By this, we mean that some types of inputs cannot be increased in tiny increments, but rather must be increased in large jumps.

A doctor, for example, needs the use of an X-ray machine in order to serve her patients. Unless she can share with other doctors (which may not be possible), she must buy one or more *whole* machines—she cannot buy a half or a fifth of an X-ray machine. Suppose a single machine can service up to 500 patients per month and costs \$2,000 per month (in interest payments or foregone investment income). Then the more patients the doctor sees (up to 500), the lower will be the cost of the machine per patient. For example, if she sees 100 patients each month, the cost per patient will be $\$2,000/100 = \20 . If she sees 500 patients, the cost per patient drops to $\$2,000/500 = \4 . If much of the doctor’s plant and equipment are lumpy in this way, her *LRATC* curve might continue to decline over some range of output.

We see this phenomenon in many types of businesses: Plant and equipment must be purchased in large lumps, and a low cost per unit is achieved only at high levels of output. If you decide to start a pizza delivery business on campus, you will have to purchase or rent at least one pizza oven. If you can make 200 pizzas per day with a single oven, then your total oven costs will be the same whether you bake 1, 10, 50, 100, or 200 pizzas. The more pizzas you make, the lower will be your oven costs *per pizza*.

Other inputs besides equipment can also be lumpy in this way. Restaurants must pay a yearly license fee and are not permitted to buy part of a license if their output is small. An answering service must have a receptionist on duty at all times, even if only a few calls come in each day. A theater must have at least one ticket seller and one projectionist, regardless of how many people come to see the show. In all of these cases, an increase in output allows the firm to spread the cost of lumpy inputs over greater amounts of output, lowering cost *per unit of output*.

Making more efficient use of lumpy inputs will have more impact on *LRATC* at low levels of output when these inputs make up a greater proportion of the firm’s to-

tal costs. At higher levels of output, the impact is smaller. For example, suppose a restaurant must pay a yearly license fee of \$1,000. If output doubles from 1,000 to 2,000 meals per year, license costs per meal served will fall from \$1 to \$0.50. But if output doubles from 10,000 to 20,000 meals, license costs per meal drop from \$0.10 to \$0.05—a hardly noticeable difference. Thus, spreading lumpy inputs across more output—like the gains from specialization—is more likely to create economies of scale at relatively low levels of output. This is another reason why the typical *LRATC* curve—as illustrated in Figure 8—slopes downward at relatively low levels of output.

A look at Table 6 on p. 174 tells us that there are, indeed, economies of scale for Spotless at low levels of output. It costs \$100 to wash 30 cars and \$195 to wash 90 cars. So as output triples from 30 to 90, costs increase by only \$95/\$100, or 95 percent, so *LRATC* falls. Spotless is clearly enjoying economies of scale. Indeed, Figure 8 shows that it will experience economies of scale for all output levels up to 130 units.

Diseconomies of Scale. As output continues to increase, most firms will reach a point where bigness begins to cause problems. This is true even in the long run, when the firm is free to increase its plant size as well as its workforce. For example, a large firm may require more layers of management, so communication and decision making become more time consuming and costly. It may also be more difficult to screen out misfits among new hires and to monitor those already working at the firm, so there is an increase in mistakes, shirking of responsibilities, and even theft from the firm. All of these problems contribute to rises in *LRTC* as output increases, and so work in the opposite direction to the forces helping to create economies of scale. Eventually, these problems may become so serious that a doubling of output will cause *more* than a doubling of total cost. When this happens, *LRATC* will rise. More generally,

when long-run total cost rises more than in proportion to output, there are diseconomies of scale, and the LRATC curve slopes upward.

While economies of scale are more likely at low levels of output, *diseconomies of scale* are more likely at higher output levels. In Figure 8, you can see that Spotless does not experience diseconomies of scale until it is washing more than 185 cars per day.

Constant Returns to Scale. In Figure 8, you can see that for output levels between 130 and 185, the smoothed-out *LRATC* curve is roughly flat. Over this range of output, *LRATC* remains approximately constant as output increases. Here, output and *LRTC* rise by roughly the same proportion:

When both output and long-run total cost rise by the same proportion, production is characterized by constant returns to scale, and the LRATC curve is flat.

Why would a firm experience constant returns to scale? We have seen that as output increases, the impact of specialization and more efficient use of lumpy inputs will diminish, while the problems of bigness become more serious. At some level of production, these forces may just cancel out, so that an increase in output does not change *LRATC* at all. The firm will then have constant returns to scale until the problems of bigness begin to dominate. Constant returns to scale are most likely to occur at some *intermediate* range of output.

In sum, when we look at the behavior of *LRATC*, we often expect a pattern like the following: economies of scale (decreasing *LRATC*) at low levels of output,

Diseconomies of scale Long-run average total cost increases as output increases.

Constant returns to scale Long-run average total cost is unchanged as output increases.

constant returns to scale (constant *LRATC*) at some intermediate levels of output, and diseconomies of scale (increasing *LRATC*) at high levels of output. This is why *LRATC* curves are typically U-shaped.

Of course, even U-shaped *LRATC* curves will have different appearances at different firms. Indeed, *LRATC* curves need not be U-shaped at all. In later chapters, you will see that the shape of an *LRATC* curve has much to tell us about the economy: about the size of firms, the nature of competition among them, and the problems faced by government regulators. But you have already learned enough about production and costs to understand an important and ongoing problem in another country. It's a problem caused—at least in part—by a past failure to grasp the logic of cost curves.

Using the THEORY



COST CURVES AND ECONOMIC REFORM IN RUSSIA

At the beginning of this chapter, it was suggested that Russian manufacturing firms might be different from their counterparts in other countries. Now we can discuss one of the important differences: Russian firms seem to require more inputs per unit of output—more labor, capital, energy, and raw materials—than, say, an American firm producing a similar product. There are many reasons for this: an antiquated capital stock, poor infrastructure (e.g., roads and telecommunications), and the absence of management skills suited to the privately owned firm. But the problem has been exacerbated by a simple misunderstanding of production and cost made by generations of Soviet leaders—a misunderstanding whose legacy continues to haunt many Russian firms and industries even 10 years after reform.

First, a fact: the Soviet economy relied heavily on *monopolies*—single enterprises that were the sole producers of a good for the entire country. As late as 1991, Soviet government economists reported that out of 7,664 major product lines, 5,884 were manufactured by a single producer!² Most of these enterprises remain intact today, and as a result, they operate on a much larger scale than their Western counterparts.

Why so many monopolies? There were essentially two reasons. First, in a command economy, it was easier for the state to monitor and control fewer large enterprises than a greater number of small ones. Second, there was an ideological bias toward bigness: Soviet leaders from Vladimir Lenin to Leonid Brezhnev viewed the capitalist practice of having many firms, each producing the same item, as unnecessarily wasteful and duplicative. Why, they asked, should several firms make different brands of toothpaste that are more or less the same when a single manufacturer can produce toothpaste? Why have several competing automobile companies, each with its own separate design divisions, management teams, and distribution network, when a single enterprise would need only *one* design division, *one* management team, and *one* distribution network? Soviet leaders believed that avoiding wasteful duplication would enable their industries to operate more efficiently than those in the capitalist world, using fewer inputs for any level of output. In effect, they believed that their enterprises could continue to enjoy *economies of scale* until each one was producing for the entire Soviet market. This view of costs is illustrated in Figure 9, where the *LRATC* curve slopes downward over the entire range of output.³

² Marshall Goldman, *What Went Wrong with Perestroika* (New York: W. W. Norton, 1992), p. 154.

³ In later chapters, you will see that there are, indeed, firms whose *LRATC* curve slopes downward everywhere. But these firms are the exception, not the rule.

Had the Soviet leadership been right—had Figure 9 been an accurate portrayal of a typical firm's *LRATC* curve—then having monopoly enterprises producing for the entire market might have made sense.

But they were not right: With the transformation to a market economy, it has become apparent that production in Russia—as elsewhere—is mostly characterized by U-shaped *LRATC* curves, like the one in Figure 10. For years, the huge Russian enterprises have been operating in the region of *diseconomies of scale*, at output levels like Q_2 . As a result, their per-unit costs have been pushed higher than necessary. If production had been organized differently—with several firms, each producing a smaller quantity of output, like Q_1 —costs per unit could have been lower.

THE SOVIET VIEW OF PRODUCTION

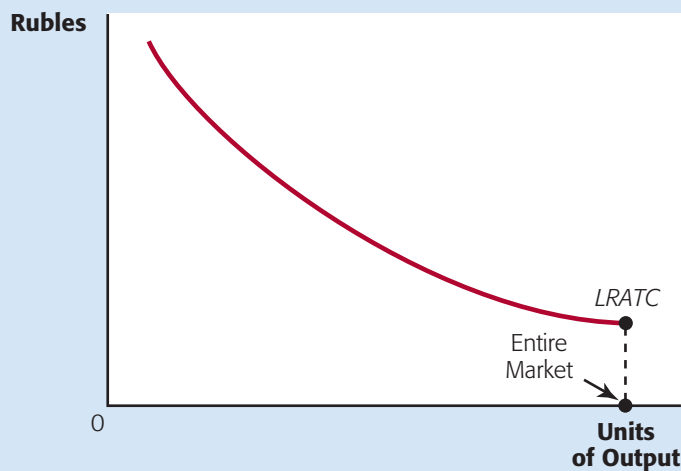


FIGURE 9

Planners in the former Soviet Union imagined—incorrectly—that their enterprises could enjoy economies of scale until each was producing for the entire market.

SOVIET PRODUCTION: THE REALITY

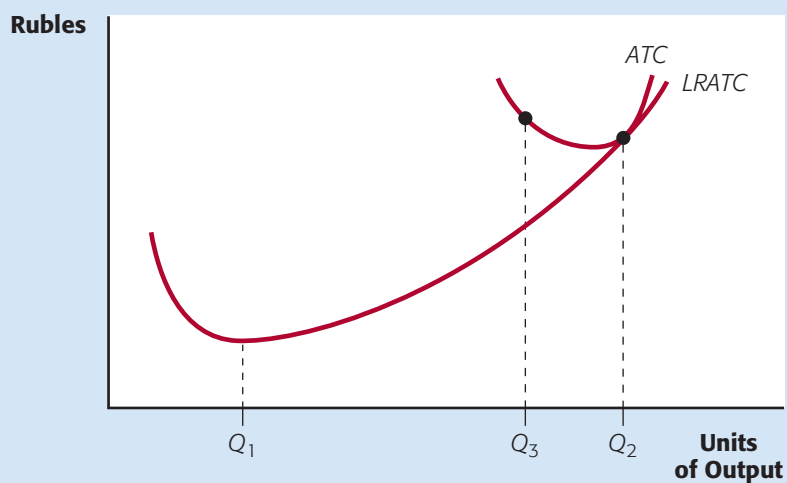


FIGURE 10

Newly privatized Russian firms have been operating in the region of *diseconomies of scale*—at levels like Q_2 . If they reduce output in the short run to Q_3 , they will find average costs rising along *ATC*.

Today, many of these huge Russian firms are in a serious jam. Since free trade has opened up, they must compete with Western firms, which can produce at substantially lower costs per unit. One recent study has found that the average Russian industrial firm requires five times the labor, raw materials, and other inputs of its U.S. counterpart.⁴ What can these Russian firms do?

Unfortunately, not much. In the short run, when plant size is fixed, any reduction in output would move the firm along its *ATC* curve, not its *LRATC* curve. Thus, reducing output—say, to Q_3 —might *raise* per-unit costs rather than lower them.

But even over a longer time horizon, these firms will not find it easy to change their scale of operations and move along their *LRATC* curves. Unlike many large Western firms—which have grown large by taking over smaller firms or building multiple plants in different parts of the country—many huge Russian firms were *built* as single-plant enterprises. The factory, the equipment, the technology—all were designed for a single plant, so it is not so easy to reduce plant size by spinning off a part of the operation. In many cases, the entire facility would have to be re-designed from the ground up in order for the firm to move down its *LRATC* curve and lower its costs per unit. But this would prove too costly.

Many Russian firms have found an easier way: They sell their products at world prices, suffering losses because of their higher costs, and then appeal to the government for subsidies to cover their losses. This means that billions of rubles in government funds that *could* be used for needed government services are instead keeping inefficient firms afloat.

Interestingly, one sector of the Russian economy—food retailing—suffers from the *opposite* problem: firms that are too *small*. In Russia today, almost all food purchases are from tiny groceries, butchers and bakers. But these little shops—which cannot take advantage of economies of scale—operate at only about 25 percent of the efficiency of larger supermarkets. That is, while supermarkets operate at or near Q_1 in Figure 10, the tiny food shops operate to the left of that output level, with substantially higher average costs. But the Russian government taxes supermarkets at a higher rate, eliminating their advantage. As a result, the higher-cost tiny shops are kept afloat, and larger, low-cost supermarkets are unable to compete with them.

⁴ Lewis, William W., “In Russia’s Economy, It’s Survival of the Weakest,” *Wall Street Journal*, November 4, 1999, p. A30.

S U M M A R Y

Business firms combine inputs to produce outputs. While some production takes place in the household, production through business firms with employees allows gains from specialization (each worker may specialize in one aspect of production), lower transaction costs, and reduced risk for employees.

A firm’s *production function* describes the maximum output it can produce using different quantities of inputs. In the *short run*, at least one of the firm’s inputs is fixed. In the *long run*, all inputs can be varied.

A firm’s *cost of production* is the opportunity cost of its owners—everything they must give up in order to produce output. In the short run, some costs are *fixed* and independ-

ent of the level of production. Other costs—*variable costs*—change as production increases. *Marginal cost* is the change in total cost from producing one more unit of output. The *marginal cost curve* has a U shape, reflecting the underlying marginal product of labor. A variety of average cost curves can be defined. The *average variable cost curve* and the *average total cost curve* are each U-shaped, reflecting the relationship between average and marginal cost.

In the long run, all costs are variable. The firm’s *long-run total cost curve* indicates the cost of producing each quantity of output with the least-cost input mix. The related *long-run average total cost (LRATC) curve* is formed by combining

portions of different ATC curves—each portion representing a different plant size. The shape of the *LRATC* curve reflects the nature of returns to scale. It slopes downward when there

are economies of scale, slopes upward when there are diseconomies of scale, and is flat when there are constant returns to scale.

KEY TERMS

business firm	fixed input	explicit costs	long-run total cost
profit	variable input	implicit costs	long-run average total cost
sole proprietorship	total product	fixed costs	plant
partnership	marginal product of labor	variable costs	economies of scale
corporation	increasing marginal returns to labor	total fixed cost	diseconomies of scale
transaction costs	diminishing marginal returns to labor	total variable cost	constant returns to scale
diversification	law of diminishing marginal returns	total cost	
technology	returns	average fixed cost	
production function	sunk costs	average variable cost	
long run		average total cost	
short run		marginal cost	

REVIEW QUESTIONS

- What are the three types of business firm? Discuss the pros and cons of each type.
- Why is most production activity carried out by firms, rather than by independent contractors?
- A home builder incurs the following costs. Which are examples of transaction costs? Why?
 - Cost of lumber
 - Lawyer's fees for handling the legal work connected with the purchase of land
 - Interest expense on a loan to buy new equipment
 - Opportunity cost of time spent gathering bids from subcontractors
- During the late 1990s, there were numerous mergers of firms. In some cases, these firms produced the same products. In other cases, the merger brought together firms that made totally different products. Explain a possible motive for the mergers in each case.
- Given the advantages of larger firm size, why don't we expect firms to grow larger without limit?
- Discuss the distinction between the short run and the long run as those terms relate to production.
- Which of the following inputs would likely be classified as fixed and which as variable over a time horizon of one month? Why?
 - Ovens to the Nabisco bakery
 - Wood to the La-Z-Boy Chair Co.
 - Oranges to Minute Maid Juice Co.
 - Labor to a McDonald's hamburger franchise
 - Cars to Hertz Rent-a-Car Co.
- Explain the difference between the total output of a firm and the marginal product of labor (*MPL*) at that firm. How are they related?
- Classify the following as fixed or variable costs for a time horizon of six months. Justify your categorization.
 - General Motors' outlay for steel
 - Pillsbury's rent on its corporate headquarters
 - The cost of newsprint for the *New York Times*
- At home on Vulcan one summer, Mr. Spock spent all his time working on an invention to give McCoy a severe shock whenever he said, "Damn it, Jim, I'm a doctor!" Alas, the invention didn't work. Spock consoled himself with the idea that, since he had used Starfleet's equipment and lab, at least his failed attempt hadn't cost him anything. Is his thinking "logical"? Explain.
- Can long-run total cost (*LRTC*) ever be greater than short-run total cost (*TC*)? Why or why not?
- Explain the U shape of a typical long-run average cost curve. Specifically, why is the curve downward sloping at lower levels of output and upward sloping at higher?
- Explain the dilemma faced by many Russian enterprises today.

PROBLEMS AND EXERCISES

1. The following table shows total output (in tax returns completed per day) of the accounting firm of Hoodwink and Finagle:

Number of Accountants	Number of Returns per Day
0	0
1	5
2	12
3	17
4	20
5	22

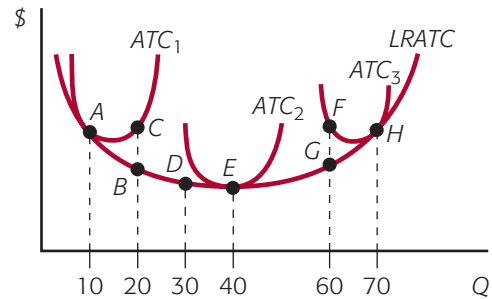
Assuming the quantity of capital (computers, adding machines, desks, etc.) remains constant at all output levels:

- Calculate the marginal product of each accountant.
 - Over what range of employment do you see increasing returns to labor? Diminishing returns?
 - Explain why *MPL* might behave this way in the context of an accounting firm.
2. The following table gives the short-run and long-run total costs for various levels of output of Consolidated National Acme, Inc.:

Q	TC_1	TC_2
0	0	350
1	300	400
2	400	435
3	465	465
4	495	505
5	560	560
6	600	635
7	700	735

- Which column, TC_1 or TC_2 , gives long-run total cost, and which gives short-run total cost? How do you know?
 - For each level of output, find short-run *TFC*, *TVC*, *AFC*, *AVC*, and *MC*.
 - At what output level would the firm's short-run and long-run input mixes be the same?
 - Starting from producing two units, Consolidated's managers decide to double production to four units. So they simply double all of their inputs in the long run. Comment on their managerial skill.
 - Over what range of output do you see economies of scale? Diseconomies of scale? Constant returns to scale?
3. Ludmilla's House of Schnitzel is currently producing 10 schnitzels a day at point A on the following diagram.

Ludmilla's business partner, Hans (an impatient sort), wants her to double production immediately.



- What point will likely illustrate Ludmilla's cost situation for the near future? Why?
 - If Ludmilla wants to keep producing 20 schnitzels, at what point does she want to be eventually? How can she get there?
 - Eventually, Ludmilla and company do very well, expanding until they find themselves making 70 schnitzels a day. But after a few years, Ludmilla discovers that profit was greater when she produced 20 schnitzels per day. She wants to scale back production to 20 schnitzels per day—laying off workers, selling off equipment, renting less space, and producing fewer schnitzels. Hans wants to reduce output by just cutting back on flour and milk and laying off workers. Who's right? Discuss the situation with reference to the relevant points on the diagram.
 - Does the figure tell us what output Ludmilla should aim for? Why or why not?
4. In a recent year, a long, hard winter gave rise to stronger-than-normal demand for heating oil. The following summer was characterized by strong demand for gasoline by vacationers. Show what these two events might have done to the short-run *MC*, *AVC*, and *ATC* curves of Continental Airlines. (*Hint*: How would these events affect the price of oil?)
5. Clean 'n' Shine is a competitor to Spotless Car Wash. Like Spotless, it must pay \$75 per day for each automated line it uses. But Clean 'n' Shine has been able to tap into a lower-cost pool of labor, paying its workers only \$50 per day. Clean 'n' Shine's production technology is given in the table below. To determine its short-run cost structure, fill in the blanks in the table.
- Over what range of output does Clean 'n' Shine experience increasing marginal returns to labor? Over what range does it experience decreasing marginal returns to labor?

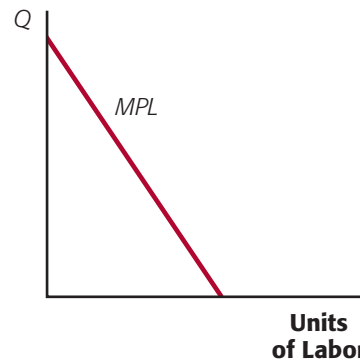
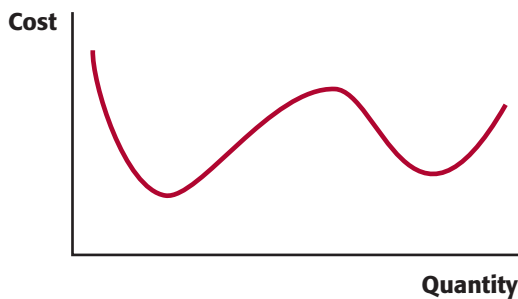
Short-Run Costs for Clean 'n' Shine Car Wash

(1) Output (per Day)	(2) Capital	(3) Labor	(4) TFC	(5) TVC	(6) TC	(7) MC	(8) AFC	(9) AVC	(10) ATC
0	1	0	\$___	\$___	\$___	\$___	—	—	—
30	1	1	\$___	\$___	\$___	\$___	\$___	\$___	\$___
70	1	2	\$___	\$___	\$___	\$___	\$___	\$___	\$___
120	1	3	\$___	\$___	\$___	\$___	\$___	\$___	\$___
160	1	4	\$___	\$___	\$___	\$___	\$___	\$___	\$___
190	1	5	\$___	\$___	\$___	\$___	\$___	\$___	\$___
210	1	6	\$___	\$___	\$___	\$___	\$___	\$___	\$___

- b. As output increases, do average fixed costs behave as described in the text? Explain.
- c. As output increases, do marginal cost, average variable cost, and average total cost behave as described in the text? Explain.
- d. Looking at the numbers in the table, but without drawing any curves, is the relationship between MC and AVC as described in the text? What about the relationship between MC and ATC?

C H A L L E N G E Q U E S T I O N S

- Draw the long-run total cost and long-run average cost curves for a firm that experiences:
 - Constant returns to scale over all output levels
 - Diseconomies of scale over low levels of output, constant returns to scale over intermediate levels of output, and economies of scale over high output levels. Does this pattern of costs make sense? Why or why not?
- A firm has the strange ATC curve drawn below. Sketch in the marginal cost curve this firm must have. (*Hint:* What do you know about the marginal-average relationship relating to cost?)
- The following curve shows the *marginal* product of labor for a firm at different levels of output.



- Show what the corresponding total product curve would look like.
- Do the total and marginal product curves for this firm ever exhibit diminishing marginal returns to labor? Increasing marginal returns to labor?

EXPERIENTAL EXERCISES

1. The terms “diminishing returns” and “economies of scale” are often used in the popular press. Use an Internet search engine such as Excite (<http://www.excite.com>) to search for these two terms. Check the first five sites you find and, in each case, determine whether the term is being used correctly. If it is not, see if you can determine the source of the writer’s confusion.



2. The “Technology” column in the Marketplace section of the *Wall Street Journal* describes many interesting technological innovations. Pick one and see if you can determine how it might affect a firm’s average cost curves. In the short run, will it affect the firm’s variable costs only, fixed costs only, or both types of costs? In the long run, what effect will the innovation have on the firm’s *LRATC* curve?

HOW FIRMS MAKE DECISIONS: PROFIT MAXIMIZATION

CHAPTER

7

In early 1996, the managers of Nintendo America, Inc., knew that they had a winner on their hands: the Nintendo 64 video-game player. With this new product, players would be able to jump, fly, and even swim through a variety of three-dimensional fantasy worlds, with images more spectacular and action much faster than in any competing product.

Then came the hard questions. Where should the new product be produced: Japan, the United States, or perhaps Hong Kong? How should the company raise the funds to pay the costs of production? When should it bring the product to market? How much should it spend on advertising, and in which types of media? And finally, what price should the company charge, and how many units should it plan to produce?

These last decisions—how much to produce and what price to charge—are the focus of this chapter. In the end, Nintendo planned to produce 500,000 units, and decided to charge \$199. But why didn't it charge a lower price that would allow it to sell more output? Or a higher price that would give it more profit on each unit sold?

Although this chapter concentrates on firms' decisions about price and output level, the tools you will learn apply to many other firm decisions. How much should MasterCard spend on advertising? How late should Starbucks keep its coffee shops open? How many copies should *Newsweek* give away free to potential subscribers? Should movie theaters offer Wednesday afternoon showings that only a few people attend? This chapter will help you understand how firms answer these sorts of questions.

THE GOAL OF PROFIT MAXIMIZATION

To analyze decision making at the firm, let's start with a very basic question: What is the firm trying to maximize?

Economists have given this question a lot of thought. Some firms—especially large ones—are complex institutions in which many different groups of people work together. A firm's owners will usually want the firm to earn as much profit as

CHAPTER OUTLINE

The Goal of Profit Maximization

Understanding Profit

Two Definitions of Profit
Why Are There Profits?

The Firm's Constraints

The Demand Constraint
The Cost Constraint

The Profit-Maximizing Output Level

The Total Revenue and Total Cost Approach
The Marginal Revenue and Marginal Cost Approach
Profit Maximization Using Graphs
What About Average Costs?

The Importance of Marginal Decision Making: A Broader View

Dealing with Losses

The Short Run and the Shutdown Rule
The Long Run: The Exit Decision

The Goal of the Firm Revisited

The Principal-Agent Problem
The Principal-Agent Problem at the Firm
The Assumption of Profit Maximization

Using the Theory: Getting It Wrong and Getting It Right

Getting It Wrong: The Failure of Franklin National Bank
Getting It Right: The Success of Continental Airlines

possible. But the workers and managers who actually run the firm may have other agendas. They may try to divert the firm away from profit maximization in order to benefit themselves. For now, let's assume that workers and managers are faithful servants of the firm's owners. That is,

Identify Goals and Constraints



We will view the firm as a single economic decision maker whose goal is to maximize its owners' profit.

Why do we make this assumption? Because it has proven so *useful* in understanding how firms behave. Toward the end of the chapter, we'll come back to the important topic of different groups within the firm and the potential conflicts among them.

UNDERSTANDING PROFIT

Profit is defined as the firm's *sales revenue* minus its *costs of production*. There is widespread agreement over how to measure the firm's revenue, the flow of money into the firm. But there are two different conceptions of the firm's costs, and each of them leads to a different definition of profit.

TWO DEFINITIONS OF PROFIT

One conception of costs is the one used by accountants. With a few exceptions, accountants consider only *explicit* costs, where money is actually paid out.¹ If we deduct only the costs recognized by accountants, we get one definition of profit:

$$\text{Accounting profit} = \text{Total revenue} - \text{Accounting costs.}$$

But economics, as you have learned, has a much broader view of cost—*opportunity cost*. For the firm's owners, opportunity cost is the total value of *everything* sacrificed to produce output. This includes not only the explicit costs recognized by accountants—such as wages and salaries and outlays on raw materials—but also *implicit costs*, when something is given up but no money changes hands. For example, if an owner contributes his own time or money to the firm, there will be foregone wages or foregone investment income—both implicit costs for the firm.

This broader conception of costs leads to a second definition of profit:

$$\begin{aligned} \text{Economic profit} &= \text{Total revenue} - \text{All costs of production} \\ &= \text{Total revenue} - (\text{Explicit costs} + \text{Implicit costs}) \end{aligned}$$

The difference between economic profit and accounting profit is an important one; when they are confused, some serious (and costly) mistakes can result. An example might help make the difference clear.

Suppose you own a firm that produces T-shirts, and you want to calculate your profit over the year. Your bookkeeper provides you with the following information:

Accounting profit Total revenue minus accounting costs.

Economic profit Total revenue minus all costs of production, explicit and implicit.

¹ One exception is *depreciation*, a charge for the gradual wearing out of the firm's plant and equipment. Accountants include this as a cost even though no money is actually paid out.

Total Revenue from Selling T-shirts	\$300,000
Cost of raw materials	\$ 80,000
Wages and salaries	150,000
Electricity and phone	20,000
Advertising cost	40,000
Total Explicit Cost	290,000
Accounting Profit	\$ 10,000

From the looks of things, your firm is earning a profit, so you might feel pretty good. Indeed, if you look only at *money* coming in and *money* going out, you have indeed earned a profit—\$10,000 for the year . . . in *accounting* profit.

But suppose that in order to start your business you invested \$100,000 of your own money—money that could have been earning \$6,000 in interest if you'd put it in the bank instead. Also, you are using two extra rooms in your own house as a factory—rooms that could have been rented out for \$4,000 per year. Finally, you are managing the business full time, without receiving a separate salary, and you could instead be working at a job earning \$40,000 per year. All of these costs—the interest, rent, and salary you *could* have earned—are implicit costs that have not been taken into account by your bookkeeper. They are part of the opportunity cost of your firm, because they are sacrifices you made to operate your business.

Now let's look at this business from the economist's perspective and calculate your *economic* profit.

Total Revenue from Selling T-shirts	\$300,000
Cost of raw materials	\$ 80,000
Wages and salaries	150,000
Electricity and phone	20,000
Advertising cost	40,000
Total Explicit Costs	\$290,000
Investment income foregone	\$ 6,000
Rent foregone	4,000
Salary foregone	40,000
Total Implicit Costs	\$ 50,000
Total Costs	\$340,000
Economic Profit	-\$ 40,000

From an economic point of view, your business is not profitable at all, but is actually losing \$40,000 per year! But wait—how can we say that your firm is suffering a loss when it takes in more money than it pays out? Because, as we've seen, your *opportunity cost*—the value of what you are giving up to produce your output—includes more than just money costs. When *all* costs are considered—implicit as well as explicit—your total revenue is not sufficient to cover what you have sacrificed to run your business. You would do better by shifting your time, your money, and your spare room to some alternative use.

Which of the two definitions of profit is the correct one? Either one of them, depending on the reason for measuring it. For tax purposes, the government is interested in profits as measured by accountants. The government cares only about the money you've earned, not what you *could* have earned had you done something else with your money or your time.

However, for our purposes—understanding the behavior of firms—economic profit is clearly better. Should your T-shirt factory stay in business? Should it expand

or contract in the long run? Will other firms be attracted to the T-shirt industry? It is economic profit that will help us answer these questions, because it is economic profit that you and other owners care about.

The proper measure of profit for understanding and predicting the behavior of firms is economic profit. Unlike accounting profit, economic profit recognizes all the opportunity costs of production—both explicit costs and implicit costs.

Let's apply these ideas to Microsoft Corporation. In the year ending in June 1999, Microsoft had an accounting profit of \$7.8 billion. But this was not its *economic* profit. Microsoft's owners—its shareholders—had invested about \$30 billion in the company—money that *could* have earned interest in a bank or some other financial investment. The foregone investment earnings must be included as part of the opportunity cost paid by Microsoft's owners. Let's suppose that they could have earned 5 percent by putting their money into some other investment. Then, the foregone investment income was $\$30 \text{ billion} \times 0.05 = \1.5 billion . Assuming that the foregone investment income was the only implicit cost for Microsoft, we then deduct it from the accounting profit of \$7.8 billion to obtain an *economic* profit of $\$7.8 \text{ billion} - \$1.5 \text{ billion} = \$6.3 \text{ billion}$.²

WHY ARE THERE PROFITS?

When you look at the income received by households in the economy, you see a variety of payments. Those who provide firms with land receive *rent*—the payment for land. Those who provide labor receive a wage or salary. And those who lend firms money so they can purchase capital equipment receive interest. The firm's profit goes to its owners. But what do the owners of the firm provide that earns them this payment?

Economists view profit as a payment for two contributions that are just as necessary for production as are land, labor, or machinery. These two contributions are *risk-taking* and *innovation*.

Consider a restaurant that happens to be earning profit for its owner. The land, labor, and capital the restaurant uses to produce its meals did not simply come together magically. Someone—the owner—had to be willing to take the initiative to set up the business, and this individual assumed the risk that the business might fail and the initial investment be lost. Because the consequences of loss are so severe, the reward for success must be large in order to induce an entrepreneur to establish a business.

On a larger scale, Ted Turner risked hundreds of millions of dollars in the late 1970s when he created Cable News Network (CNN). Now that CNN has turned out to be so successful, it is easy to forget how risky the venture was at the outset. At the time, many respected financial analysts forecast that the project would fail and Turner would be driven into bankruptcy.

Profits are also a reward for *innovation*. Ted Turner was the first to create a 24-hour global news network, just as Steven Jobs and Steven Wozniak—when they formed the Apple Computer Company in the 1970s—were the first to produce a usable personal computer for the mass market. These are obvious innovations.

But innovations can also be more subtle, and they are more common than you might think. When you pass by a successful laundromat, you may not give it a sec-

² Source: Morningstar.com database, 11/25/99. Some of Microsoft's owners are also workers or managers in the firm. We do not count their sacrifice of time as part of the owners' implicit cost because these owners receive a separate salary to compensate for their time.

ond thought. But someone, at some time, had to be the first one to realize, “I bet a laundromat in this neighborhood would do well”—an innovation. There can also be innovations in the production process, such as that improvement in mass production that made the disposable contact lens possible.

In almost any business, if you look closely, you will find that some sort of innovation was needed to get things started. Innovation, like taking on the risk of losing substantial wealth, makes an essential contribution to production. Profit is, in part, a reward to those who innovate.

THE FIRM'S CONSTRAINTS

If the firm were free to earn whatever level of profit it wanted, it would earn virtually infinite profit. This would make the owners very happy. Unfortunately for owners, though, the firm is not free to do this; it faces *constraints* on both its revenue and its costs.

THE DEMAND CONSTRAINT

One constraint on the firm's profit arises from a familiar concept: the demand curve. This curve always tells us the quantity of a good buyers wish to buy at different prices. But which buyers? And from which firms are they buying? Depending on how we answer these questions, we might be talking about any of several different types of demand curves.

Market demand curves—like the ones you studied in Chapters 3 and 4—tell us the quantity demanded by *all* consumers from *all* firms in a market. The *individual demand curve* you studied in Chapter 5 referred to the quantity of a good demanded by *one* consumer only. In this chapter, we look at yet another kind of demand curve:

The demand curve facing the firm tells us, for different prices, the quantity of output that customers will choose to purchase from that firm.

Notice that this new demand curve—the demand curve facing the firm—refers to only *one* firm, and to *all buyers* who are potential customers of that firm.

Let's consider the demand curve faced by Ned, the owner and manager of Ned's Beds—a manufacturer of bed frames. Figure 1 lists the different prices that Ned could charge for each bed frame and the number of them (per day) he can sell at each price. The figure also shows a graph of the demand curve facing Ned's firm. For each price (on the vertical axis), it shows us the quantity of output the firm can sell (on the horizontal axis). Notice that, like the other types of demand curves we have studied, the demand curve facing the firm slopes downward. In order to sell more bed frames, Ned must lower his price.³

The definition of the demand curve facing the firm suggests that once it selects a price, the firm has also determined how much output it will sell. But, as you saw a few chapters ago, we can also flip the demand relationship around: Once the firm

³ The downward-sloping demand curve tells us that Ned's Beds sells its output in an *imperfectly competitive market*—a market where the firm can *set* its price. Most firms operate in this type of market. If a manager thinks, “I'd like to sell more output, but then I'd have to lower my price, so let's see if it's worth it,” we know he operates in an imperfectly competitive market. In a *perfectly competitive market*, by contrast, the firm would have to accept the market price as given—a case we'll take up in the next chapter.



Microsoft's accounting profit is greater than its economic profit.



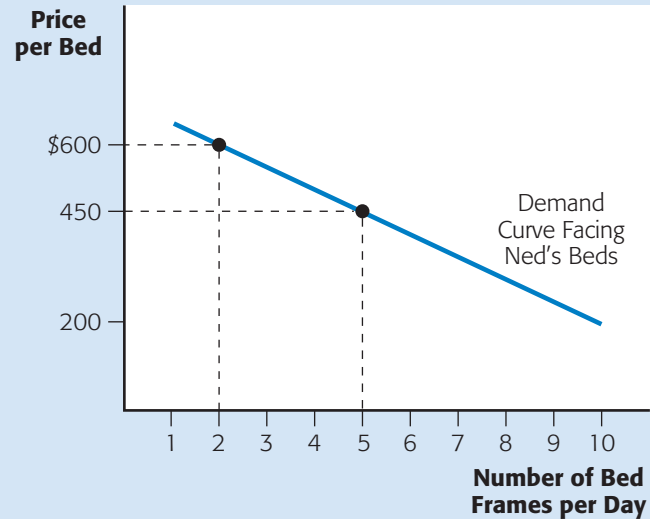
Identify Goals and Constraints

Demand curve facing the firm A curve that indicates, for different prices, the quantity of output that customers will purchase from a particular firm.

FIGURE 1

THE DEMAND CURVE FACING THE FIRM

(1) Price	(2) Output	(3) Total Revenue	(4) Total Cost	(5) Profit
> \$650	0	0	\$ 300	-\$ 300
\$650	1	\$ 650	\$ 700	-\$ 50
\$600	2	\$1,200	\$ 900	\$ 300
\$550	3	\$1,650	\$1,000	\$ 650
\$500	4	\$2,000	\$1,150	\$ 850
\$450	5	\$2,250	\$1,350	\$ 900
\$400	6	\$2,400	\$1,600	\$ 800
\$350	7	\$2,450	\$1,900	\$ 550
\$300	8	\$2,400	\$2,250	\$ 150
\$250	9	\$2,250	\$2,650	-\$ 400
\$200	10	\$2,000	\$3,100	-\$1,100



The table presents information about Ned's Beds. Data from the first two columns are plotted in the figure to show the demand curve facing the firm. At any point along that demand curve, the product of price and quantity equals total revenue, which is given in the third column of the table.

has selected an output level, it has also determined the maximum price it can charge. This leads to an alternative definition:

The demand curve facing the firm shows us the maximum price the firm can charge to sell any given amount of output.

Looking at Figure 1 from this perspective, we see that the horizontal axis shows alternative levels of output and the vertical axis shows the price Ned should charge if he wishes to sell each quantity of output.

These two different ways of defining the firm's demand curve show us that it is, indeed, a constraint for the firm. The firm can freely determine *either* its price *or* its level of output. But once it makes the choice, the other variable is automatically determined by the firm's demand curve. Thus, the firm has only *one* choice to make. Selecting a particular price *implies* a level of output, and selecting an output level *implies* a particular price. Economists typically focus on the choice of output level, with the price implied as a consequence. We will follow that convention in this textbook.

Total revenue The total inflow of receipts from selling a given amount of output.

Total Revenue. A firm's **total revenue** is the total inflow of receipts from selling output. Each time the firm chooses a level of output, it also determines its total revenue. Why? Because once we know the level of output, we also know the highest price the firm can charge. Total revenue—which is the number of units of output times the price per unit—follows automatically.

The third column in Figure 1 lists the total revenue of Ned's Beds. Each entry is calculated by multiplying the quantity of output (column 2) by the price per unit (column 1). For example, if Ned's firm produces 2 bed frames per day, he can charge \$600 for each of them, so total revenue will be $2 \times \$600 = \$1,200$. If Ned increases output to 3 units, he must lower the price to \$550, earning a total revenue of $3 \times \$550 =$

\$1,650. Because the firm's demand curve slopes downward, Ned must lower his price each time his output increases, or else he will not be able to sell all he produces. With more units of output, but each one selling at a lower price, total revenue could rise or fall. Scanning the total revenue column, we see that for this firm, total revenue first rises and then begins to fall. This will be discussed in greater detail later on.

THE COST CONSTRAINT

Every firm struggles to reduce costs, but there is a limit to how low costs can go. These limits impose a second constraint on the firm. Where do the limits come from? They come from concepts that you learned about in Chapter 6. Let's review them briefly.

First, the firm has a given production function, which is determined by its production technology. The production function tells us all the different ways in which the firm can produce any given level of output. In the long run, when all inputs are variable, the firm can use *any* method in its production function. In the short run, it is even more constrained: Not only is it limited by its production function, but it can only use *some* of the methods in that production function, because one or more of its inputs are *fixed*.

Second, the firm must pay *prices* for each of the inputs that it uses, and we assume there is nothing the firm can do about those prices. Together, the production function and the prices of the inputs determine what it will cost to produce any given level of output. And once the firm chooses the *least cost* method available, it has driven the cost of producing that output level as low as it can go.

The firm uses its production function, and the prices it must pay for its inputs, to determine the least cost method of producing any given output level. Therefore, for any level of output the firm might want to produce, it must pay the cost of the "least cost method" of production. This is the firm's cost constraint.

The fourth column of Figure 1 lists Ned's total cost—the lowest possible cost of producing each quantity of output. More output always means greater costs, so the numbers in this column are always increasing. For example, at an output of zero, total cost is \$300. This tells us we are looking at costs in the short run, over which some of the firm's costs are *fixed*. (What would be the cost of producing 0 units if this were the long run?) If output increases from 0 to 1 bed frame, total cost rises from \$300 to \$700. This increase in total costs—\$400—is caused by an increase in *variable* costs, such as labor and raw materials.

We can sum up our discussion of the firm's constraints as follows:

The firm faces constraints that limit its ability to earn profit. For each level of output the firm might choose, its demand curve determines the price it can charge and the total revenue it will receive. Its production technology and the price of its inputs determine the total cost it must bear.

THE PROFIT-MAXIMIZING OUTPUT LEVEL

In this section, we ask a very simple question: How does a firm find the level of output that will earn it the greatest possible profit? We'll look at this question from several angles, each one giving us further insight into the behavior of the firm.



Identify Goals and Constraints

THE TOTAL REVENUE AND TOTAL COST APPROACH

At any given output level, we know (1) how much revenue the firm will earn and (2) its cost of production. We can then easily calculate profit, which is just the difference between total revenue (TR) and total cost (TC).

In the total revenue and total cost approach, the firm calculates Profit = $TR - TC$ at each output level and selects the output level where profit is greatest.

Let's see how this works for Ned's Beds. Column 5 of Figure 1 lists total profit at each output level. If the firm were to produce no bed frames at all, total revenue (TR) would be 0, while total cost (TC) would be \$300. Total profit would be $TR - TC = 0 - \$300 = -\300 . We would say that the firm earns a profit of negative \$300 or a **loss** of \$300 per day. Producing one bed frame would raise total revenue to \$650 and total cost to \$700, for a loss of \$50. Not until the firm produces 2 bed frames does total revenue rise above total cost and the firm begin to make a profit. At 2 bed frames per day, TR is \$1,200 and TC is \$900, so the firm earns a profit of \$300. Remember that as long as we have been careful to include *all* costs in TC —implicit as well as explicit—the profits and losses we are calculating are *economic* profits and losses.

In the total revenue and total cost approach, finding the profit-maximizing output level is straightforward: We just scan the numbers in the profit column until we find the largest value—\$900—and the output level at which it is achieved—5 units per day. We conclude that the profit-maximizing output for Ned's Beds is 5 units per day.

THE MARGINAL REVENUE AND MARGINAL COST APPROACH

There is another way to find the profit-maximizing level of output. This approach, which uses *marginal* concepts, gives us some powerful insights into the firm's decision-making process. Recall that *marginal* cost is the *change* in total cost from producing one more unit of output. Now, let's consider a similar concept for revenue.

Marginal revenue (MR) is the change in total revenue from producing one more unit of output. Mathematically, MR is calculated by dividing the change in total revenue (ΔTR) by the change in output (ΔQ): $MR = \Delta TR / \Delta Q$.

Table 1 reproduces the TR and TC columns from Figure 1, but adds columns for marginal revenue and marginal cost. (In the table, output is always changing by one unit, so we can use ΔTR alone as our measure of marginal revenue.) For example, when output changes from 2 to 3 units, total revenue rises from \$1,200 to \$1,650. For this output change, $MR = \$450$. As usual, marginals are placed *between* different output levels because they tell us what happens as output *changes* from one level to another.

There are two important things to notice about marginal revenue. First, when MR is *positive*, an increase in output causes total revenue to *rise*. In the table, MR is positive for all increases in output from 0 to 7 units. When MR is

Loss A negative profit—when total cost exceeds total revenue.

Marginal revenue The change in total revenue from producing one more unit of output.



You may be tempted to forget about profit and think that the firm should produce where its total revenue is maximized. As you can see in Figure 1, total revenue is greatest when the firm produces 7 units per day, but at this output level, profit is not as high as it could be. The firm does better by producing only 5 units. True, revenue is lower at 5 units, but so are costs. It is the difference between revenue and cost that matters, not revenue alone.

TABLE 1

Output	Total Revenue	Marginal Revenue	Total Cost	Marginal Cost	Profit
0	0		\$ 300		-\$ 300
1	\$ 650	\$650	\$ 700	\$400	-\$ 50
2	\$1,200	\$550	\$ 900	\$200	\$ 300
3	\$1,650	\$450	\$1,000	\$100	\$ 650
4	\$2,000	\$350	\$1,150	\$150	\$ 850
5	\$2,250	\$250	\$1,350	\$200	\$ 900
6	\$2,400	\$150	\$1,600	\$250	\$ 800
7	\$2,450	\$ 50	\$1,900	\$300	\$ 850
8	\$2,400	-\$ 50	\$2,250	\$350	\$ 550
9	\$2,250	-\$150	\$2,650	\$400	\$ 150
10	\$2,000	-\$250	\$3,100	\$450	-\$ 400

**MORE DATA FOR
NED'S BEDS**

negative, an increase in output causes total revenue to *fall*, as occurs for all increases beyond 7 units.

The second thing to notice about *MR* is a bit more complicated: Each time output increases, *MR* is *smaller* than the price the firm charges at the new output level. For example, when output increases from 2 to 3 units, the firm's total revenue rises by \$450—even though it sells the third unit for a price of \$550. This may seem strange to you. After all, if the firm increases output from 2 to 3 units, and it gets \$550 for the third unit of output, why doesn't its total revenue rise by \$550?

The answer is found in the firm's downward-sloping demand curve, which tells us that to sell more output, the firm must cut its price. Look back at Figure 1 (p. 192). When output increases from 2 to 3 units, the firm must lower its price from \$600 to \$550. Moreover, the new price of \$550 will apply to *all three* units the firm sells.⁴ This means it *gains* some revenue—\$550—by selling that third unit. But it also *loses* some revenue—\$100—by having to lower the price by \$50 on each of the two units of output it could have otherwise sold at \$600. Marginal revenue will always equal the *difference* between this gain and loss in revenue—in this case, $\$550 - \$100 = \$450$.

When a firm faces a downward-sloping demand curve, each increase in output causes a revenue gain—from selling additional output at the new price—and a revenue loss—from having to lower the price on all previous units of output. Marginal revenue is therefore less than the price of the last unit of output.

⁴ Some firms can charge two or more different prices for the same product. We'll explore some examples in Chapter 9.

Using *MR* and *MC* to Maximize Profits. Now we'll see how marginal revenue, together with marginal cost, can be used to find the profit-maximizing output level. The logic behind the *MC* and *MR* approach is this:

An increase in output will always raise profit as long as marginal revenue is greater than marginal cost ($MR > MC$).

Notice the word *always*. Let's see why this rather sweeping statement must be true. Table 1 tells us that when output rises from 2 to 3 units, *MR* is \$450, while *MC* is \$100. This change in output causes both total revenue and total cost to rise, but it causes revenue to rise by *more* than cost ($\$450 > \100). As a result, profit must increase. Indeed, looking at the profit column, we see that increasing output from 2 to 3 units *does* cause profit to increase, from \$300 to \$650.⁵

The converse of this statement is also true:

An increase in output will always lower profit whenever marginal revenue is less than marginal cost ($MR < MC$).

For example, when output rises from 5 to 6 units, *MR* is \$150, while *MC* is \$250. For this change in output, both total revenue and total cost rise, but cost rises *more*, so profit must go down. In Table 1, you can see that this change in output does indeed cause profit to decline, from \$900 to \$800.

These insights about *MR* and *MC* lead us to the following simple guideline the firm should use to find its profit-maximizing level of output:

To find the profit-maximizing output level, the firm should increase output whenever $MR > MC$, and decrease output when $MR < MC$.

Let's apply this rule to Ned's Beds. In Table 1 we see that when moving from 0 to 1 unit of output, *MR* is \$650, while *MC* is only \$400. Since *MR* is larger than *MC*, making this move will increase profit. Thus, if the firm is producing 0 beds, it should always increase to 1 bed. Should it stop there? Let's see. If it moves from 1 to 2 beds, *MR* is \$550, while *MC* is only \$200. Once again, $MR > MC$, so the firm should increase to 2 beds. You can verify from the table that if the firm finds itself producing 0, 1, 2, 3, or 4 beds, $MR > MC$ for an increase of 1 unit, so it will always make greater profit by increasing production.

Until, that is, output reaches 5 beds. At this point, the picture changes: From 5 to 6 beds, *MR* is \$150, while *MC* is \$250. For this move, $MR < MC$, so profits would decrease. Thus, if the firm is producing 5 beds, it should *not* increase to 6. The same is true at every other output level beyond 5 units: The firm should *not* raise its output, since $MR < MC$ for each increase. We conclude that Ned maximizes his profit by producing 5 beds per day—the same answer we got using the *TR* and *TC* approach earlier.⁶

⁵ You may have noticed that the rise in profit (\$350) is equal to the difference between *MR* and *MC* in this example. This is no accident. *MR* tells us the *rise* in revenue; *MC* tells us the *rise* in cost. The difference between them will always be the *rise* in profit.

⁶ It sometimes happens that *MR* is precisely equal to *MC* for some change in output, although this does not occur in Table 1. In this case, increasing output would cause *both* cost and revenue to rise by equal amounts, so there would be *no* change in profit. The firm should not care whether it makes this change in output or not.

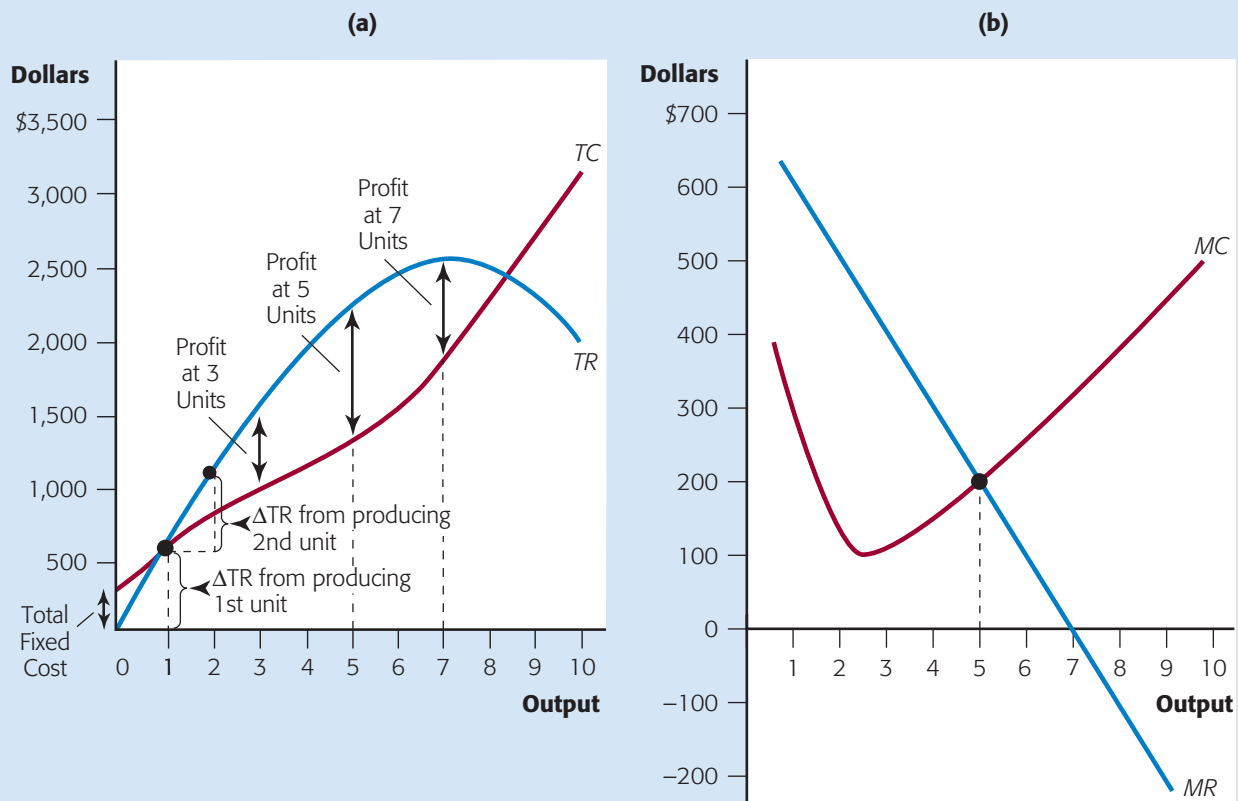
PROFIT MAXIMIZATION USING GRAPHS

Both approaches to maximizing profit (using totals or using marginals) can be seen even more clearly when we use graphs. In Figure 2(a) and (b), the data from Table 1 have been plotted—the TC and TR curves in the left panel, and the MC and MR curves in the right one.

Note the important relationship between the MR and TR curves. MR tells us the *change* in total revenue as output increases. Thus, as long as the MR curve lies above the horizontal axis ($MR > 0$), TR must be increasing, and the TR curve must slope upward. In the figure, $MR > 0$, and the TR curve slopes upward from zero to 7 units. When the MR curve dips below the horizontal axis ($MR < 0$), TR is decreasing, so the TR curve begins to slope downward. In the figure, this occurs beyond 7 units of output. As output increases in Figure 2, MR is first positive and then turns negative, so the TR curve will first *rise* and then *fall*.

The TR and TC Approach Using Graphs. Now let's see how we can use the TC and TR curves to guide the firm to its profit-maximizing output level. We know that the firm earns a profit at any output level where $TR > TC$ —where the TR curve lies

PROFIT MAXIMIZATION

FIGURE 2


Panel (a) shows the firm's total revenue (TR) and total cost (TC) curves. Profit is the vertical distance between the two curves at any level of output. Profit is maximized when that vertical distance is greatest—at 5 units of output. Panel (b) shows the firm's marginal revenue (MR) and marginal cost (MC) curves. (As long as MR lies above the horizontal axis, the TR curve slopes upward.) Profit is maximized at the level of output closest to where the two curves cross—at 5 units of output.

above the TC curve. In Figure 2(a), you can see that all output levels between 2 and 8 units are profitable for the firm. The *amount* of profit is simply the *vertical distance* between the TR and TC curves, whenever the TR curve lies above the TC curve. Since the firm cannot sell part of a bed frame, it must choose whole numbers for its output, so the profit-maximizing output level is simply the whole-number quantity at which this vertical distance is greatest—5 units of output. Of course, the TR and TC curves in Figure 2 were plotted from the data in Table 1, so we should not be surprised to find the same profit-maximizing output level—5 units—that we found before when using the table.

We can sum up our graphical rule for using the TR and TC curves this way:

To maximize profit, the firm should produce the quantity of output where the vertical distance between the TR and TC curves is greatest and the TR curve lies above the TC curve.

The MR and MC Approach Using Graphs. Figure 2 also illustrates the MR and MC approach to maximizing profits. As usual, the marginal data in panel (b) are plotted *between* output levels, since they tell us what happens as output changes from one level to another.

In the diagram, as long as output is less than 5 units, the MR curve lies above the MC curve ($MR > MC$), so the firm should produce more. For example, if we consider the move from 4 to 5 units, we compare the MR and MC curves at the midpoint between 4 and 5. Here, the MR curve lies above the MC curve, so increasing output from 4 to 5 will increase profit.

But now suppose the firm is producing 5 units and considering a move to 6. At the midpoint between 5 and 6 units, the MR curve has already crossed the MC curve, and now it lies *below* the MC curve. For this move, $MR < MC$, so raising output would *decrease* the firm's profit. The same is true for every increase in output beyond 5 units: The MR curve always lies below the MC curve, so the firm will decrease its profits by increasing output. Once again, we find that the profit-maximizing output level for the firm is 5 units.

Notice that the profit-maximizing output level—5 units—is the level closest to where the MC and MR curves cross. This is no accident. For each change in output that *increases* profit, the MR curve will lie above the MC curve. The first time that an output change *decreases* profit, the MR curve will cross the MC curve and dip below it. Thus, the MC and MR curves will always cross closest to the profit-maximizing output level.

With this graphical insight, we can summarize the MC and MR approach this way:

To maximize profit, the firm should produce the level of output closest to the point where $MC = MR$ —that is, the level of output at which the MC and MR curves intersect.

This rule is very useful, since it allows us to look at a diagram of MC and MR curves and *immediately* identify the profit-maximizing output level. In this text, you will often see this rule. When you read, “The profit-maximizing output level is where MC equals MR ,” translate to “The profit-maximizing output level is closest to the point where the MC curve crosses the MR curve.”

An Important Proviso. There is one important exception to this rule. Sometimes the MC and MR curves cross at two different points. In this case, the profit-

maximizing output level is the one at which the *MC curve crosses the MR curve from below*.

Figure 3 shows why. At point A, the *MC curve crosses the MR curve from above*. Our rule tells us that the output level at this point— Q_1 —is *not* profit maximizing. Why not? Because at output levels

lower than Q_1 , $MC > MR$, so profit *falls* as output increases toward Q_1 . Also, profit *rises* as output increases *beyond* Q_1 , since $MR > MC$ for these moves. Since it never pays to increase *to* Q_1 , and profit rises when increasing *from* Q_1 , we know that Q_1 cannot possibly maximize the firm's profit.

But now look at point B, where the *MC curve crosses the MR curve from below*. You can see that when we are at an output level lower than Q^* , it always pays to increase Q , since $MR > MC$ for these moves. You can also see that, once we have arrived at Q^* , further increases will reduce profit, since $MC > MR$. Q^* is thus the profit-maximizing output level for this firm—the output level at which the *MC curve crosses the MR curve from below*.



MR and MC is as small as possible—not as large as possible.

A common error is assuming that the firm should produce the level of output at which the difference between *MR* and *MC* is as large as possible, like 2 or 3 units of output in Figure 2. Let's see why this is wrong. If the firm produces 2 or 3 units, it would leave many profitable increases in output unexploited—increases where $MR > MC$. As long as *MR* is even a tiny bit larger than *MC*, it pays to increase output, since doing so will add more to revenue than to cost. The firm should be satisfied only when the difference between

WHAT ABOUT AVERAGE COSTS?

You may have noticed that this chapter has discussed *most* of the cost concepts introduced in Chapter 6. But it has not yet referred to *average cost*. There is a good reason for this. We have been concerned about how much the firm should produce if it wishes to earn the greatest possible level of profit. To achieve this goal, the firm should produce more output whenever doing this *increases* profit, and it needs to know only *marginal cost* and *marginal revenue* for this purpose. The different types of average cost (*ATC*, *AVC*, and *AFC*) are simply irrelevant. Indeed, a common

TWO POINTS OF INTERSECTION

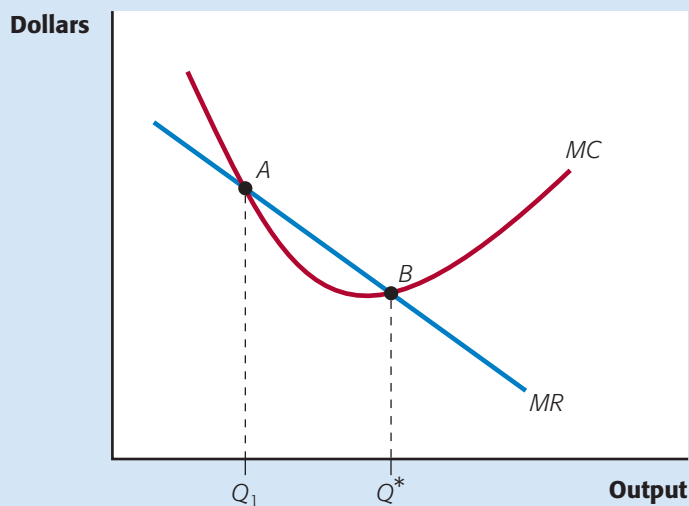


FIGURE 3

Sometimes the *MR* and *MC* curves intersect twice. The profit-maximizing level of output is always found where *MC* crosses *MR* from below.

error—sometimes made even by business managers—is to use *average* cost in place of *marginal* cost in making decisions.

For example, suppose a yacht maker wants to know how much his total cost will rise in the short run if he produces another unit of output. It is tempting—*but wrong*—for the yacht maker to reason this way: “My cost per unit (*ATC*) is currently \$50,000 per yacht. Therefore, if I increase production by 1 unit, my total cost will rise by \$50,000; if I increase production by 2 units, my total cost will rise by \$100,000, and so on.”

There are two problems with this approach. First, *ATC* includes many costs that are *fixed* in the short run—including the cost of all fixed inputs such as the factory and equipment and the design staff. These costs will *not* increase when additional yachts are produced, and they are therefore irrelevant to the firm’s decision making in the short run. Second, *ATC* *changes* as output increases. The cost per yacht may rise above \$50,000 or fall below \$50,000, depending on whether the *ATC* curve is upward or downward sloping at the current production level. Note that the first problem—fixed costs—could be solved by using *AVC* instead of *ATC*. The second problem—changes in average cost—remains even when *AVC* is used.

The correct approach, as we’ve seen in this chapter, is to use the *marginal cost* of a yacht and to consider increases in output one unit at a time. Alternatively, the firm can cut to the chase and produce where its *MC* curve crosses its *MR* curve from below. Average cost doesn’t help at all; it only confuses the issue.

Does this mean that all of your efforts to master *ATC* and *AVC*—their definitions, their relationship to each other, and their relationship to *MC*—were a waste of time? Far from it. As you’ll see, average cost will prove *very* useful in the chapters to come. You’ll learn that whereas marginal values tell the firm *what* to do, averages tell the firm *how well* it has done. But average cost should *not* be used in place of marginal cost as a basis for decisions.

THE IMPORTANCE OF MARGINAL DECISION MAKING: A BROADER VIEW

The *MC* and *MR* approach for finding the profit-maximizing output level is actually a very specific application of a more general principle:

The marginal approach to profit states that a firm should take any action that adds more to its revenue than to its cost.

Marginal approach to profit A firm maximizes its profit by taking any action that adds more to its revenue than to its cost.

In this chapter, the action we’ve been considering is to increase output by 1 unit, and we’ve learned that the firm should take this action whenever $MR > MC$. Since *MR* is how much this action *adds* to revenue and *MC* is how much it *adds* to cost, you can see that all along we have indeed been using a particular application of the more general marginal approach to profit. But this principle can be applied to *any other decision* facing the firm.

How can we be so bold as to say *any* decision facing the firm? Any action we can imagine that increases the firm’s revenue more than its costs will, *by definition*, increase its profits. Suppose that having the president of the company sing “The Star-Spangled Banner” each morning while standing on his head would add more to revenue than to cost. Then—to earn maximum profit—the firm should have the president do just that. But we needn’t dwell on absurd actions, since there are plenty of realistic ones to illustrate this principle.

PROFIT-MAXIMIZING CHOICE OF ADVERTISING

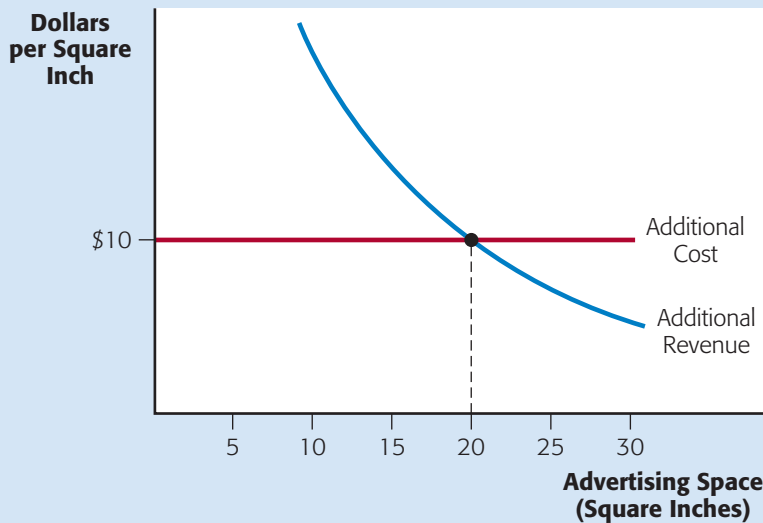


FIGURE 4

The firm's profit-maximizing level of advertising is found where the cost of an additional square inch of advertising space just equals the additional revenue that square inch will generate.

Consider the manager of a movie theater who must decide what size advertisement to take out in the local paper. Suppose each square inch of advertising space costs \$10. Figure 4 shows the theater's additional cost curve for advertising, which will be a horizontal line at \$10—each time the advertisement increases by 1 square inch, total cost *rises* by \$10.⁷ Suppose also that the larger the ad, the greater the revenue from selling tickets but that the additional revenue declines as the ad grows larger and larger in square-inch increments. In this case, the marginal revenue curve for advertising is downward sloping, like the one drawn in the figure. The marginal approach to profit tells us that the firm should keep increasing the size of the ad as long as the additional revenue from doing so is greater than the additional cost of doing so. Notice that we are assuming that the added viewers attracted by additional ads don't increase the theater's other costs. As you can see in the diagram, the firm should select the ad size where the two curves cross, or 20 square inches. Any ad larger or smaller than this would cause the theater to earn a smaller profit.

The marginal approach to profit explains much firm behavior that we see in the real world. It tells us that a profit-maximizing firm will take any action that adds more to its revenue than it adds to its costs, whether that action is lobbying the government for special treatment, extending the hours that a store is open, recalling a defective product, or giving free samples of merchandise. (You may want to think about how firms would decide on each of these actions using the marginal approach to profits.) Two more examples of this technique are discussed in greater detail in the "Using the Theory" section at the end of this chapter.

⁷ Don't be confused by this horizontal line. The "additional cost" curves we've considered so far have all been *MC* curves. These were U-shaped because they tracked the additional cost of producing more *output* when there were increasing and then diminishing returns to labor. In this example, we are looking at the additional cost *not* of producing more output, but of buying more *advertising space*. Since ad space costs a constant \$10 per square inch, the additional cost of 1 more square inch will be constant at \$10 as well.

DEALING WITH LOSSES

So far, we have dealt only with the pleasant case of profitable firms and how they select their profit-maximizing output level. But what about a firm that cannot earn a positive profit at *any* output level? What should it do? The answer depends on what time horizon we are looking at.

THE SHORT RUN AND THE SHUTDOWN RULE

In the short run, the firm must pay for its fixed inputs, because there is not enough time to sell them or get out of lease and rental agreements. But the firm can *still* make decisions about production. And one of its options is to *shut down*—to stop producing output, at least temporarily.

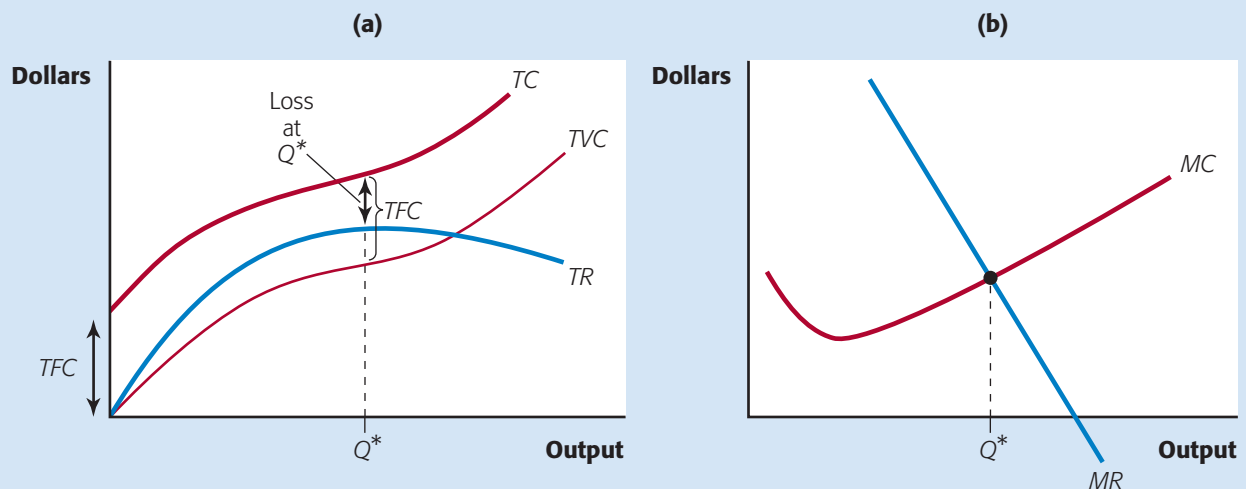
At first glance, you might think that a loss-making firm should always shut down its operation in the short run. After all, why keep producing if you are not making any profit? In fact, it makes sense for some unprofitable firms to continue operating.

Imagine a firm with the TC and TR curves shown in panel (a) of Figure 5 (ignore the TVC curve for now). No matter at what output level the firm produces, the TC curve lies above the TR curve, so it will suffer a loss—a negative profit. For this firm, the goal is still profit maximization. But now, the highest profit will be the one with the *least negative value*. In other words, profit maximization becomes *loss minimization*.

If the firm keeps producing, then the smallest possible loss is at an output level of Q^* , where the distance between the TC and TR curves is smallest. Q^* is also the output level we would find by using our marginal approach to profit (increasing output whenever that adds more to revenue than to costs). This is why, in panel (b) of Figure 5, the MC and MR curves must intersect at (or very close to) Q^* .

FIGURE 5

LOSS MINIMIZATION



The firm shown here cannot earn a positive profit at *any* level of output. If it produces anything, it will minimize its loss by producing where the vertical distance between TR and TC is smallest. Because TR exceeds TVC at Q^* , the firm will produce there in the short run.

The question is: Should this firm produce at Q^* and suffer a loss? The answer is yes—if the firm would lose even *more* if it stopped producing and shut down its operation. Remember that, in the short run, a firm must continue to pay its total fixed cost (TFC) no matter what level of output it produces—even if it produces nothing at all. If the firm shuts down, it will therefore have a loss equal to its TFC , since it will not earn any revenue. But if, by producing some output, the firm can cut its loss to something *less* than TFC , then it should stay open and keep producing.

To understand the shutdown decision more clearly, let's think about the firm's total variable costs. Business managers often call TVC the firm's *operating cost*, since the firm only pays these variable costs when it continues to operate. If a firm, by staying open, can earn *more* than enough revenue to cover its operating costs, then it is making an *operating profit* ($TR > TVC$). It should not shut down because its operating profit can be used to help pay its fixed costs. But if the firm cannot even cover its operating cost when it stays open—that is, if it would suffer an *operating loss* ($TR < TVC$)—it should definitely shut down. Continuing to operate only *adds* to the firm's loss, increasing the total loss beyond fixed costs.

This suggests the following guideline—called the **shutdown rule**—for a loss-making firm:

Let Q^ be the output level at which $MR = MC$. Then, in the short run:*
If $TR > TVC$ at Q^ , the firm should keep producing.*
If $TR < TVC$ at Q^ , the firm should shut down.*
If $TR = TVC$ at Q^ , the firm should be indifferent between shutting down and producing.*

Look back at Figure 5. At Q^* , the firm is making an operating profit, since its TR curve is above its TVC curve. This firm, as we've seen, should continue to operate.

Figure 6 is drawn for a different firm, one with the same TC and TVC curves as the firm in Figure 5, but with a lower TR curve. This firm *cannot* earn an operating profit, since its TR curve lies below its TVC curve everywhere—even at Q^* . This firm should shut down.

Shutdown rule In the short run, the firm should continue to produce if total revenue exceeds total variable costs; otherwise, it should shut down.

SHUT DOWN

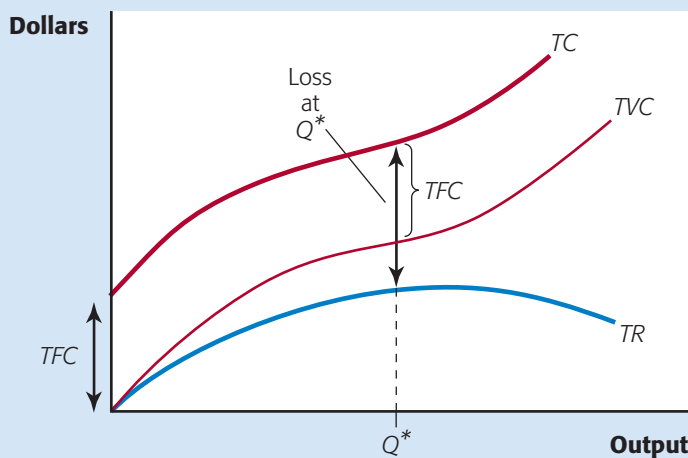


FIGURE 6

At Q^* , this firm's total variable cost exceeds its total revenue. The best policy is to shut down, produce nothing, and suffer a loss equal to TFC in the short run.

The shutdown rule is a powerful predictor of firms' decisions to stay open or cease production in the short run. It tells us, for example, why some seasonal businesses—such as ice cream shops in summer resort areas—shut down in the winter, when TR drops so low that it becomes smaller than TVC . And it tells us why producers of steel, automobiles, agricultural goods, and television sets will often keep producing output for some time even when they are losing money.

THE LONG RUN: THE EXIT DECISION

The shutdown rule applies only in the short run, a time horizon too short for the firm to escape its commitments to pay for fixed inputs such as plant and equipment. In fact, we only use the term *shut down* when referring to the short run.

But a firm can also decide to stop producing in the long run. In that case, we say the firm has decided to **exit** the industry.

The long-run decision to exit is different than the short-run decision to shut down. That's because in the long run, there *are* no fixed costs, since all inputs can be varied. Therefore, a firm that exits—by reducing all of its inputs to zero—will have *zero* costs (an option not available in the short run). And since exit also means zero revenue, a firm that exits will earn zero profit. When would a firm decide to exit and earn zero profit? When its only other alternative is to earn *negative* profit.

A firm should exit the industry in the long run when—at its best possible output level—it has any size loss at all.

We will look more closely at the exit decision and other long-run considerations in the next chapter.

Exit A permanent cessation of production when a firm leaves an industry.

Identify Goals and Constraints



THE GOAL OF THE FIRM REVISITED

So far in this chapter, we've assumed that a firm will make decisions that maximize its profit. That is, we've assumed that the firm is operated on behalf of its owners, who receive its profits. But in large firms, the owners hire managers to run the firm, and the managers, in turn, hire workers. Can we be sure that the workers and managers will maximize profit, as the owners desire? Or can workers and managers pursue their *own* goals that reduce the firm's profit?

THE PRINCIPAL-AGENT PROBLEM

The relationships among workers, managers, and owners within the firm are examples of what economists call *principal-agent relationships*.

A principal is a person or group who hires someone to do a job. An agent is the person hired to do that job.

Principal A person or group that hires someone to do a job.

Agent A person hired to do a job.

Principal-agent relationships can be seen everywhere in the economy. Your economics professor, for example, is an *agent* of the university and its trustees, who are the principals. If someone cleans your house or apartment, that person is your agent, and you are the principal. Principal-agent relationships are a natural consequence of specialization. If each of us specializes in one type of good or service, we will always find ourselves producing for others. Thus, principal-agent relationships, by enabling specialization in production, enable us all to enjoy high standards of living. But they are often problematic.

The principal-agent problem arises when an agent has (1) interests that conflict with the principal's, and (2) the ability to pursue those interests.

In practice, the principal-agent problem is unlikely to arise when the agent can be closely monitored. For example, when you hire someone to clean your house, a quick inspection will tell you whether the job was done properly. But in other cases, this kind of monitoring may be difficult or impossible. If you hire a baby-sitter, only the baby-sitter knows for sure how he or she treated the child in your absence. Similarly, most people who take their cars in for repairs do not have the expertise to tell whether the mechanic (their agent) has honestly diagnosed the problem or has even done all the work listed in the bill. In each of these cases, the agent knows something important that the principal does not, and this situation prevents proper monitoring and creates a principal-agent problem.

THE PRINCIPAL-AGENT PROBLEM AT THE FIRM

In business firms, there are two important principal-agent relationships—one between owners and managers and another between managers and workers. Each of these relationships can be plagued by conflicting goals. Managers are certainly interested in keeping their jobs and being promoted, giving them strong incentive to please owners by striving for high profits. But managers may also be interested in *other* things that can inflate the firm's costs and reduce its profits. These include high salaries, long vacations, the prestige of managing a great number of employees, or perks such as first-class air travel, large offices with nice views, and extravagant expense accounts. Managers may also use company property such as telephones, photocopiers, and computers for personal use.

What about workers? They are interested in keeping their jobs and being promoted and thus have some incentive to please their managers. But they, too, may have other goals that conflict with profit maximization—for example, not working too hard, taking long lunch breaks, and (like managers) using company property for personal use.

We can see that agents at the firm have interests that *conflict* with those of their principals. That is one of the requirements for a principal-agent problem. But is the second requirement satisfied? Do the agents have the *ability* to pursue their interests?

Indeed they do. Owners can, to some extent, monitor the actions of managers by reading quarterly reports of the firm's profits. But owners can never know as much as managers do about conditions and events at the firm, and so they cannot always tell whether management's decisions have led to the *highest profit possible*. Similarly, managers can observe how much output workers are producing, but because they cannot be everywhere at once, they cannot know whether workers are performing *as well as they could be*.

The principal-agent problem is likely to be more serious in large firms than in smaller ones. If you own a small ice cream shop and hire a helper to scoop ice cream while you deal with customers, you will know how hard your employee is working and the length of his breaks, and you will have a good chance of catching him if he begins stealing ice cream to bring home to his friends. But a large firm like Ben & Jerry's, with hundreds of managers and thousands of workers, has a serious problem on its hands. The stockholders cannot be sure that management expense accounts are being used solely for company purposes or that each promotion or pay hike is justified. And if management wants to expand the firm—say, by starting a new product line—are they doing it with higher profits in mind, or merely to increase opportunities for promotion within the firm? Similarly, managers have a

Principal-agent problem The situation that arises when an agent has interests that conflict with the principal's, and has the ability to pursue those interests.

hard time preventing hourly workers from slacking off when they aren't being watched or ensuring that each worker punches in his or her own time card.

Economists have thought a lot about the principal-agent problem. Using models that view the firm as a collection of different groups—workers, managers, and owners—economics is discovering new ways in which the principal-agent problem affects the behavior of business firms, as well as nonprofit organizations like hospitals and universities and government agencies like police departments and the military. These models help us understand why firms can and do deviate from profit-maximizing behavior and how different types of supervision and pay arrangements can help solve the principal-agent problem in different types of organizations.

THE ASSUMPTION OF PROFIT MAXIMIZATION

From the preceding discussion, it might seem that the profit-maximizing assumption underlying most of this chapter is somewhat naive. After all, because of the principal-agent problem, the firm may *not* always maximize profits, even though that is what the owners want. Why, then, did we base our theory of firm behavior on such a simple assumption? Why not go back and view the firm in light of the principal-agent problem?

For one very good reason: The assumption of profit maximization, while not completely accurate, is reasonably accurate in most cases. While profit maximization is not the *only* goal of decision makers at the firm, it seems to be the driving force behind most management decisions. Remember that an economic model *abstracts* from reality. To stay simple and comprehensible, it leaves out many real-world details and includes only what is relevant for the purpose at hand. If the purpose is to explain *conflict within the firm*, or *deviations from profit-maximizing behavior*, the principal-agent problem would be a central element of the model. But when the purpose is to explain how firms decide what price to charge and how much to produce, how much to spend on advertising, or whether to shut down or continue operating in the short run, the assumption of profit maximization has proven to be sufficient. It explains what firms actually do with reasonable—and often remarkable—accuracy.

Even in larger firms, profit maximization seems to be the most important force driving firm decisions. How can this be, when principal-agent problems can be so serious in large firms? In part, because the market provides some *solutions* to the principal-agent problem of large firms. Although far from perfect, they prevent the firm from straying *too* far from the profit-maximizing course.

For example, if a firm's managers significantly inflate costs and reduce profit, the company's stock will become less attractive to potential buyers, and its price will fall. Management will then face one of two consequences: (1) a **stockholder revolt**, in which owners, seeing that their firm is less profitable than others in the industry, replace the management team with another that promises to do better, or (2) a **hostile takeover**, in which outsiders buy up a majority of the firm's shares at low prices, often with the goal of sacking the current managers and replacing them with better ones. (The term *hostile* is from the viewpoint of the current managers.) In large corporations, poor management decisions can reduce profits by millions or even billions of dollars. Since there is so much at stake, stockholder revolts and hostile takeovers are not at all uncommon.

A recent example of a stockholder revolt occurred at USAir. With negative profits each year from 1990 to 1994, the company was headed toward bankruptcy. Stockholders blamed the management. In January 1996, the airline's board of directors, elected by its shareholders, took the first step toward replacing the manage-

Stockholder revolt When owners, dissatisfied with the profits they are earning, replace the firm's management team.

Hostile takeover When outsiders buy up a firm's shares with the goal of replacing the management team and increasing profits.

ment team by bringing on Stephen Wolf as the new chief executive. Wolf undertook a series of dramatic actions that included placing a large order for new planes, aggressive cost cutting, and a change of the firm's name to US Airways. Over the next several years, more changes were implemented, profits increased dramatically, and the market value of US Airways stock grew by 1,000 percent.

A hostile takeover is more complicated than a stockholder revolt, because the current managers, who are the likely losers in a hostile takeover, can attempt a variety of measures to foil it. A recent example—and one that illustrates some of the vocabulary you will see in the business pages of your local newspaper—is the case of John Labatt Ltd., the Canadian beer maker. When Labatt's profits dwindled in late 1994 and the price of its stock dropped to 19 Canadian dollars per share, the financial press blamed poor management decisions. Sensing the firm was ripe for a hostile takeover, Labatt's managers tried to forestall it with a variety of actions designed to make the firm less attractive to outsiders, such as selling off valuable assets or taking on especially risky new projects. But at the company's annual shareholder meeting, the shareholders voted to reject these moves. They *wanted* a takeover so they could sell their shares at a price reflecting the firm's *potential* profit under new management. Sure enough, the hostile takeover attempt came in early 1995 when an outsider—Onex Corporation—offered to buy all Labatt shares for 24 Canadian dollars each. Labatt's management responded by trying to arrange a **friendly takeover** by another firm deemed less likely to fire them. The **white knight** that came to their rescue was Interbrew SA, a Belgian beer maker. Interbrew was eager to expand into Canada to fulfill its own strategic plan, and it made an even more generous offer to Labatt's shareholders—28.5 Canadian dollars per share. In June 1995, Labatt's board of directors, representing the shareholders, happily agreed to the acquisition. A new president was soon put in place, and the firm went on to increase both its domestic market share and its exports to the United States.

The threat of being fired is a powerful incentive for managers to worry about profits, but many firms use positive incentives as well. End-of-year *bonuses*—payments in addition to regular wages or salary—are often tied to total profit at the firm. In many cases, these bonuses are a substantial portion of a manager's total compensation. **Stock options**—which give managers the right to buy shares of the company's stock at a prespecified price—are another positive incentive. If the managers perform well, the market value of the firm's stock will rise. The managers can then *exercise* their stock options—purchasing the stock at the prespecified low price—and, if they choose, they can immediately sell the stock at the higher, market price and pocket the difference.

To see how stock options work, let's use the case of Floyd Hall (no relation to any author of this book). In June 1995, Hall was hired as Kmart corporation's new president. His yearly salary was about \$1 million. But he was also given stock options to buy 3 million shares of Kmart stock at the then-current price of \$12.38 per share. If he and his management team could increase the company's profits and convince potential stockholders that earnings growth would continue, then Kmart stock would become more attractive, and its price would rise. For example, if the stock rose \$20 per share, then Floyd Hall would be able to exercise his options: He could buy 3 million shares at \$12.38 each (a total of \$37 million), and then immediately sell them at \$20 each (a total of \$60 million), making a tidy gain of \$23 million.

Needless to say, Hall tried to do everything he could to raise Kmart's profits over the next several years. The results were mixed. Profits *did* grow, but Hall was unable to convince stockholders and potential stockholders that rapid growth in earnings would *continue*, especially with Kmart facing aggressive competitors like Wal-Mart



The Wharton School of the University of Pennsylvania maintains a Web page on corporate governance (<http://www-management.wharton.upenn.edu/leadership/governance/>). Click on "Corporate Control" for more information on factors that help discipline corporate managers.

Friendly takeover When a firm's management arranges a takeover by another firm deemed unlikely to fire them.

White knight A firm that undertakes a friendly takeover.

Stock options Rights to purchase shares of stock at a prespecified price.



The Foundation for Enterprise Development (<http://www.fed.org/about/index.html>) provides information about employee ownership and stock options as means of motivating employees.

and Target. In April 2000, Kmart stock was actually selling at around \$9 per share—significantly *below* the price when Hall had been hired. Hall's stock options were therefore worthless: What would be the point of exercising the right to buy shares at \$12.38 when they could be purchased in the market for \$9? Nevertheless, since Hall still owned the options, he had a powerful incentive to keep trying. (And if he failed to raise the stock's price, he would probably not retain his position much longer. By the end of 1999, at least two large supermarket chains—Safeway and Kroger—were reportedly eyeing Kmart as a potential takeover candidate.)

Incentives like bonuses and stock options on the one hand and threats of stockholder revolt or hostile takeover on the other are usually enough to keep management's eye on company profits. When this carrot-and-stick approach doesn't work, then *actual* revolts or takeovers—and the dumping or disciplining of management—ensure that poor managers do not survive for long. At any given time, therefore, we can expect *most* managers to try to maximize profits *most* of the time.

Similar mechanisms help ensure that hourly workers contribute to maximum profit at the firm. There are, indeed, plenty of opportunities to shirk or otherwise frustrate management's goals, but these can be pursued *only up to a point*, or the worker can expect to be fired. Television's Homer Simpson has on numerous occasions spilled coffee into the control panel of the nuclear reactor he operates, stolen expensive equipment for home use, and taken snoozes while the reactor goes into meltdown. Nobody in the real world would survive in a job with his record.

For all of the reasons just discussed, assuming that firms maximize profit for their owners is not too far off the mark. The principal-agent problem does exist, and it helps us understand many aspects of firm behavior, such as conflicts that arise within the firm, the structure of pay, and the methods used to supervise workers and managers. However, if our goal is to achieve a reasonably accurate prediction of firm decisions, profit maximization works pretty well.

Ask a physicist to predict when a bowling ball dropped from the top of the Empire State building will hit the ground, and her calculations will assume it is falling in a perfect vacuum. Ask an economist to predict how much output a firm will produce and what price it will charge, and he will assume the firm's only goal is to maximize profit. In both cases the assumptions lead to very accurate—if not perfectly accurate—predictions.

Using the THEORY

GETTING IT WRONG AND GETTING IT RIGHT

Today, almost all managers have a good grasp of the concepts you've learned in this chapter, largely because microeconomics has become an important part of every business school curriculum. But if we go back a few decades—to when fewer managers had business degrees—we can find two examples of how management's failure to understand the basic theory of the firm led to serious errors. In one case, ignorance of the theory caused a large bank to go bankrupt; in the other, an airline was able to outperform its competitors because *they* remained ignorant of the theory.

GETTING IT WRONG: THE FAILURE OF FRANKLIN NATIONAL BANK

In the mid-1970s, Franklin National Bank—one of the largest banks in the United States—went bankrupt. The bank's management had made several errors, but we will focus on the most serious one.

First, a little background. A bank is very much like any other business firm: It produces output (in this case a service, making loans) using a variety of inputs (land, labor, capital, and raw materials). The price of the bank's output is the interest rate it charges to borrowers. For example, with a 5 percent interest rate, the price of each dollar in loans is 5 cents per year.

Unfortunately for banks, they must also *pay* for the money they lend out. The largest source of funds is customer deposits, for which the bank must pay interest. If a bank wants to lend out *more* than its customers have deposited, it can obtain funds from a second source, the *federal funds market*, where banks lend money to one another. To borrow money in this market, the bank will usually have to pay a higher interest rate than it pays on customer deposits.

In mid-1974, John Sadlik, Franklin's chief financial officer, asked his staff to compute the average cost to the bank of a dollar in loanable funds. At the time, Franklin's funds came from three sources, each with its own associated interest cost:

Source	Interest Cost
Checking Accounts	2.25 percent
Savings Accounts	4 percent
Borrowed Funds	9–11 percent

What do these numbers tell us? First, each dollar deposited in a Franklin *checking* account cost the bank 2.25 cents per year,⁸ while each dollar in a *savings* account cost Franklin 4 cents. Also, Franklin—like other banks at the time—had to pay between 9 and 11 cents on each dollar borrowed in the federal funds market. When Franklin's accountants were asked to figure out the average cost of a dollar in loans, they divided the total cost of funds by the number of dollars lent out. The number they came up with was 7 cents.

This average cost of 7 cents per dollar is an interesting number, but, as we know, it should have *no relevance to a profit-maximizing firm's decisions*. And this is where Franklin went wrong. At the time, all banks—including Franklin—were charging interest rates of 9 to 9.5 percent to their best customers. But Sadlik decided that since money was costing an *average* of 7 cents per dollar, the bank could make a tidy profit by lending money at 8 percent—earning 8 cents per dollar. Accordingly, he ordered his loan officers to approve any loan that could be made to a reputable borrower at 8 percent interest. Needless to say, with other banks continuing to charge 9 percent or more, Franklin National Bank became a very popular place from which to borrow money.

But where did Franklin get the additional funds it was lending out? That was a problem for the managers in *another* department at Franklin, who were responsible for *obtaining* funds. It was not easy to attract additional checking and savings account deposits, since, in the 1970s, the interest rate banks could pay was regulated by the government. That left only one alternative: the federal funds market. And this is exactly where Franklin went to obtain the funds pouring out of its lending department. Of course, these funds were borrowed not at 7 percent, the average cost of funds, but at 9 to 11 percent, the cost of borrowing in the federal funds market.

⁸ This cost was not actually a direct interest payment to depositors, since in the 1970s banks generally did not pay interest on checking accounts. But banks *did* provide free services such as check clearing, monthly statements, free coffee, and even gifts to their checking account depositors, and the cost of these freebies was computed to be 2.25 cents per dollar of deposits.

To understand Franklin's error, let's look again at the average cost figure it was using. This figure included an irrelevant cost: the cost of funds obtained from customer deposits. This cost was irrelevant to the bank's lending decisions, since *additional* loans would not come from these deposits, but rather from the more expensive federal funds market. Further, this average figure was doomed to rise as Franklin expanded its loans. How do we know this? The *marginal* cost of an additional dollar of loans—9 to 11 cents per dollar—was greater than the *average* cost—7 cents. As you know, whenever the marginal is greater than the average, it pulls the average up. Thus, Franklin was basing its decisions on an average cost figure that not only included irrelevant sunk costs but was bound to increase as its lending expanded.

More directly, we can see Franklin's error through the lens of the marginal approach. The *marginal revenue* of each additional dollar lent out at 8 percent was 8 cents, while the *marginal cost* of each additional dollar—since it came from the federal funds market—was 9 to 11 cents. *MC* was greater than *MR*, so Franklin was actually losing money each time its loan officers approved another loan! Not surprisingly, these loans—which never should have been made—caused Franklin's profits to *decrease*, and within a year the bank had lost hundreds of millions of dollars. This, together with other management errors, caused the bank to fail.⁹

GETTING IT RIGHT: THE SUCCESS OF CONTINENTAL AIRLINES

Continental Airlines was doing something that seemed like a horrible mistake. All other airlines at the time were following a simple rule: They would only offer a flight if, on average, 65 percent of the seats could be filled with paying passengers, since only then could the flight break even. Continental, however, was flying jets filled to just 50 percent of capacity and was actually expanding flights on many routes. When word of Continental's policy leaked out, its stockholders were angry, and managers at competing airlines smiled knowingly, waiting for Continental to fail. Yet Continental's profits—already higher than the industry average—continued to grow. What was going on?

There *was*, indeed, a serious mistake being made—but by the *other* airlines, not Continental. This mistake should by now be familiar to you: using average cost instead of marginal cost to make decisions. The “65 percent of capacity” rule used throughout the industry was derived more or less as follows: The total cost of the airline for the year (*TC*), was divided by the number of flights during the year (*Q*) to obtain the average cost of a flight ($TC/Q = ATC$). For the typical flight, this came to about \$4,000. Since a jet had to be 65 percent full in order to earn ticket sales of \$4,000, the industry regarded any flight that repeatedly took off with less than 65 percent as a money loser and canceled it.

As usual, there are two problems with using *ATC* in this way. First, an airline's average cost per flight includes many costs that are fixed and are therefore irrelevant to the decision to add or subtract a flight. These include the cost of running the reservations system, paying interest on the firm's debt, and fixed fees for landing rights at airports—none of which would change if the firm added or subtracted a flight. Also, average cost ordinarily *changes* as output changes, so it is wrong to assume it is constant in decisions about *changing* output.



⁹ For more information on the failure of Franklin National Bank, see Sanford Rose, “What Really Went Wrong at Franklin National?” *Fortune*, October 1974, pp. 118–226.

Continental's management, led by its vice-president of operations, had decided to try the marginal approach to profit. Whenever a new flight was being considered, every department within the company was asked to determine the *additional* cost they would have to bear. Of course, the only additional costs were for additional *variable* inputs, such as additional flight attendants, ground crew personnel, in-flight meals, and jet fuel. These additional costs came to only about \$2,000 per flight. Thus, the *marginal* cost of an additional flight—\$2,000—was significantly less than the marginal revenue of a flight filled to 65 percent of capacity—\$4,000. The marginal approach to profits tells us that when $MR > MC$, output should be increased, which is just what Continental was doing. Indeed, Continental correctly drew the conclusion that the marginal revenue of a flight filled at even 50 percent of capacity—\$3,000—was *still* greater than its marginal cost, and so offering the flight would increase profit. This is why Continental was expanding routes even when it could fill only 50 percent of its seats.

In the early 1960s, Continental was able to outperform its competitors by using a secret—the marginal approach to profits. Today, of course, the secret is out, and all airlines use the marginal approach when deciding which flights to offer.¹⁰

¹⁰ For more information about Continental's strategy, see "Airline Takes the Marginal Bone," *Business Week*, April 20, 1963, pp. 111–114.

S U M M A R Y

In economics, we view the firm as a single economic decision maker with the goal of maximizing the owners' profit. Profit is total revenue minus *all* costs of production—explicit and implicit. In their pursuit of maximum profit, firms face two constraints. One is embodied in the demand curve the firm faces; it indicates the maximum price the firm can charge to sell any amount of output. This constraint determines the firm's revenue at each level of production. The other constraint is imposed by costs: More output always means greater costs. In choosing the profit-maximizing output, the firm must consider both revenues and costs.

One approach to choosing the optimal level of output is to measure profit as the difference between total revenue and total cost at each level of output, and then select the output level at which profit is greatest. An alternate approach uses *marginal revenue* (MR)—the change in total revenue from producing one more unit of output—and *mar-*

ginal cost (MC)—the change in total cost from producing one more unit. The firm should increase output whenever $MR > MC$, and lower output when $MR < MC$. The profit-maximizing output level is the one closest to the point where $MR = MC$.

If profit is negative, but total revenue exceeds total variable cost, the firm should continue producing in the short run. Otherwise, it should shut down and suffer a loss equal to its fixed cost.

All of this assumes that the firm will be run with the owners' best interests in mind. However, a principal-agent problem may exist in which workers or managers pursue their own interests to the detriment of the owners' interests. Still, firms' owners have come up with a variety of incentives to keep managers' and workers' eyes on profits. The assumption of profit maximization, while not completely accurate, is accurate enough to be useful.

K E Y T E R M S

accounting profit
economic profit
demand curve facing the firm
total revenue

loss
marginal revenue
marginal approach to profit
shutdown rule
exit

principal agent
principal-agent problem
stockholder revolt
hostile takeover

friendly takeover
white knight
stock options

R E V I E W Q U E S T I O N S

1. What is the difference between accounting profit and economic profit?
2. Can a firm earn an accounting profit at the same time it is suffering an economic loss? If so, give a numerical example. Can a firm earn an economic profit at the same time it is suffering an accounting loss? Again, if this is possible, give a numerical example.
3. Name two contributions to the production process for which profit is a payment. Pick a local business, and briefly explain how the entrepreneur behind it has made each of these contributions.
4. What are the three kinds of demand curve we have studied so far in this book? What does each tell us?
5. What are the constraints on the firm's ability to earn profit? How does each constraint arise?
6. How does the firm select the level of output where profit is greatest in:
 - a. The total revenue and total cost approach?
 - b. The marginal revenue and marginal cost approach? How is each approach illustrated graphically?
7. What are the two conditions necessary for the principal-agent problem to arise?
8. Discuss the following statement: "The assumption that a firm's only goal is profit maximization is completely unrealistic. Different groups within a company typically pursue their own agendas, which frequently have nothing to do with profit."
9. What is the difference, if any, between a hostile takeover and a stockholder revolt?
10. What forces help ensure that firms actually do seek to maximize their profits?

P R O B L E M S A N D E X E R C I S E S

1. You have a part-time work/study job at the library that pays \$10 per hour, 3 hours per day on Saturdays and Sundays. Some friends want you to join them on a weekend ski trip leaving Friday night and returning Monday morning. They estimate your share of the gas, motel, lift tickets, and other expenses to be around \$30. What is your total cost (considering both explicit and implicit costs) for the trip?
2. Until recently, you worked for a software development firm at a yearly salary of \$35,000. Now, you decide to open your own business. Planning to be the next Bill Gates, you quit your job, cash in a \$10,000 savings account (which pays 5 percent interest), and use the money to buy the latest computer hardware to use in your business. You also convert a basement apartment in your house, which you have been renting for \$250 a month, into a workspace for your new software firm.

You lease some office equipment for \$3,600 a year and hire two part-time programmers, whose combined salary is \$25,000 a year. You also figure it costs around \$50 a month to provide heat and light for your new office.

 - a. What are the total annual explicit costs of your new business?
 - b. What are the total annual implicit costs?
 - c. At the end of your first year, your accountant cheerily informs you that your total sales for the year amounted to \$55,000. She congratulates you on a profitable year. Are her congratulations warranted? Why or why not?
3. The following data are price/quantity/cost combinations for Titan Industry's mainframe computer division:

Quantity	Price per Unit	Total Cost of Production
0	above \$225,000	\$200,000
1	\$225,000	\$250,000
2	\$175,000	\$275,000
3	\$150,000	\$325,000
4	\$125,000	\$400,000
5	\$90,000	\$500,000

 - a. What is the marginal revenue if output rises from 2 to 3 units? (*Hint:* Calculate total revenue at each output level first.) What is the marginal cost if output rises from 4 to 5 units?
 - b. What quantity should Titan produce to maximize total revenue? Total profit?
 - c. What is Titan's fixed cost? How do Titan's marginal costs behave as output increases? Provide a plausible explanation as to why a computer manufacturer's marginal costs might behave in this way.
4. Discuss how serious you think the principal-agent problem would be in each of the following situations:
 - a. You leave your computer at a shop for repair.
 - b. You and a friend buy and run a business together.
 - c. A couple hires you to house-sit while they're in Europe for 2 months.

- d. An employee owns shares in the company for which he works. His supervisor is out sick for a week.
5. Each entry in this table shows marginal revenue and marginal cost when a firm increases output to the given quantity:

Quantity	MR	MC
10	30	40
11	29	35
12	27	30
13	25	25
14	23	20
15	21	15
16	19	19
17	17	23

What is the profit-maximizing level of output?

6. The following tables give information about demand and total cost for two firms. In the short run, how much should each produce?

Quantity	Price	Total Cost
0	above \$125	\$250
1	\$125	\$400
2	\$100	\$500
3	\$ 75	\$550
4	\$ 50	\$600
5	\$ 25	\$700

Quantity	Price	Total Cost
0	above \$500	\$ 500
1	\$500	\$ 700
2	\$400	\$ 900
3	\$300	\$1,100
4	\$200	\$1,300
5	\$100	\$1,500

CHALLENGE QUESTIONS

- A firm's *marginal profit* can be defined as the change in its profit when output increases by one unit.
 - Compute the marginal profit for each change in Ned's Beds' output in Table 1.
 - State a complete rule for finding the profit-maximizing output level in terms of marginal profit.
- Howell Industries specializes in precision plastics. Their latest invention promises to revolutionize the electronics industry, and they have already made and sold 75 of the miracle devices. They have estimated average costs as given in the following table:

Unit	AC
74	\$10,000
75	\$12,000
76	\$14,000

Backus Electronics has just offered Howell \$150,000 if they will produce the 76th unit. Should Howell accept the offer and manufacture the additional device?

EXPERIENTIAL EXERCISES

- Make a rough estimate of the economic profit earned by a corporation of your choice. To do this, go to the Morningstar.com database at <http://www.morningstar.com>. Choose "Stocks" from the tabs at the top of the page, then enter the name of your firm in the "Quotes and Reports" box. Choose "Historical Overview" on the left. When you get the data, find the firm's most recent Net Income. This is equivalent to its accounting profit. To estimate economic profit, you will have to estimate the opportunity cost of the funds the owners have invested in the firm. In the pie chart at the bottom of the Web page, find "Shareholders Equity." Multiply this by



the current interest rate—you can use 6.5 percent—to approximate the owners' implicit costs. Subtract this from Net Income and you have a rough approximation to the firm's economic profit. Is your firm earning a positive profit, or suffering a loss?

- Use Infotrac or the *Wall Street Journal* to find an example of the principal-agent problem. In your example, who is the agent and who is the principal? Is there evidence that the agent's interests conflict with the principal's, and that the agent has the ability to pursue his or her interests? How can this problem be resolved?

CHAPTER

8

PERFECT COMPETITION

CHAPTER OUTLINE

What Is Perfect Competition?

The Three Requirements for Perfect Competition
Is Perfect Competition Realistic?

The Perfectly Competitive Firm

Goals and Constraints of the Competitive Firm
Cost and Revenue Data for a Competitive Firm
Finding the Profit-Maximizing Output Level
Measuring Total Profit
The Firm's Short-Run Supply Curve

Competitive Markets in the Short Run

The (Short-Run) Market Supply Curve
Short-Run Equilibrium

Competitive Markets in the Long Run

Profit and Loss and the Long-Run Equilibrium
The Notion of Zero Profit in Perfect Competition
Perfect Competition and Plant Size
A Summary of the Competitive Firm in the Long Run

What Happens When Things Change?

A Change in Demand
Market Signals and the Economy

Using the Theory: Changes in Technology

Market structure The characteristics of a market that influence how trading takes place.

No one knows exactly how many different types of goods and services are offered for sale in the United States, but the number must be somewhere in the tens of millions. Each of these goods is traded in a market, where buyers and sellers come together, and these markets have several things in common. Sellers want to sell at the *highest* possible price; buyers seek the *lowest* possible price; and all trade is *voluntary*. But here, the similarity ends.

When we observe buyers and sellers in action, we see that different goods and services are sold in vastly different ways. Take advertising, for example. Every day, we are inundated with sales pitches on television, radio, and newspapers for a long list of products: toothpaste, perfume, automobiles, Internet Web sites, cat food, banking services, and more. But have you ever seen a farmer on television, trying to convince you to buy *his* wheat, rather than the wheat of other farmers? Do shareholders of major corporations like General Motors sell their stock by advertising in the newspaper? Why, in a world in which virtually *everything* seems to be advertised, do we not see ads for wheat, corn, crude oil, gold, copper, shares of stock, or foreign currency?

Or consider profits. Anyone starting a business hopes to make as much profit as possible. Yet some companies—Microsoft, Quaker Oats, and Pepsico, for example—earn sizable profit for their owners year after year, while at other companies, such as Trans World Airlines and most small businesses, economic profit is generally low.

We could say, “That’s just how the cookie crumbles,” and attribute all of these observations to pure randomness. But economics is all about *explaining* such things—finding patterns amidst the chaos of everyday economic life. When economists turn their attention to differences in trading, such as these, they think immediately about *market structure*:

By market structure, we mean all the characteristics of a market that influence the behavior of buyers and sellers when they come together to trade.

To determine the structure of any particular market, we begin by asking three simple questions:

1. *How many* buyers and sellers are there in the market?
2. Is each seller offering a *standardized product*, more or less indistinguishable from that offered by other sellers, or are there significant differences between the products of different firms?
3. Are there any *barriers to entry or exit*, or can *outsiders* easily enter and leave this market?

The answers to these questions help us to classify a market into one of four basic types: *perfect competition*, *monopoly*, *monopolistic competition*, or *oligopoly*. The subject of this chapter is perfect competition. In the next two chapters, we'll look carefully at the other market structures.

WHAT IS PERFECT COMPETITION?

Does the phrase “perfect competition” sound familiar? It should, because you encountered it earlier, in Chapter 3. There you learned (briefly) that the famous supply and demand model explains how prices are determined in *perfectly competitive markets*. Now we're going to take a much deeper and more comprehensive look at perfectly competitive markets. By the end of this chapter, you will understand very clearly how perfect competition and the supply and demand model are related.

Let's start with the word “competition” itself. When you hear that word, you may think of an intense, personal rivalry, like that between two boxers competing in a ring or two students competing for the best grade in a small class. But there are other, less personal forms of competition. If you took the SAT exam to get into college, you were competing with thousands of other test takers in rooms just like yours, all across the country. But the competition was *impersonal*: You were trying to do the best that you could do, trying to outperform others in general, but not competing with any one individual in the room. In economics, the term “competition” is used in the latter sense. It describes a situation of diffuse, impersonal competition in a highly populated environment. The market structure you will learn about in this chapter—perfect competition—is an example of this notion.

THE THREE REQUIREMENTS OF PERFECT COMPETITION

Perfect competition is a market structure with three important characteristics:

1. *There are large numbers of buyers and sellers, and each buys or sells only a tiny fraction of the total quantity in the market.*
2. *Sellers offer a standardized product.*
3. *Sellers can easily enter into or exit from the market.*

These three conditions probably raise more questions than they answer, so let's see what each one really means.

A Large Number of Buyers and Sellers. In perfect competition, there must be many buyers and sellers. How many? It would be nice if we could specify a number—like 32,456—for this requirement. Unfortunately, we cannot, since what constitutes a large number of buyers and sellers can be different under different conditions. What is important is this:



Characterize the Market

Perfect competition A market structure in which there are many buyers and sellers, the product is standardized, and sellers can easily enter or exit the market.

In a perfectly competitive market, the number of buyers and sellers is so large that no individual decision maker can significantly affect the price of the product by changing the quantity it buys or sells.

Think of the world market for wheat. On the selling side, there are hundreds of thousands of individual wheat farmers—more than 250,000 in the United States alone. Each of these farmers produces only a tiny fraction of the total market quantity. If any one of them were to double, triple, or even quadruple its production, the impact on total market quantity and market price would be negligible. The same is true on the buying side: There are so many small buyers that no one of them can affect the market price by increasing or decreasing its quantity demanded.

Most agricultural markets conform to the large-number/small-participant requirement, as do markets for precious metals such as gold and silver and markets for the stocks and bonds of large corporations. For example, more than 2 million shares of General Motors stock are bought and sold *every day*, at a price (as this is written) of about \$70 per share. A decision by a single stockholder to sell say, \$1 million dollars worth of this stock—about 14,000 shares—would cause only a barely noticeable change in quantity supplied on any given day.

But now think about the U.S. market for athletic shoes. Here, four large producers—Nike, Reebok, Adidas, and FILA—account for 75 percent of total sales. If any one of these producers decided to change its output by even 10 percent, the impact on total quantity supplied—and market price—would be *very* noticeable. The market for athletic shoes thus fails the large-number/small-participant requirement, so it is not an example of perfect competition.

A Standardized Product Offered by Sellers. In a perfectly competitive market, buyers do not perceive significant differences between the products of one seller and another. For example, buyers of wheat will ordinarily have no preference for one farmer's wheat over another's, so wheat would surely pass the standardized product test. The same is true of many other agricultural products—corn syrup and soybeans. It is also true of commodities like crude oil or pork bellies, precious metals like gold or silver, and financial instruments such as the stocks and bonds of a particular firm. (One share of AT&T stock is indistinguishable from another.)

When buyers *do* notice significant differences in the outputs of different sellers, the market is not perfectly competitive. For example, most consumers perceive differences among the various brands of coffee on the supermarket shelf and may have strong preferences for one particular brand. Coffee, therefore, fails the standardized product test of perfect competition. Other goods and services that would fail this test include personal computers, automobiles, houses, colleges, and medical care.

Easy Entry into and Exit from the Market. Entry into a market is rarely free—a new seller must always incur *some* costs to set up shop, begin production, and establish contacts with customers. But a perfectly competitive market has no *significant* barriers to discourage new entrants: Any firm wishing to enter can do business on the same terms as firms that are already there. For example, anyone with the right background in farming can begin planting and growing wheat by paying the same costs as veteran wheat farmers. The same is true of anyone wishing to open up a dry cleaning shop, a new restaurant, or an E-commerce consulting firm for companies that want to sell more effectively over the Internet. Each of these examples would pass the free-entry test of perfect competition.

In many markets, however, there are significant barriers to entry. These are often imposed by government. Sometimes, the government imposes *absolute* restrictions on the number of market participants allowed. For example, the number of taxicabs licensed to operate in New York City is fixed, determined by the city government. From the 1930s until 1996, this number was set at 11,787. In the late 1990s, the city finally increased the number of taxi licenses, but only by a few hundred, bringing the total to 12,187. Unless the city issues more licenses in the future, true entry into this market will be impossible—the licenses may change hands, but the total number of legally operated taxis cannot increase. Another example of government barriers to entry is *zoning laws*. These place strict limits on how many businesses such as movie theaters, supermarkets, or hotels can operate in a local area.

Barriers to entry can also arise without any government action, simply because existing sellers have an important advantage new entrants cannot duplicate. The brand loyalty enjoyed by existing producers of breakfast cereals, instant coffee, and soft drinks would require a new entrant to wrest customers away from existing firms—a very costly undertaking. Or significant economies of scale may give existing firms a cost advantage over new entrants. We will discuss these and other barriers to entry in more detail in later chapters.

In addition to easy entry, perfect competition requires easy *exit*: A firm suffering a long-run loss must be able to sell off its plant and equipment and leave the industry for good, without obstacles. Some markets satisfy this requirement, and some do not. Plant-closing laws or union agreements can require lengthy advance notice and high severance pay when workers are laid off. Or capital equipment may be so highly specialized—like an assembly line designed to produce just one type of automobile—that it cannot be sold off if the firm decides to exit the market. These and other barriers to exit do not conform to the assumptions of perfect competition.

IS PERFECT COMPETITION REALISTIC?

The three assumptions a market must satisfy to be perfectly competitive (or just “competitive,” for short) are rather restrictive. Do any markets satisfy all these requirements? How broadly can we apply the model of perfect competition when we think about the real world?

First, remember that perfect competition is a *model*—an abstract representation of reality. No model can capture *all* of the details of a real-world market, nor should it. Still, in some cases, the model fits remarkably well. We have seen that the market for wheat, for example, passes all three tests for a competitive market: many buyers and sellers, standardized output, and easy entry and exit. Indeed, most agricultural markets satisfy the strict requirements of perfect competition quite closely, as do many financial markets and some markets for consumer goods and services.

But in the vast majority of markets, one or more of the assumptions of perfect competition will, in a strict sense, be violated. This might suggest that the model can be applied only in a few limited cases. Yet when economists look at real-world markets, they use perfect competition more often than any other market structure. Why is this?

First, with perfect competition, we can use simple techniques to make some strong predictions about a market’s response to changes in consumer tastes, technology, and government policies. While other types of market structure models also yield valuable predictions, they are often more cumbersome and their predictions less definitive. Second, economists believe that many markets—while not strictly perfectly competitive—come *reasonably* close. The more closely a real-world market fits the model, the more accurate our predictions will be when we use it.

We can even—with some caution—use the model to analyze markets that violate all three assumptions. Take the worldwide market for television sets. There are about a dozen major sellers in this market. Each of them knows that its output decisions will have *some* effect on the market price, but no one of them can have a *major* impact on price. Consumers do recognize the difference between one brand and another, but their preferences are not very strong, and most recognize that quality has become so standardized that all brands are actually close substitutes for each other. And there are indeed barriers to entry—existing firms have supply and distribution networks that would be difficult for new entrants to replicate—but these barriers are not so great that they would keep out new entrants in the face of high potential profit. Thus, although the market for televisions does not strictly satisfy any of the requirements of perfect competition, it is not *too* far off on any one of them. The model will not perform as accurately for televisions as it does for wheat, but, depending on how much accuracy we need, it may do just fine.

In sum, perfect competition can approximate conditions and yield accurate-enough predictions in a wide variety of markets. This is why you will often find economists using the model to analyze the markets for crude oil, consumer electronic goods, fast-food meals, medical care, and higher education, even though in each of these cases one or more of the requirements may not be strictly satisfied.

THE PERFECTLY COMPETITIVE FIRM

When we stand at a distance and look at conditions in a competitive market, we get one view of what is occurring; when we stand close and look at the individual competitive *firm*, we get an entirely different picture. But these two pictures are very closely related. After all, a market is a collection of individual decision makers, much as a human body is a collection of individual cells. In a perfectly competitive market, the individual cells of firms and consumers and the overall body of the market affect each other through a variety of feedback mechanisms. This is why, in learning about the competitive firm, we must also discuss the competitive market in which it operates.

Figure 1(a) applies the tools you have already learned—supply and demand—to the competitive market for gold. The market demand curve slopes downward: As price falls, buyers will want to purchase more. The supply curve slopes upward: As price rises, the total quantity supplied by firms in the market will rise. The intersection of the supply and demand curves determines the equilibrium price of gold, which, in the figure, is \$400 per troy ounce.¹ This is all familiar territory. But now let's switch lenses and see how Small Time Gold Mines—an individual mining company—views this market.

Identify Goals and Constraints



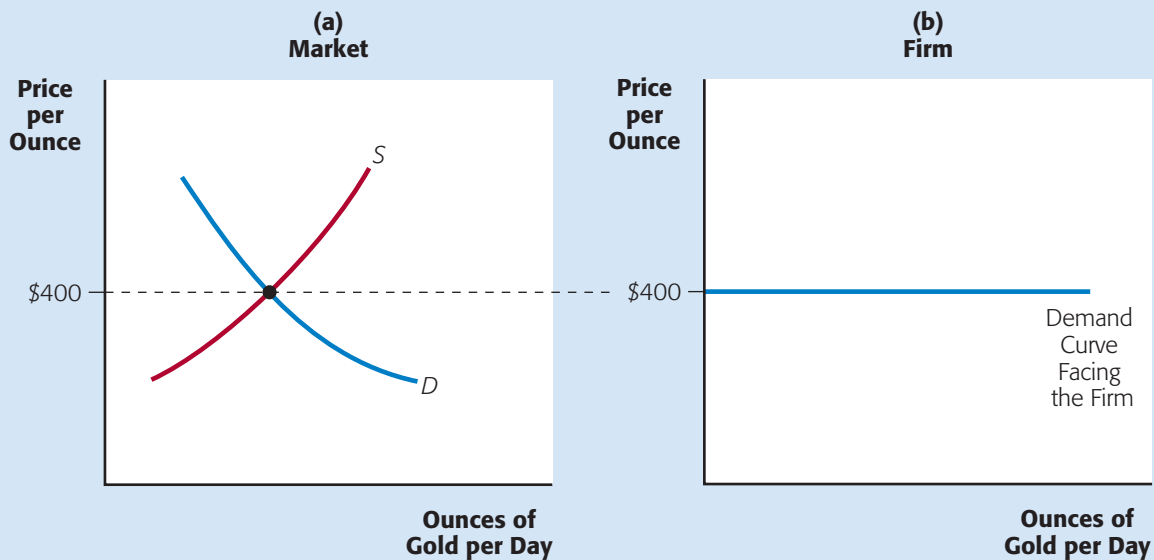
GOALS AND CONSTRAINTS OF THE COMPETITIVE FIRM

Small Time, like any business firm, strives to maximize profit. And, like any firm, it faces constraints. For example, it must use some given production technology to produce its output, and must pay some given prices for its inputs. As a result, Small Time firm faces a familiar cost constraint, just like Spotless Car Wash in Chapter 6 and Ned's Beds in Chapter 7, Small Time Gold Mines faces a total cost of production for any level of output it might want to produce. In addition to *total* cost, Small

¹ Gold is sold by the troy ounce, which is about 10 percent heavier than a regular ounce.

THE COMPETITIVE INDUSTRY AND FIRM

FIGURE 1



In panel (a) the market supply and demand curves intersect to determine a market price of \$400 per ounce. The typical firm in panel (b) can sell all it wants at that price. The demand curve facing the competitive firm is a horizontal line at the market price.

Time has *ATC*, *AVC*, and *MC* curves, and these have the familiar shapes you learned about in the previous two chapters.

A perfectly competitive firm faces a cost constraint like any other firm. The cost of producing any given level of output depends on the firm's production technology and the prices it must pay for its inputs.

In addition to a cost constraint, Small Time Gold Mines faces a demand constraint, as does any firm. But there is something different about the demand constraint for a perfectly competitive firm like Small Time.

The Demand Curve Facing a Perfectly Competitive Firm. Panel (b) of Figure 1 shows the demand curve facing Small Time Gold Mines. Notice the special shape of this curve: It is horizontal, or infinitely price elastic. This tells us that no matter how much gold Small Time produces, it will always sell it at the same price—\$400 per troy ounce. Why should this be?

First, in perfect competition, output is standardized—buyers do not distinguish the gold of one mine from that of another. If Small Time were to charge a price even a tiny bit higher than other producers, it would lose all of its customers—they would simply buy from Small Time's competitors, whose prices would be lower. The horizontal demand curve captures this effect. It tells us that if Small Time raises its price above \$400, it will not just sell *less* output, it will sell *no* output.

Second, Small Time is only a tiny producer relative to the entire gold market. No matter how much it decides to produce, it cannot make a noticeable difference in market quantity supplied and so cannot affect the market price. Once again, the

horizontal demand curve describes this effect perfectly: The firm can increase its production without having to lower its price.

All of this means that Small Time has no control over the price of its output—it simply accepts the market price as given:

Price taker Any firm that treats the price of its product as given and beyond its control.

In perfect competition, the firm is a price taker—it treats the price of its output as given.

The horizontal demand curve facing the firm and the resulting price-taking behavior of firms are hallmarks of perfect competition. If a manager thinks, “If we produce more output, we will have to lower our price,” then the firm faces a *downward-sloping* demand curve and is not a competitive firm. The manager of a competitive firm will always think, “We can sell all the output we want at the going price, so how much should we produce?”

Notice that, since a competitive firm takes the market price as given, its only decision is *how much output to produce and sell*. Once it makes that decision, we can determine the firm’s cost of production, as well as the total revenue it will earn (the market price times the quantity of output produced). Let’s see how this works in practice with Small Time Gold Mines.

COST AND REVENUE DATA FOR A COMPETITIVE FIRM

Table 1 shows cost and revenue data for Small Time. In the first two columns are different quantities of gold that Small Time could produce each day and the maximum

TABLE 1

COST AND REVENUE DATA FOR SMALL TIME GOLD MINES

(1) Output (Troy Ounces of Gold per Day)	(2) Price (per Troy Ounce)	(3) Total Revenue	(4) Marginal Revenue	(5) Total Cost	(6) Marginal Cost	(7) Profit
0	\$400	\$ 0		\$ 550		–\$550
1	\$400	\$ 400	\$400	\$1,000	\$450	–\$600
2	\$400	\$ 800	\$400	\$1,200	\$ 50	–\$400
3	\$400	\$1,200	\$400	\$1,250	\$100	–\$ 50
4	\$400	\$1,600	\$400	\$1,350	\$150	\$250
5	\$400	\$2,000	\$400	\$1,500	\$250	\$500
6	\$400	\$2,400	\$400	\$1,750	\$350	\$650
7	\$400	\$2,800	\$400	\$2,100	\$450	\$700
8	\$400	\$3,200	\$400	\$2,550	\$550	\$650
9	\$400	\$3,600	\$400	\$3,100	\$650	\$500
10	\$400	\$4,000	\$400	\$3,750		\$250

price that it could charge. Because Small Time is a competitive firm—a price taker—the price remains constant at \$400 per ounce, no matter *how* much gold it produces.

Run your finger down the total revenue and marginal revenue columns. Since price is always \$400, each time the firm produces another ounce of gold, total revenue rises by \$400. This is why marginal revenue—the additional revenue from selling one more ounce of gold—remains constant at \$400.

Figure 2 plots Small Time’s total revenue and marginal revenue. Notice that the total revenue (TR) curve in panel (a) is a *straight line* that slopes upward—each time output increases by one unit, TR rises by the same \$400. That is, the slope of the TR curve is equal to the price of output.

PROFIT MAXIMIZATION IN PERFECT COMPETITION

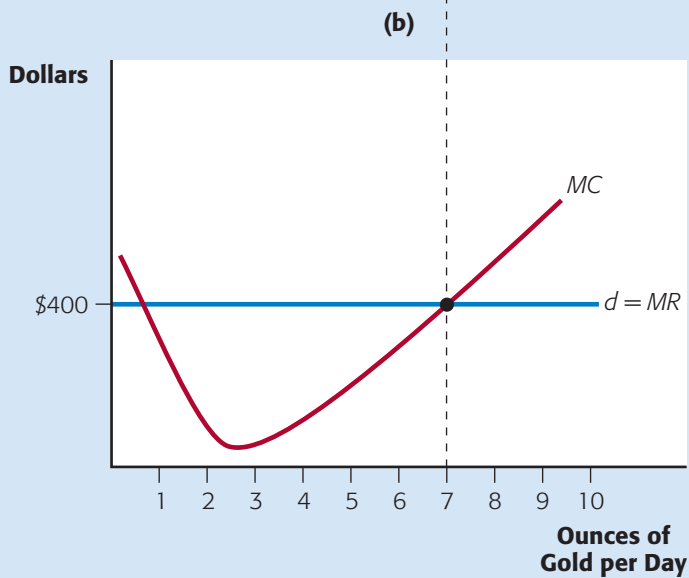
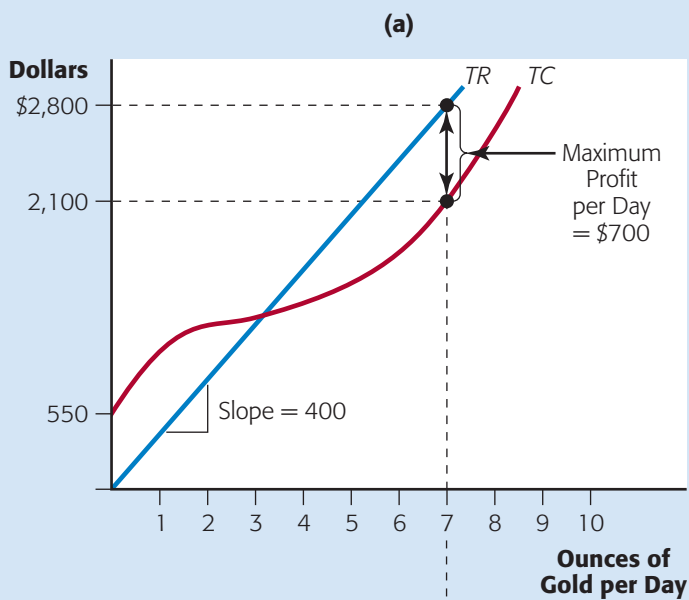


FIGURE 2

Panel (a) shows a competitive firm’s total revenue (TR) and total cost (TC) curves. TR is a straight line with slope equal to the market price. Profit is maximized at 7 ounces per day, where the vertical distance between TR and TC is greatest. Panel (b) shows that profit is maximized where the marginal cost (MC) curve intersects the horizontal demand (d) and marginal revenue (MR) curves.

The marginal revenue (MR) curve is a *horizontal* line at the market price. In fact, the MR curve is the same horizontal line as the demand curve. Why? Remember that marginal revenue is the additional revenue the firm earns from selling an additional unit of output. For a price-taking competitive firm, that additional revenue will always be the price per unit—no matter how many units it is already selling.

For a competitive firm, marginal revenue at each quantity is the same as the market price. For this reason, the marginal revenue curve and the demand curve facing the firm are the same—a horizontal line at the market price.

In panel (b), we have labeled the horizontal line “ $d = MR$,” since this line is both the firm’s demand curve (d) and its marginal revenue curve (MR).²

Columns 5 and 6 of Table 1 show total cost and marginal cost for Small Time. There is nothing special about cost data for a competitive firm. In Figure 2, you can see that marginal cost (MC)—as usual—first falls and then rises. Total cost, therefore, rises first at a decreasing rate and then at an increasing rate. (You may want to look at Chapter 6 to review why this cost behavior is so common.)

FINDING THE PROFIT-MAXIMIZING OUTPUT LEVEL

A competitive firm—like any other firm—wants to earn the highest possible profit, and to do so, it should use the principles you learned in Chapter 7. Although the diagrams look a bit different for competitive firms, the ideas behind them are the same. We can use either Table 1 or Figure 2 to find the profit-maximizing output level. And we can use the techniques you have already learned: the total-revenue and total-cost approach, or the marginal-revenue and marginal-cost approach.

The Total Revenue and Total Cost Approach. In the TR and TC approach, profit at each output level—entered in the last column of Table 1—is equal to $TR - TC$. Scan the profit entries until you find the highest value—\$700 per day. The output level at which the firm earns this profit—7 ounces per day—is the profit-maximizing output level. Alternatively, use the graph in panel (a) of Figure 2. Profit is the distance between the TR and TC curves, and this distance is greatest when the firm is producing 7 units of output, verifying what we found in the table.

The Marginal Revenue and Marginal Cost Approach. In the MR and MC approach, the firm should continue to increase output as long as marginal revenue is greater than marginal cost. You can verify, using Table 1, that if the firm is initially producing 1, 2, 3, 4, 5, or 6 units, $MR > MC$, so producing more will raise profit. Once the firm is producing 7 units, however, $MR < MC$, so further increases in output will reduce profit. Alternatively, using the graph in panel (b) of Figure 2, we look for the output level at which $MR = MC$. As the graph shows, there are two output levels at which the MR and MC curves intersect. However, we can rule out the first crossing point because there, the MC curve crosses the MR curve from *above*. Remember that the profit-maximizing output is found where the MC curve crosses the MR curve from *below*. Once again, this occurs at 7 units of output.

² In this and later chapters, lower-case letters for quantities and demand curves refer to the individual firm, and upper-case letters to the entire market. For example, the demand curve facing the firm is labeled d , while the market demand curve is labeled D .

You can see that finding the profit-maximizing output level for a competitive firm requires no new concepts or techniques; you have already learned everything you need to know in Chapter 7. In fact, the only difference is one of appearance. Ned's Beds—our firm in Chapter 7—did *not* operate under perfect competition. As a result, both its demand curve and its marginal revenue curve sloped *downward*. Small Time, however, operates under perfect competition, so its demand and *MR* curves are the same horizontal line.

MEASURING TOTAL PROFIT

You have already seen one way to measure a firm's total profit on a graph: the vertical distance between the *TR* and *TC* curves. In this section, you will learn another graphical way to measure profit.

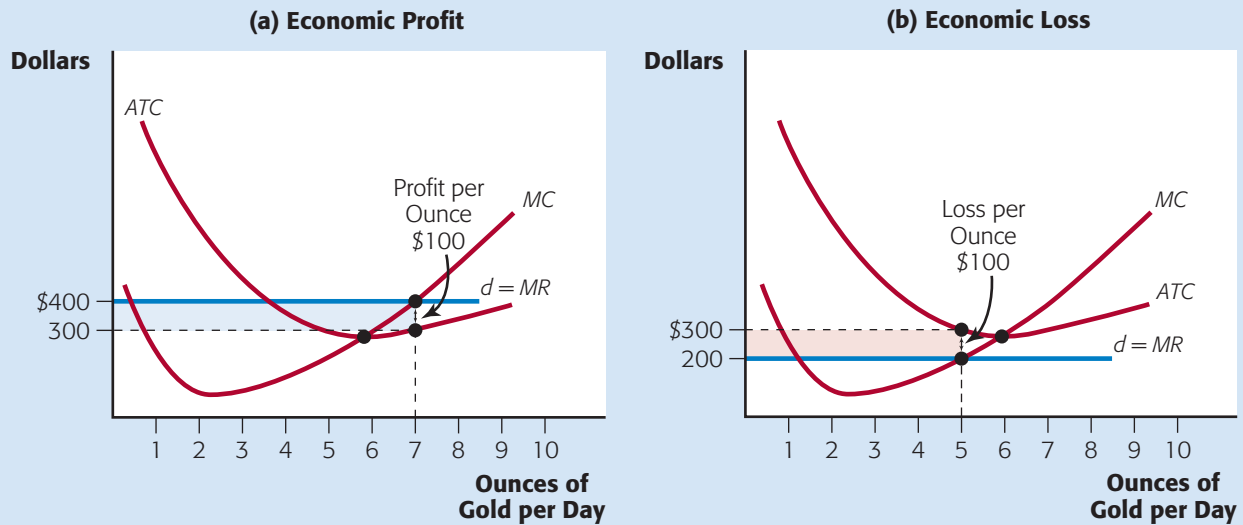
To do this, we start with the firm's *profit per unit*, which is the revenue it gets on each unit minus the cost per unit. Revenue per unit is just the price (*P*) of the firm's output, and cost per unit is our familiar *ATC*, so we can write:

$$\text{Profit per unit} = P - ATC.$$

In Figure 3(a), Small Time's *ATC* curve has been plotted from the data in Table 1 (p. 220). When the firm is producing at the profit-maximizing output level, 7 units, its *ATC* is \$300. Since the price of output is \$400, *profit per unit* = $P - ATC = \$400 - \$300 = \$100$. This is just the vertical distance between the firm's demand curve and its *ATC* curve at the profit-maximizing output level.

MEASURING PROFIT OR LOSS

FIGURE 3



The competitive firm in panel (a) produces where marginal cost equals marginal revenue, or 7 units of output per day. Profit per unit at that output level is equal to revenue per unit (\$400) minus cost per unit (\$300), or \$100 per unit. Total profit (indicated by the blue-shaded rectangle) is equal to profit per unit times the number of units sold, $\$100 \times 7 = \700 . In panel (b), the firm faces a lower market price of \$200 per ounce. The best it can do is to produce 5 ounces per day and suffer a loss shown by the red area. It loses \$100 per ounce on each of those 5 ounces produced, so the total loss is \$500—the area of the red-shaded rectangle.

Once we know Small Time's profit per unit, it is easy to calculate its *total* profit: Just multiply profit per unit by the number of units sold. Small Time is earning \$100 profit on each ounce of gold, and it sells 7 ounces in all, so total profit is $\$100 \times 7 = \700 .

Now look at the blue-shaded rectangle in Figure 3(a). The height of this rectangle is profit per unit, and the width is the number of units produced. The *area* of the rectangle—height \times width—equals Small Time's profit:

A firm earns a profit whenever $P > ATC$. Its total profit at the best output level equals the area of a rectangle with height equal to the distance between P and ATC , and width equal to the level of output.

In the figure, Small Time is fortunate: At a price of \$400, there are several output levels at which it can earn a profit. Its problem is to select the one that makes its profit as large as possible. (We should all have such problems.)

But what if the price had been lower than \$400—so low, in fact, that Small Time could not make a profit at *any* output level? Then the best it can do is to choose the smallest possible loss. Just as we did in the case of profit, we can measure the firm's total loss using the *ATC* curve.

Panel (b) of Figure 3 reproduces Small Time's *ATC* and *MC* curves from panel (a). This time, however, we have assumed a lower price for gold—\$200—so the firm's $d = MR$ curve is the horizontal line at \$200. Since this line lies everywhere below the *ATC* curve, profit per unit ($P - ATC$) is always negative: Small Time cannot make a positive profit at *any* output level.

With a price of \$200, the *MC* curve crosses the *MR* curve from below at 5 units of output. Thus, unless Small Time decides to shut down (we'll discuss shutting down later), it should produce 5 units. At that level of output, *ATC* is \$300, and profit per unit is $P - ATC = \$200 - \$300 = -\$100$, a *loss* of \$100 per unit. The total loss is loss per unit times the number of units produced, or $-\$100 \times 5 = -\500 . This is equal to the area of the red-shaded rectangle in Figure 3(b), with height equal to \$100 and width equal to 5 units:

A firm suffers a loss whenever $P < ATC$ at the best level of output. Its total loss equals the area of a rectangle with height equal to the distance between P and ATC , and width equal to the level of output.

THE FIRM'S SHORT-RUN SUPPLY CURVE

A competitive firm is a price taker: It takes the market price as given and then decides how much output it will produce at that price. If the market price changes for any reason, the price taken as given will change as well. The firm will then have to find a new profit-maximizing output level. Let's see how the firm's choice of output changes as the market price rises or falls.

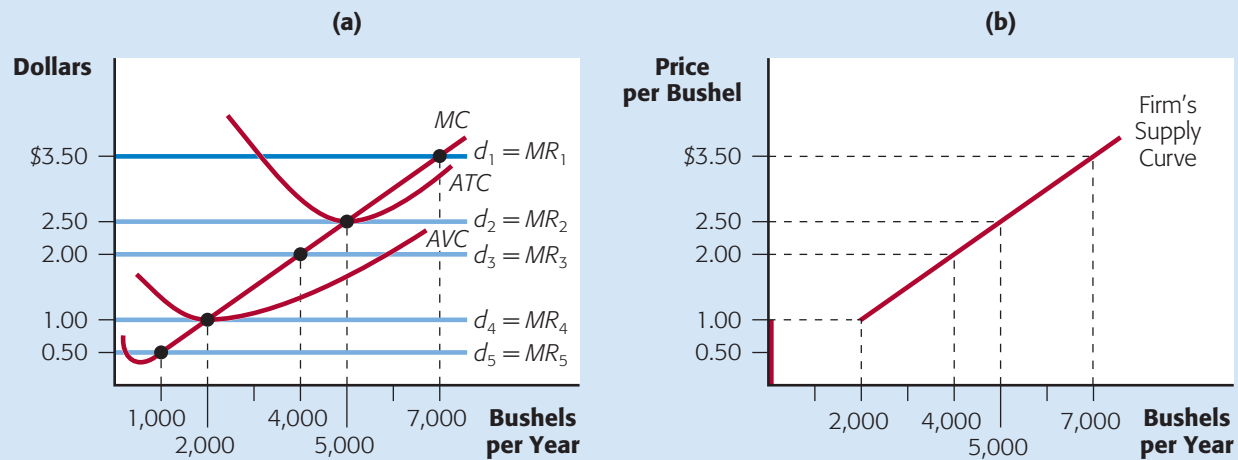
Figure 4(a) shows *ATC*, *AVC*, and *MC* curves for a competitive producer of wheat. The figure also shows five hypothetical demand curves the firm might face, each corresponding to a different market price for wheat. If



It is tempting—but *wrong*—to think that the firm should produce where profit *per unit* ($P - ATC$) is greatest. The firm's goal is to maximize *total* profit, not profit per unit. Using Table 1 or Figure 3(a), you can verify that while Small Time's profit *per unit* is greatest at 6 units of output, its *total* profit is greatest at 7 units.

SHORT-RUN SUPPLY UNDER PERFECT COMPETITION

FIGURE 4



Panel (a) shows a typical competitive firm facing various market prices. For prices between \$1 and \$3.50 per bushel, the profit-maximizing quantity is found by sliding along the MC curve. Below \$1 per bushel, the firm is better off shutting down, because $P < AVC$, and so $TR < TVC$. Panel (b) shows that the firm's supply curve consists of two segments. Above the shut-down price of \$1 per bushel it follows the MC curve; below that price, it is coincident with the vertical axis.

the market price were \$3.50 per bushel, the firm would face demand curve d_1 , and its profit-maximizing output level—where $MC = MR$ —would be 7,000 bushels per year. If the price dropped to \$2.50 per bushel, the firm would face demand curve d_2 , and its profit-maximizing output level would drop to 5,000 bushels. You can see that the profit-maximizing output level is always found by traveling from the price, across to the firm's MC curve, and then down to the horizontal axis. In other words,

as the price of output changes, the firm will slide along its MC curve in deciding how much to produce.

But there is one problem with this: If the firm is suffering a loss—a loss large enough to justify shutting down—then it will *not* produce along its MC curve; it will produce zero units instead. Thus, in order to know for certain how much output the firm will produce, we must bring in the shutdown rule you learned in Chapter 7.

Suppose the price in Figure 4(a) drops down to \$2 per bushel. At this price, the best output level is 4,000 bushels, and the firm suffers a loss, since $P < ATC$. Should the firm shut down? Let's see. At 4,000 bushels, it is also true that $P > AVC$, since the demand curve lies above the AVC curve at this output level. Multiplying both sides of the last inequality by Q gives us

$$P \times Q > AVC \times Q.$$

Since $AVC \times Q$ is just TVC , this inequality is the same as

$$TR > TVC.$$

As we know from Chapter 7, a firm should *never* shut down when $TR > TVC$. Thus, at a price of \$4, the firm will stay open and produce 4,000 units of output.

Now, suppose the price drops all the way down to \$0.50 per bushel. At this price, $MR = MC$ at 1,000 bushels, but notice that here $P < AVC$. Once again, we multiply both sides by Q to obtain

$$P \times Q < AVC \times Q$$

or

$$TR < TVC.$$

A firm should *always* shut down when $TR < TVC$, so at a price of \$0.50, this firm will produce *zero* units of output.

Finally, let's consider a price of \$1. At this price, $MR = MC$ at 2,000 bushels, and here we have $P = AVC$ or $TR = TVC$. At \$1, therefore, the firm will be indifferent between staying open and shutting down. We call this price the firm's **shutdown price**, since it will shut down at any price lower and stay open at any price higher. The output level at which the firm will shut down must occur at the *minimum* of the AVC curve. Why? Note that as the price of output decreases, the best output level is found by sliding along the MC curve, until MC and AVC cross. At that point, the firm will shut down. But—as you learned in Chapter 6— MC will always cross AVC at its minimum point.

Now let's recapitulate what we've found about the firm's output decision. For all prices above the minimum point on the AVC curve, the firm will stay open and will produce the level of output at which $MR = MC$. For these prices, the firm slides along its MC curve in deciding how much output to produce. But for any price below the minimum AVC , the firm will shut down and produce zero units. We can summarize all of this information in a single curve—the **firm's supply curve**—which tells us how much output the firm will produce at any price:

The competitive firm's supply curve has two parts. For all prices above the minimum point on its AVC curve, the supply curve coincides with the MC curve. For all prices below the minimum point on the AVC curve, the firm will shut down, so its supply curve is a vertical line segment at zero units of output.

In panel (b) of Figure 4, we have drawn the supply curve for our hypothetical wheat farmer. As price declines from \$3.50 to \$1, output is determined by the firm's MC curve. For all prices *below* \$1—the shutdown price—output is zero and the supply curve coincides with the vertical axis.

COMPETITIVE MARKETS IN THE SHORT RUN

Recall that the short run is a time period too short for the firm to vary *all* of its inputs: The quantity of at least one input remains fixed. For example, in the short run, a wheat farmer will be stuck with a certain plot of land and a certain number of tractors. Now let's extend the concept of the short run from the firm to the market as a whole. It makes sense that if the short run is insufficient time for a firm to vary its fixed inputs, then it is also insufficient time for a *new* firm to acquire those fixed inputs and *enter* the market. Similarly, it is too short a period for firms to reduce their fixed inputs to zero and *exit* the market. We conclude that

Shutdown price The price at which a firm is indifferent between producing and shutting down.

Firm's supply curve A curve that shows the quantity of output a competitive firm will produce at different prices.

in the short run, the number of firms in the industry is fixed.

THE (SHORT-RUN) MARKET SUPPLY CURVE

Once we know how to find the supply curve of each *individual* firm in a market, we can easily determine the short-run **market supply curve**—showing the amount of output that all sellers in the market will offer at each price.

To obtain the market supply curve, we add up the quantities of output supplied by all firms in the market at each price.

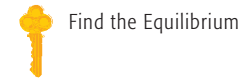
Market supply curve A curve indicating the quantity of output that all sellers in a market will produce at different prices.

To keep things simple, suppose there are 100 identical wheat farms and that each one has the supply curve shown in Figure 5(a)—the same supply curve we derived in Figure 4. Then at a price of \$3.50, each firm would produce 7,000 bushels. With 100 such firms, the market quantity supplied will be $7,000 \times 100 = 700,000$ bushels. At a price of \$2.50, each firm would supply 5,000 bushels, so market supply would be 500,000. Continuing in this way, we can trace out the market supply curve shown in panel (b) of Figure 4. Notice that once the price drops below \$1—the shutdown price for each firm—the market supply curve jumps to zero.

The market supply curve in the figure is a *short-run* market supply curve, since it gives us the combined output level of just those firms *already* in the industry. As we move along this curve, we are assuming that two things are constant: (1) the fixed inputs of each firm and (2) the number of firms in the market.

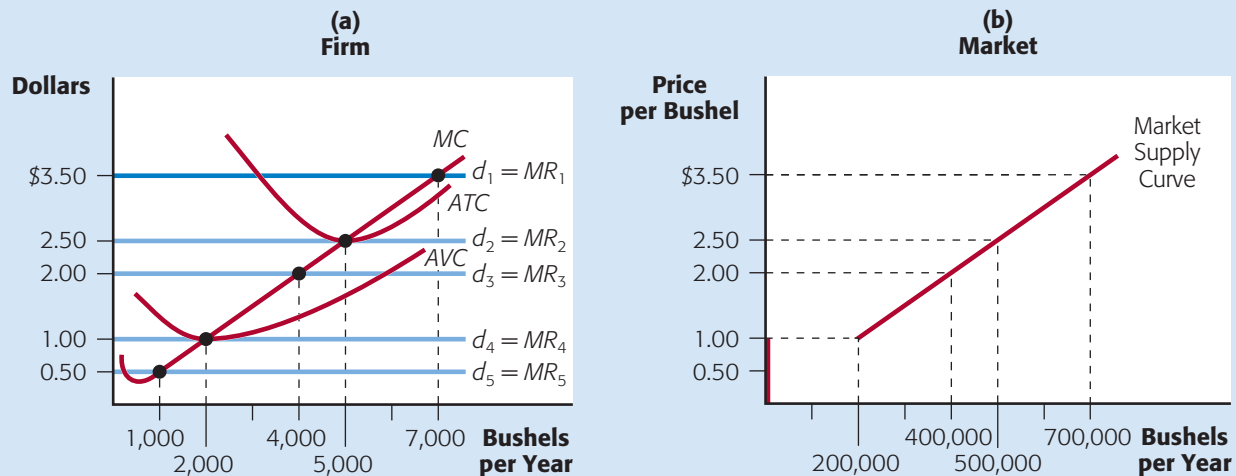
SHORT-RUN EQUILIBRIUM

How does a perfectly competitive market achieve equilibrium? We’ve already addressed this question in Chapter 3, in our study of supply and demand. But now



DERIVING THE MARKET SUPPLY CURVE

FIGURE 5



The market supply curve of panel (b) is obtained by adding up the quantities of output supplied by all firms in the market at each price, as shown in panel (a).

we'll take a much closer look, paying attention to the individual firm and individual consumer as well as the market.

Figure 6 puts together the pieces we've discussed so far, including those from Chapter 5 on consumer choice, to paint a complete picture of how a competitive market arrives at a short-run equilibrium. On the right side, we add up the quantities supplied by all firms to obtain the market supply curve. On the left side, we add up the quantities demanded by all consumers to obtain the market demand curve. The market supply and demand curves show *if/then* relationships: *If* the price were such and such, *then* firms would supply this much and consumers would buy that much. Up to this point, the prices and quantities are purely hypothetical. But once we bring the two curves together and find their intersection point, we know the *equilibrium* price—the price at which trading will actually take place. Finally, we confront each firm and each consumer with the equilibrium price to find the actual quantity each consumer will buy and the actual quantity each firm will produce.

FIGURE 6

PERFECT COMPETITION

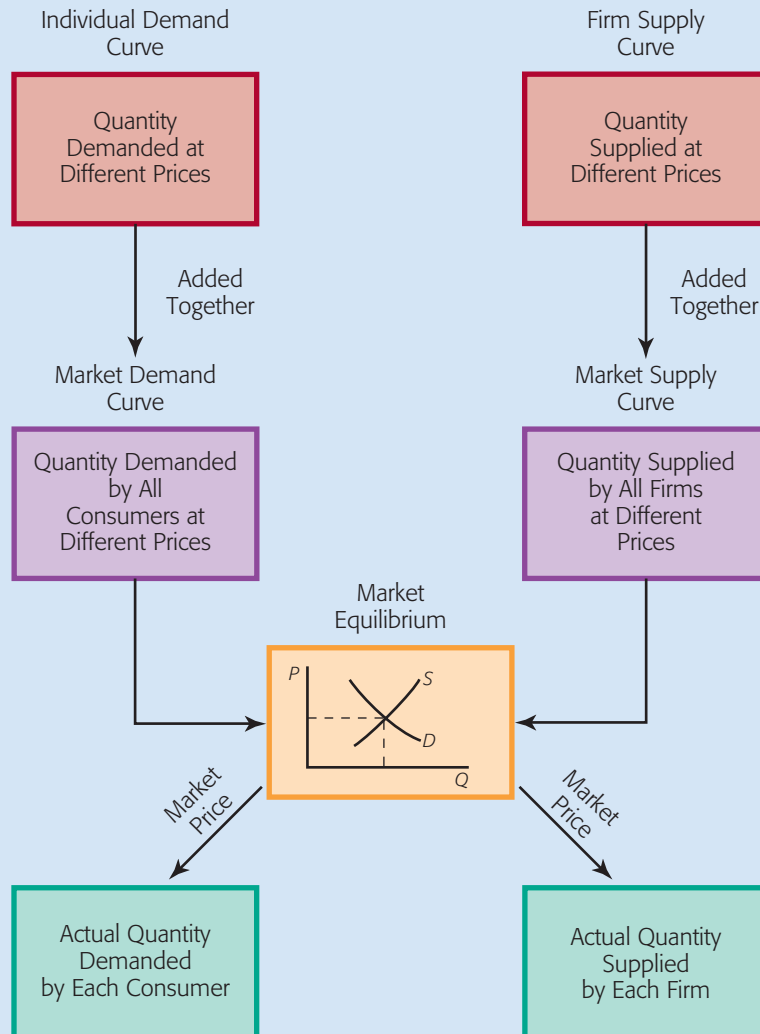


Figure 7 gets more specific, illustrating two possible short-run equilibria in the wheat market. In panel (a), if the market demand curve were D_1 , the short-run equilibrium price would be \$3.50. Each firm would face the horizontal demand curve d_1 (panel (b)) and decide to produce 7,000 bushels. With 100 such firms, the equilibrium market quantity would be 700,000 bushels. Notice that, at a price of \$3.50, each firm is enjoying an economic profit, since $P > ATC$.

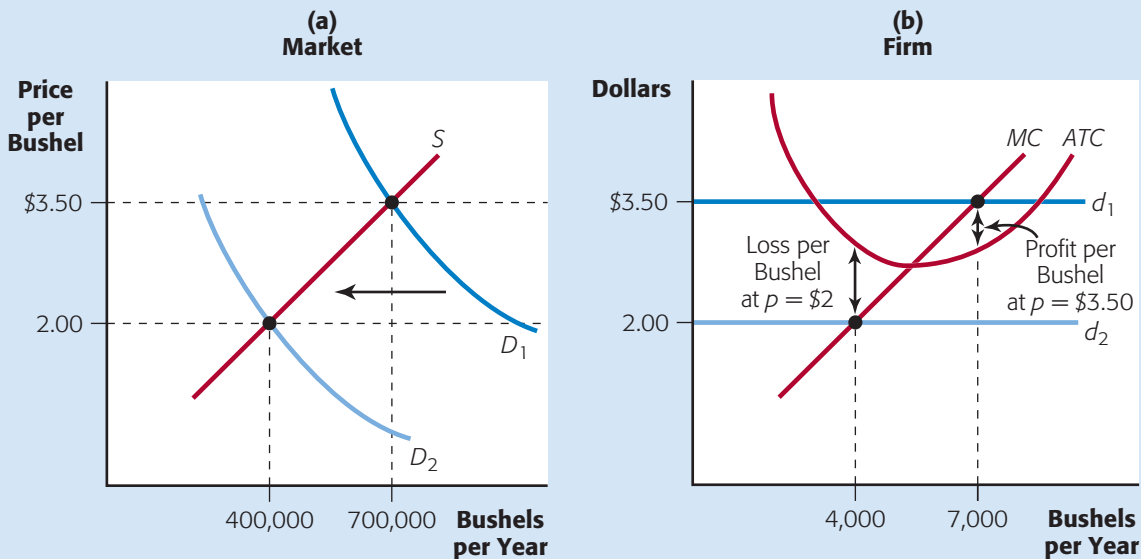
If the market demand curve were D_2 instead, the equilibrium price would be \$2. Each firm would face demand curve d_2 , and produce 4,000 bushels. With 100 firms, the equilibrium market quantity would be 400,000. Here, each firm is suffering an economic loss, since $P < ATC$. These two examples show us that *in short-run equilibrium, competitive firms can earn an economic profit or suffer an economic loss.*

We are about to leave the short run and turn our attention to what happens in a competitive market over the long run. But before we do, let's look once more at how a short-run equilibrium is established. One part of this process—combining supply and demand curves to find the market equilibrium—has been familiar to you all along. But now you can better appreciate how much information is contained within each of these curves and what an impressive job the market does coordinating millions of decisions made by people who may never even meet each other.

Think about it: So many individual consumers and firms, each with its own agenda, trading in the market. Not one of them has any power to decide or even influence the market price. Rather, the price is determined by *all* of them, adjusting until *total* quantity supplied is equal to *total* quantity demanded. Then, facing this equilibrium price, each consumer buys the quantity he or she wants, each firm produces the output level that it wants, and we can be confident that all of them will

SHORT-RUN EQUILIBRIUM IN PERFECT COMPETITION

FIGURE 7



In panel (a) demand curve D_1 intersects supply curve S to determine a market price of \$3.50 per bushel. The firm in panel (b) takes that price as given, produces 7,000 bushels per year—determined at the intersection of its marginal cost curve with the horizontal demand curve, d_1 —and earns a short-run profit. If the market demand curve shifts left to D_2 , the market price falls to \$2 per bushel. The typical firm then reduces production to 4,000 bushels per year and suffers a short-run loss.

be able to realize their plans. Each buyer can find willing sellers, and each seller can find willing buyers.

In perfect competition, the market sums up the buying and selling preferences of individual consumers and producers, and determines the market price. Each buyer and seller then takes the market price as given, and each is able to buy or sell the desired quantity.

This process is, from a certain perspective, a thing of beauty, and it happens each day in markets all across the world—markets for wheat, corn, barley, soybeans, apples, oranges, gold, silver, copper, and more. And something quite similar happens in other markets that do not strictly satisfy our requirements for perfect competition—markets for television sets, books, air conditioners, fast-food meals, oil, natural gas, bottled water, blue jeans. The list is virtually endless.

COMPETITIVE MARKETS IN THE LONG RUN

So far, we've explored the short run only, and assumed that the number of firms in the market is fixed. But perfect competition becomes even *more* interesting in the long run, when entry and exit can occur. After all, the long run is a time horizon long enough for firms to vary *all* of their inputs. It should therefore be long enough for *new* firms to acquire fixed inputs and enter the market, and for firms already in the industry to sell off their fixed inputs and *exit* from the market.

But what makes firms want to enter or exit a market? The driving force behind entry is economic profit, and the force behind exit is economic loss.

PROFIT AND LOSS AND THE LONG RUN

Recall that economic profit is the amount by which total revenue exceeds *all* costs of doing business. The costs to be deducted include implicit costs like foregone investment income and foregone wages for an owner who devotes money and time to the business. Thus, when a firm earns positive economic profit, we know the owners are earning *more* than they could by devoting their money and time to some other activity.

A temporary episode of positive economic profit will not have much impact on a competitive industry, other than the temporary pleasure it gives the owners of competitive firms. But when positive profit reflects basic conditions in the industry and is expected to continue, major changes are in the works. Outsiders, hungry for profit themselves, will want to enter the market, and—since *there are no barriers to entry*—they can do so.

Similarly, if firms in the market are suffering economic losses, they are not earning enough revenue to cover all their costs, so there must be other opportunities that would more adequately compensate owners for their money or time. If this situation is expected to continue over the firm's long-run planning horizon—a period long enough to vary *all* inputs—there is only one thing for the firm to do: exit the market by selling off its plant and equipment, thereby reducing its loss to zero.

In a competitive market, economic profit and loss are the forces driving long-run change. The expectation of continued economic profit causes outsiders to enter the market; the expectation of continued economic losses causes firms in the market to exit.



Rocky Mountain Internet Service is one of more than 7,000 new Internet Service Providers that have entered the market due to high profits of existing firms.

In the real world of business, entry and exit occur literally every day. In some cases, we see entry occur through the formation of an entirely new firm. For example, in the late 1990s, the high profits of the earliest Internet service providers (ISPs)—such as America Online, CompuServe, and Prodigy—led to the establishment of more than 7,000 new ISPs by the end of the decade. Entry can also occur when an existing firm adds a new product to its line. For example, among the firms that entered the ISP market were many firms that had been established years before there *was* such a thing as an ISP, such as Sprint (which entered with Earthlink), Microsoft (the Microsoft Network), and AT&T. Although these were not *new firms*, they were *new participants* in the market for Internet service.

Exit, too, can occur in different ways. A firm may go out of business entirely, selling off its assets and freeing itself once and for all from all costs. Every year, thousands of small businesses exit markets in this way. You may know of a local video store, grocery store, or furniture shop that decided to permanently shut its doors. Restaurants, in particular, seem especially prone to long-run economic loss. It has been reported that half of all new restaurants exit the market within two years of being established.

But exit can also occur when a firm switches out of a particular product line, even as it continues to produce other things. For example, publishing companies often decide to abandon unsuccessful magazines, yet they continue to thrive by publishing other magazines and books.

LONG-RUN EQUILIBRIUM

Entry and exit—however they occur—are powerful forces in real-world competitive markets. They determine how these markets change over the long run, how much output will ultimately be available to consumers, and the prices they must pay. To explore these issues, let's see how entry and exit move a market to its long-run equilibrium from different starting points.

From Short-Run Profit to Long-Run Equilibrium. Suppose that the market for wheat is initially in a short-run equilibrium like point A in panel (a) of Figure 8, with market supply curve S_1 . The initial equilibrium price is \$4.50 per bushel. In panel (b), we see that a typical competitive firm—producing 9,000 bushels—is earning economic profit, since $P > ATC$ at that output level. As long as we remain in the short run—with no new firms entering the market—this situation will not change.

But as we enter the long run, much will change. First, economic profit will attract new entrants, increasing the number of sellers in the market and *shifting the market supply curve rightward*. (Remember, the market supply curve S_1 is drawn for a fixed number of firms; with more firms in the market, a greater quantity will be supplied at each price.) As the market supply curve shifts rightward, several things happen:

1. The market price begins to fall—from \$4.50 to \$4.00 to \$3.50 and so on.
2. As market price falls, the demand curve facing each firm shifts downward.
3. Each firm—striving as always to maximize profit—will slide down its marginal cost curve, decreasing output.³

This process of adjustment—in the market and the firm—continues until . . . well, until when? To answer this question, remember why these adjustments are

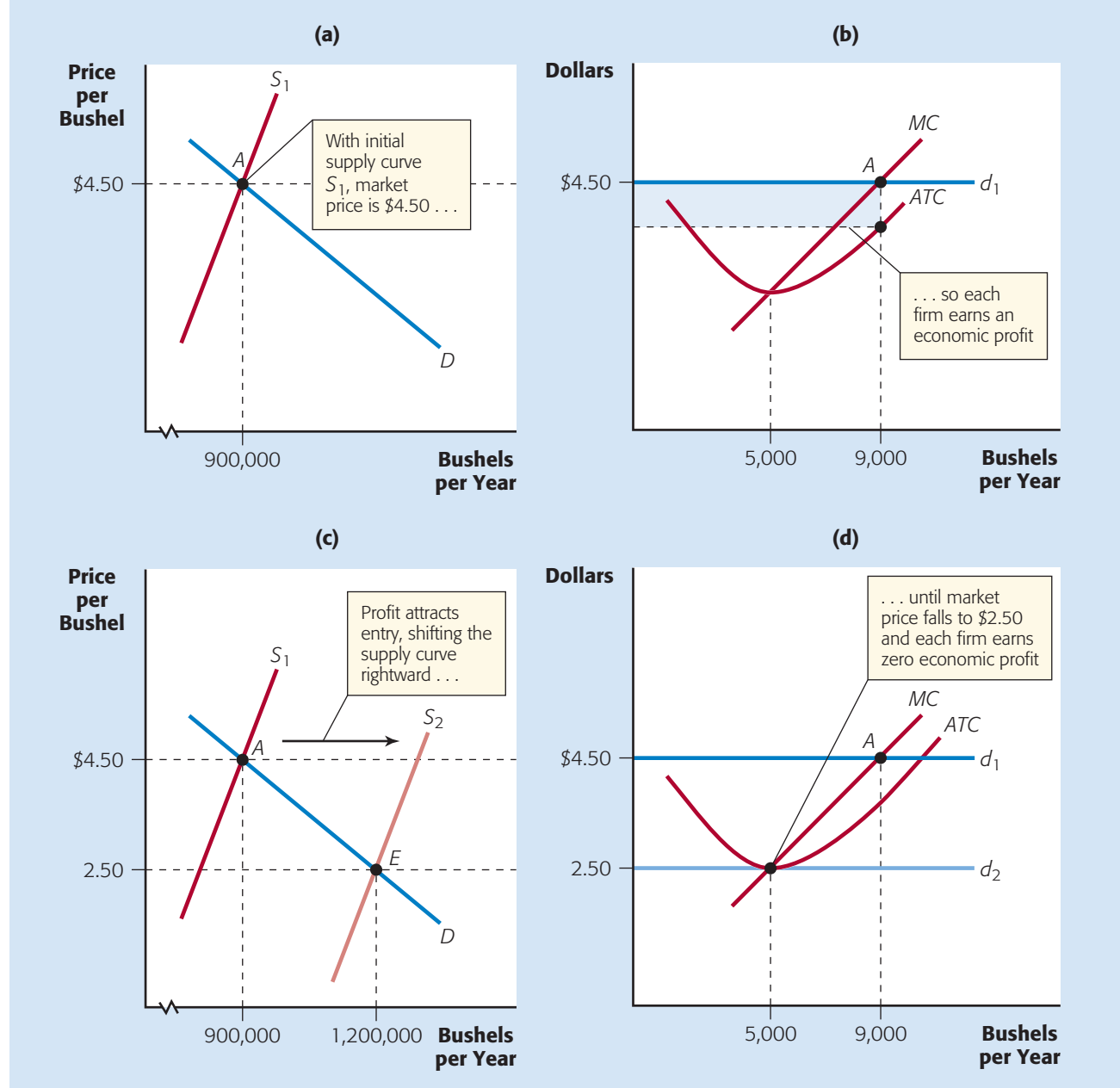


Find the Equilibrium

³ There is one other possible consequence that we ignore here: Entry into the industry—which changes the demand for the industry's inputs—may also change input prices. If this occurs, firms' ATC curves will shift. For now, we will assume that entry (and exit) do not affect input prices, so that the ATC curve does not shift.

FIGURE 8

FROM SHORT-RUN PROFIT TO LONG-RUN EQUILIBRIUM



occurring in the first place: Economic profit is attracting new entrants and shifting the market supply curve rightward. Thus, all of these changes will stop when the *reason* for entry—positive profit—no longer exists. And this, in turn, requires the market supply curve to shift rightward enough, and the price to fall enough, so that *each existing firm is earning zero economic profit*. Panels (c) and (d) in Figure 8 show the final, long-run equilibrium. First, look at panel (c), which shows long-run market equilibrium at point E . The market supply curve has shifted to S_2 , and the price has fallen to \$2.50 per bushel. Next, look at panel (d), which tells us why the market supply curve stops shifting when it reaches S_2 . With that supply curve, each

firm is producing at the lowest point of its *ATC* curve, with $P = ATC = \$2.50$, and each is earning zero economic profit. With no economic profit, there is no further reason for entry, and no further shift in the market supply curve.

In a competitive market, positive economic profit continues to attract new entrants until economic profit is reduced to zero.

Before proceeding further, take a close look at Figure 8. As the market moves to its long-run equilibrium (point *E* in panels (c) and (d)), output at each firm *decreases* from 9,000 to 5,000 bushels. But in the market as a whole, output *increases* from 900,000 to 1,200,000 bushels. How can this be? (See if you can answer this question yourself. *Hint*: entry!)

From Short-Run Loss to Long-Run Equilibrium. We have just seen how, beginning from a position of short-run profit at the typical firm, a competitive market will adjust until the profit is eliminated. But what if we begin from a position of loss? As you might guess, the same type of adjustments will occur, only in the opposite direction.

This is a good opportunity for you to test your own skill and understanding. Study Figure 8 carefully. Then see if you can draw a similar diagram that illustrates the adjustment from short-run *loss* to long-run equilibrium. Start with a market price of \$1. Use the same demand curve as in Figure 8, but draw in a new, appropriate market supply curve. Then let the market work. Show what happens in the market, and at each firm, as economic loss causes some firms to exit. If you do this correctly, you'll end up once again at a market price of \$2.50, with each firm earning zero economic profit. Your graph will illustrate the following conclusion:

In a competitive market, economic losses continue to cause exit until the losses are reduced to zero.

Distinguishing Short-Run from Long-Run Outcomes. You've seen that the equilibrium in a competitive market can be very different in the short run than in the long run. In short-run equilibrium, competitive firms can earn profits or suffer losses. But in long-run equilibrium, after entry or exit has occurred, economic profit is always zero. The distinction between short-run and long-run equilibrium is important, and not just in competitive markets. In *any* market, our analysis will depend on the time period we are considering, and the correct period depends on the question we are asking. If we want to predict what happens several years after a change in demand, we should ask what the new *long-run* equilibrium will be. If we want to know what happens a few *months* after a change in demand, we'll look for the new *short-run* equilibrium.

When economists look at a market, they automatically think of the short run versus the long run and then choose the period more appropriate for the question at hand. As you'll see, this way of thinking is applied again and again in economics.

THE NOTION OF ZERO PROFIT IN PERFECT COMPETITION

From the preceding description, you may wonder why anyone in his or her right mind would ever want to set up shop in a competitive industry or stay there for any length of time, since—in the long run—they can expect zero economic profit. Indeed, if you want to become a millionaire, you would be well advised not to buy a

wheat farm. But most wheat farmers—like most other sellers in competitive markets—do not curse their fate. On the contrary, they are likely to be quite content with the performance of their businesses. How can this be?

Remember that zero *economic* profit is not the same as zero *accounting* profit. When a firm is making zero *economic* profit, it is still making some accounting profit. In fact, the accounting profit is just enough to cover all of the owner's costs—including compensation for any foregone investment income or foregone salary. Suppose, for example, that a wheat farmer paid \$100,000 for land and works 40 hours per week. Suppose, too, that the money *could* have been invested in some other way and earned \$6,000 per year, and the farmer *could* have worked equally pleasantly elsewhere and earned \$40,000 per year. Then the farm's implicit costs will be \$46,000, and zero economic profit would mean that the farm was earning \$46,000 in *accounting profit* each year. This won't make a wheat farmer ecstatic, but it will make it worthwhile to keep working the farm. After all, if the farmer quits and takes up the next best alternative, he or she will do no better. To emphasize that zero economic profit is not an unpleasant outcome, economists often replace it with the term **normal profit**, which is a synonym for “zero economic profit,” or “just enough accounting profit to cover implicit costs.” Using this language, we can summarize long-run conditions at the typical firm this way:

Normal profit Another name for zero economic profit.

In the long run, every competitive firm will earn normal profit—that is, zero economic profit.

PERFECT COMPETITION AND PLANT SIZE

There is one more characteristic of competitive markets in the long run that we have not yet discussed: the plant size of the competitive firm. It turns out that the same forces—entry and exit—that cause all firms to earn zero economic profit *also* ensure that:

In long-run equilibrium, every competitive firm will select its plant size and output level so that it operates at the minimum point of its LRATC curve.

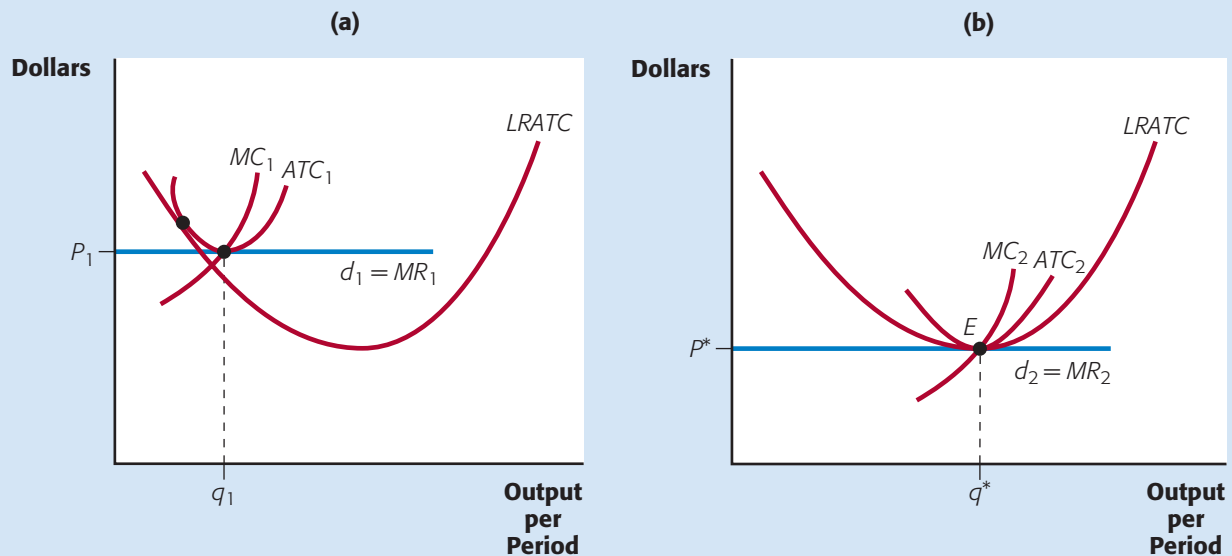
To see why, let's consider what would happen if this condition were violated. Figure 9(a) illustrates a firm in a perfectly competitive market. The firm faces a market price of P_1 and produces quantity q_1 , where $MC_1 = MR_1$. With its current plant, the firm has average costs given by ATC_1 . Note that the firm is earning zero profit, since average cost is equal to P_1 at the best output level.

But panel (a) does *not* show a true long-run equilibrium. How do we know this? First, in the long run, the typical firm will want to expand. Why? Because by increasing its plant size, it could slide down its *LRATC* curve and produce more output at a lower cost per unit. Since it is a perfectly competitive firm—a small participant in the market—it can expand in this way *without* affecting market price. As a result, the firm—after expanding—could operate on a new, lower *ATC* curve, so that *ATC* is less than *P*. That is, by expanding, the firm could earn an economic profit.

Second, this same opportunity to earn positive economic profit will attract new entrants that will establish larger plants from the outset. Expansion by existing firms and entry by new ones increases market output and bring down the market price. The process will stop—and a long-run equilibrium will be established—only when there is no potential to earn positive economic profit with *any* plant size. As

PERFECT COMPETITION AND PLANT SIZE

FIGURE 9



The firm in panel (a) faces a price of P_1 and produces quantity q_1 . It earns zero profit because price equals average cost. In the long run, this firm will want to expand. By sliding down $LRATC$, it could produce more output at a lower cost per unit and earn an economic profit. In turn, economic profit will attract entry, and that will reduce the market price. The firm's long-run equilibrium position is shown in panel (b). The firm earns zero profit by operating at minimum $LRATC$.

you can see in panel (b), this condition is satisfied only when each firm is operating at the minimum point on its $LRATC$ curve, using the plant represented by ATC_2 , and producing output of q^* . Entry and expansion must continue in this market until the price falls to P^* , because only then will each firm—doing the best that it can do—earn zero economic profit. (*Question:* In the long run, what would happen to the firm in panel (a) if it refused to increase its plant size?)

A SUMMARY OF THE COMPETITIVE FIRM IN THE LONG RUN

Panel (b) of Figure 9 summarizes everything you have learned about the competitive firm in long-run equilibrium. The typical firm—taking the market price P^* as given—produces the profit-maximizing output level q^* , where $MR = MC$. Since this is the long run, each firm will be earning zero economic profit, so we also know that $P^* = ATC$. But since $P^* = MC$ and $P^* = ATC$, it must also be true that $MC = ATC$. As you learned in Chapter 6, MC and ATC are equal only at the minimum point of the ATC curve. Thus, we know that each firm must be operating at the lowest possible point on the ATC curve for the plant it is operating. Finally, each firm selects the plant that makes its $LRATC$ as low as possible, so each operates at the minimum point on its $LRATC$ curve.

As you can see, there is a lot going on in Figure 9 (b). But we can put it all together with a very simple statement:

At each competitive firm in long-run equilibrium, $P = MC = \text{minimum } ATC = \text{minimum } LRATC$.

In Figure 9(b), this equality is satisfied when the typical firm produces at point E , where its demand, marginal cost, ATC , and $LRATC$ curves all intersect. This is a figure well worth remembering, since it summarizes so much information about competitive markets in a single picture. (Here is a useful self-test: Close the book, put away your notes, and draw a set of diagrams in which one curve at a time does *not* pass through the common intersection point of the other three. Then explain which principle of firm or market behavior is violated by your diagram. Do this separately for all four curves.)

Figure 9(b) also explains one of the important ways in which perfect competition benefits consumers: In the long run, each firm is driven to the plant size and output level at which its cost per unit is as low as possible. This lowest possible cost per unit is also the price per unit that consumers will pay. If price were any lower than P^* , it would not be worthwhile for firms to continue producing the good in the long run. Thus, given the ATC curve faced by each firm in this industry—a curve that is determined by each firm’s production technology and the costs of its inputs— P^* is the lowest possible price that will ensure the continued availability of the good. In perfect competition, consumers are getting the best deal they could possibly get.



Economists have tried to simulate the behavior of competitive markets through experiments. Vanderbilt University’s Market. Econ is an Internet-based example (<http://market.econ.vanderbilt.edu>).

What Happens When
Things Change?



WHAT HAPPENS WHEN THINGS CHANGE?

So far, you’ve learned how competitive firms make decisions, how these decisions lead to a short-run equilibrium in the market, and how the market moves from short- to long-run equilibrium through entry and exit. Now, it’s time to turn to Key Step 4: *What happens when things change?* In this section, we’ll deal with a change in demand for the product and, in the process, learn some important additional features of perfect competition. In the section titled “Using the Theory,” we’ll look at changes in technology.

A CHANGE IN DEMAND

In Figure 10, panel (a) shows a competitive market that is initially in long-run equilibrium at point A , where the market demand curve D_1 and supply curve S_1 intersect. (Ignore the other curves for now). Panel (b) shows conditions at the firm, which faces demand curve d_1 and produces the profit-maximizing quantity q_1 .

But now suppose that the market demand curve shifts rightward to D_2 and remains there. (This shift could be caused by any one of several factors. If you can’t list some of them, turn back to Chapter 3 and look again at Figure 3.) Panels (c) and (d) show what happens. In the *short run*, the shift in demand moves the market equilibrium to point B , with market output Q_{SR} and price P_{SR} . At the same time, the demand curve facing each firm shifts upward, and each firm raises output to the new profit-maximizing level q_{SR} . At this output level, $P > ATC$, so each firm is earning economic profit. Thus, the short-run impact of an increase in demand is (1) a rise in market price, (2) a rise in market quantity, and (3) economic profits.

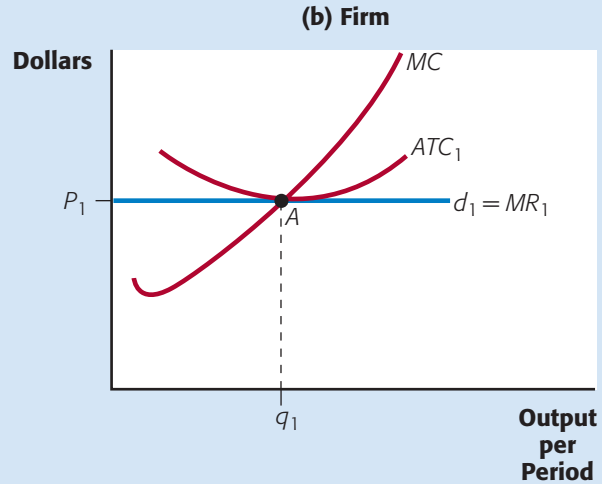
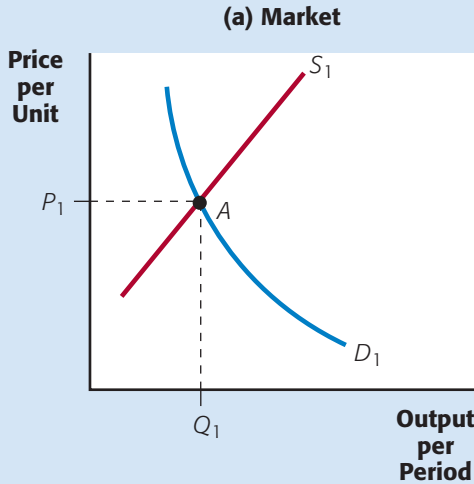
When we turn to the long run, we know that entry will occur (why?), so the market supply curve shifts rightward, bringing down the price until each firm earns zero economic profit. But how far must the price fall in order to bring this about? That is, how far can we expect the market supply curve to shift? In answering this question, we’ll add one more detail to our model that we’ve ignored until now.

Think about what happens as entry occurs in an industry. With more firms, output increases, so the industry will demand more *inputs*—more raw materials,

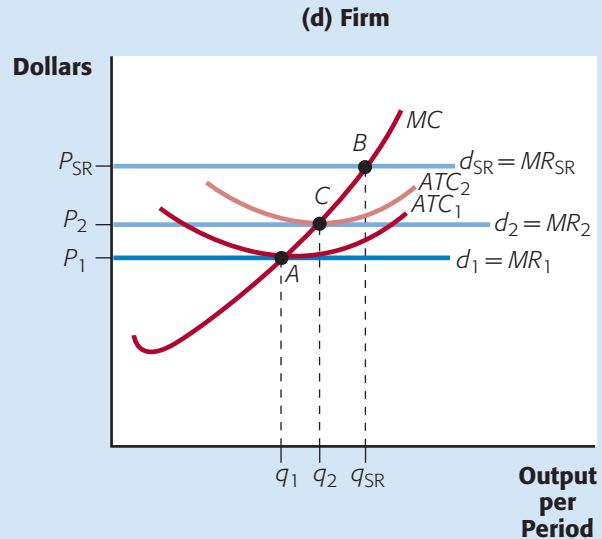
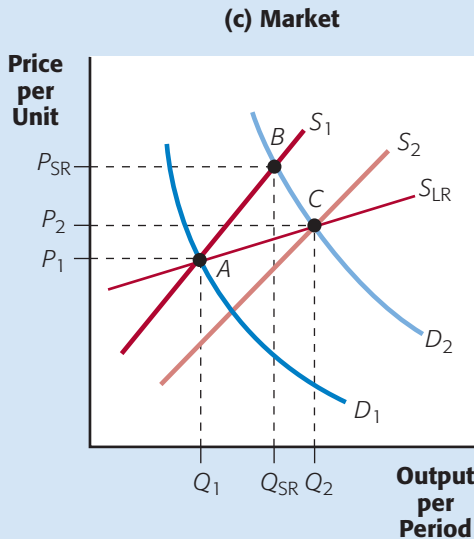
AN INCREASING-COST INDUSTRY

FIGURE 10

INITIAL EQUILIBRIUM



NEW EQUILIBRIUM



At point A in panel (a), the market is in long-run equilibrium. The typical firm in panel (b) earns zero economic profit. If demand increases market price rises. Individual firms increase output and earn an economic profit at point B. Profit attracts entry, increasing market supply and driving up ATC. When long-run equilibrium is reestablished at point C in panel (c), price is higher, but the typical firm again earns zero economic profit. The long-run market supply curve is an upward-sloping line found by connecting points like A and C in panel (c).

more labor, more capital, and more land. We can usually expect the prices of these inputs to rise.

Now, a rise in input prices will affect a firm's ATC curve. Why? Whenever we draw the ATC curve, we assume that the firm's production technology and the prices it must pay for its inputs remain constant. But when inputs become more expensive, cost per unit will be greater at any level of output. As a result, the ATC

curve will shift upward. For example, expansion of the artichoke industry would increase the demand for land suitable for growing this crop, cause the price of this land to rise, and force up the ATC curve facing each artichoke producer.⁴

Let's sum up what we know so far. After the demand curve shifts, we arrive at point B in panel (c) in the short run. At this point, the price is higher and the typical firm is earning economic profit. Profit attracts entry, so the market supply curve begins to shift rightward, bringing the price back down. At the same time, the expansion of output in the industry raises input prices and shifts the typical firm's ATC curve upward. In panel (d), the ATC curve shifts upward to the dashed curve ATC_2 .

Now comes our important conclusion: Since the ATC curve has shifted upward, zero profit will occur at a price *higher* than the initial price P_1 . In panel (d), the typical firm will earn zero profit when the price is P_2 . Thus, entry will cease, and the market supply curve will *stop* shifting rightward, when the market price reaches P_2 . In the figure, this occurs when the market supply curve reaches S_2 . The final, long-run equilibrium occurs at point C , with price P_2 , industry output at Q_2 , and the typical firm producing q_2 .

There is a lot going on in Figure 10. But we can make the story simpler if we *skip over* the short-run equilibrium at point B , and just ask: What happens in the *long run* after the demand curve shifts rightward? The answer is: The market equilibrium will move from point A to point C . A line drawn through these two points tells us, in the long run, the market price we can expect for any quantity the market provides. In Figure 10, this is the thin black line, which is called the *long-run supply curve* (S_{LR}).

Long-run supply curve A curve indicating the quantity of output that all sellers in a market will produce at different prices, after all long-run adjustments have taken place.

The long-run supply curve shows the relationship between market price and market quantity produced after all long-run adjustments have taken place.

If input prices rise when an industry expands (as in our example), then an increase in market quantity will require an increase in the price. This is why the long-run supply curve S_{LR} has an *upward slope* in Figure 10.

However, things don't *have* to end up as in Figure 10. It depends on what happens to input prices as new firms enter the industry and begin demanding inputs along with the firms already there. In Figure 10, the increase in demand for inputs causes the price of those inputs to *rise*. That's why the ATC curve shifts upward, and that's why the long-run supply curve slopes upward. This type of industry—which is the most common—is called an **increasing cost industry**.

Increasing cost industry An industry in which the long-run supply curve slopes upward because each firm's ATC curve shifts upward as industry output increases.

But there are two other (less common) possibilities. One occurs when an industry uses such a small percentage of total inputs that—even as new firms enter—there is no noticeable effect on input prices. For example, the college textbook industry uses a very tiny percentage of the nation's land, labor, and capital, and a relatively small percentage of the nation's paper and ink. This industry could expand considerably without any noticeable rise in input prices. As a result, the ATC curve would stay put as new firms entered the industry and—as you are asked to verify in the end-of-chapter challenge question—the long-run supply curve would be horizontal. This type of industry is a **constant cost industry**.

Constant cost industry An industry in which the long-run supply curve is horizontal because each firm's ATC curve is unaffected by changes in industry output.

⁴ Notice that, in Figure 10, the marginal cost curve does not shift upward. That's because we assume that only *fixed* inputs, like land or factory space, are becoming more expensive. A rise in the price of a fixed input has no effect on the MC curve, which tells us the cost of producing *additional* units of output. However, if *variable* inputs were to rise in price, then both the ATC curve *and* the MC curve would shift upward in Figure 10. That's because a rise in a variable input's price increases the cost per unit of output (ATC) *and* the cost of producing *one more* unit of output (MC).

The last possibility is that of a **decreasing cost industry**, in which entry by new firms actually *decreases* input prices. This will occur when firms that make the *inputs* experience economies of scale, so that their cost per unit—and the prices they charge—come down as they step up production. For example, videotapes are an important input for video rental stores. In the 1980s, the entry of new firms into the video rental industry led to increased demand for, and production of, videotapes. But since videotape producers enjoyed economies of scale, their increased production led to lower costs and—ultimately—to lower prices for videotapes. As a result, the entry of new firms into the video rental industry actually *decreased* average costs for all of them. In a decreasing cost industry like the video rental industry, the long-run supply curve will slope *downward*. (Challenge question 1 at the end of this chapter asks you to verify this.)

Decreasing cost industry An industry in which the long-run supply curve slopes downward because each firm's ATC curve shifts downward as industry output increases.

MARKET SIGNALS AND THE ECONOMY

The previous discussion of changes in demand included a lot of details, so let's take a moment to go over it in broad outline. You've seen that an *increase* in demand always leads to an *increase* in market output in the short run, as existing firms raise their output levels, and an even *greater* increase in output in the long run, as new firms enter the market.

We could also have analyzed what happens when demand *decreases*, but you are encouraged to do this on your own instead, drawing the diagram and tracing through the logic. If you do it correctly, you'll find that the leftward shift of the demand curve will cause a drop in output in the short run and an even greater drop in the long run.

But now let's step back from these details and see what they really tell us about the economy. We can start with a simple fact: In the real world, the demand curves for different goods and services are constantly shifting. For example, over the last decade, Americans have developed an increased taste for bottled water. The average American gulped down 6.4 gallons of the stuff in 1988, and more than twice that much—13.3 gallons—in 1998. As a consequence, the *production* of bottled water has increased dramatically. This seems like magic: Consumers want more bottled water and—presto!—the economy provides it. What our model of perfect competition shows us are the workings behind the magic, the logical sequence of events leading from our desire to consume more bottled water and its appearance on store shelves.

The secret—the trick up the magician's sleeve—is this: As demand increases or decreases in a market, *prices change*. And price changes act as *signals* for firms to enter or exit an industry. How do these signals work? As you've seen, when demand increases, the price tends to initially *overshoot* its long-run equilibrium value during the adjustment process, creating sizable temporary profits for existing firms. Similarly, when demand decreases, the price falls *below* its long-run equilibrium value, creating sizable losses for existing firms. These exaggerated, temporary movements in price, and the profits and losses they cause, are almost irresistible forces, pulling new firms into the market, or driving existing firms out. In this way, the economy is driven to produce whatever collection of goods consumers prefer.

For example, as Americans shifted their tastes toward bottled water, the market demand curve for this good shifted rightward, and the price rose. Initially, the price rose *above* its new long-run equilibrium value, leading to high profits at existing bottled water firms such as Poland Spring and Arrowhead. High profits, in turn, attracted entry—especially the entry of new brands from established firms, such as Pepsi's Aquafina and Coke's Dasani. As a result, production expanded to match the increase in demand by consumers. More of our land, labor, and capital are now used to produce bottled water. Where did these resources come from?

In large part, they were freed up from those industries that experienced a *decline* in demand. In these industries, lower prices have caused exit, freeing up land, labor, and capital to be used in other, expanding industries, such as the bottled water industry.

Market signals Price changes that cause firms to change their production to more closely match consumer demand.

In a market economy, price changes act as market signals, ensuring that the pattern of production matches the pattern of consumer demands. When demand increases, a rise in price signals firms to enter the market, increasing industry output. When demand decreases, a fall in price signals firms to exit the market, decreasing industry output.

Importantly, in a market economy, no single person or government agency directs this process. There is no central command post where information about consumer demand is assembled, and no one tells firms how to respond. Instead, existing firms and new entrants—in their *own* search for higher profits—respond to market signals and help move the overall market in the direction it needs to go. This is what Adam Smith meant when he suggested that individual decision makers act—as if guided by an *invisible hand*—for the overall benefit of society, even though, as individuals, they are merely trying to satisfy their own desires.

CHANGES IN TECHNOLOGY

Using the THEORY



Perfect competition, while it does wonders for society as a whole, is hard on the individual firm. We have seen that economic profit—when it occurs—exists only fleetingly before being eliminated by the entry of other firms. Similarly, economic loss is eliminated by exit—a rather clinical term for thousands of painful business failures each year. But these features of competition make it a powerful engine for satisfying our material desires. In this section, we look at another way in which perfect competition, while rather heartless toward the individual firm, works for the overall benefit of society: the adoption of new technology.

One industry that has experienced especially rapid technological changes in the 1990s is farming. By using genetically altered seeds, farmers are able to grow crops that are more resistant to insects and more tolerant of herbicides. This lowers the total—and average—cost of producing any given amount of the crop.

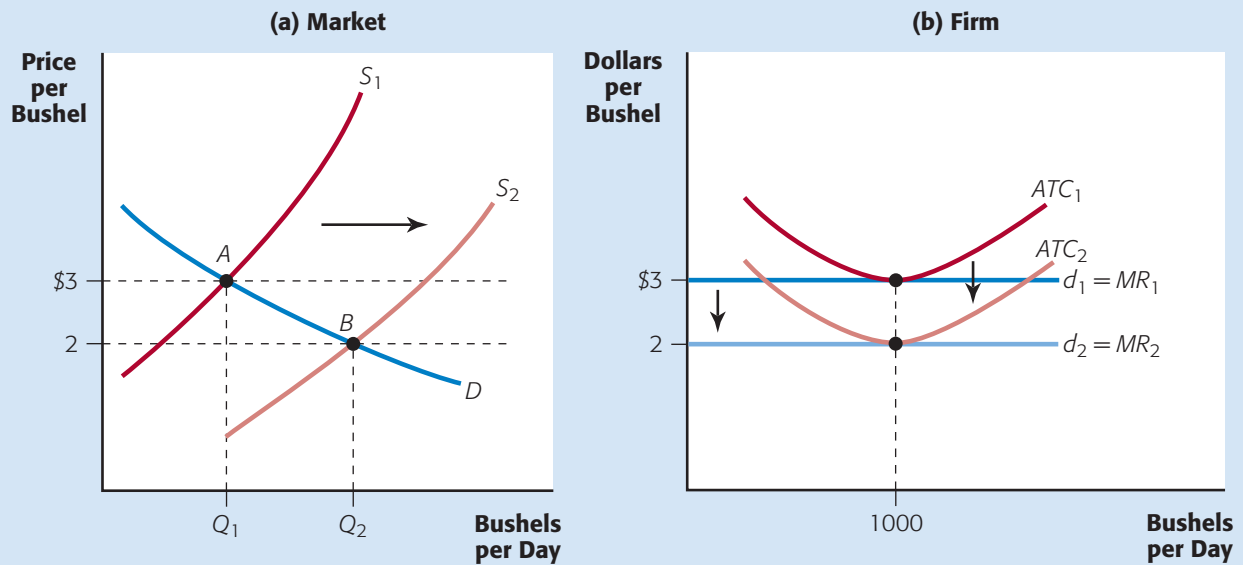
Figure 11 illustrates the market for corn, but it could just as well be the market for soybeans, cotton, or many other crops. In panel (a), the market begins at point *A*, where the price of corn is \$3.00 per bushel. In panel (b), the typical farm produces 1,000 bushels per year and—with average cost given by ATC_1 —earns zero economic profit.

Now let's see what happens when new, higher-yield corn seeds are made available. Suppose first that only one farm uses the new technology. This farm will enjoy a downward shift in its ATC curve from ATC_1 to ATC_2 . Since it is so small relative to the market, it can produce all it wants and continue to sell at \$3.00. Although we have not drawn in the farm's MC curve, you can see that the farm has several output levels from which to choose where $P > ATC$ and it can earn economic profit.

But not for long. In the long run, economic profit at this farm will cause two things to happen. First, all other farmers in the market will have a powerful incentive to adopt the new technology—to plant the new, genetically engineered seed themselves. Under perfect competition, they can do so; there are no barriers that

TECHNOLOGICAL CHANGE IN PERFECT COMPETITION

FIGURE 11



Technological change may reduce ATC . In panel (b), any farm that adopts new technology will earn an economic profit if it can produce at the old market price of \$3 per bushel. That profit will lead its competitors to adopt the same technology and will also attract new entrants. As market supply increases, price falls until each farm is once again earning zero economic profit.

prevent any farmer from using the same technology as any other. As these farms adopt the new seed technology, their ATC curves, too, will drop down to ATC_2 .

Second, outsiders will have an incentive to enter this industry with plants utilizing the new technology, shifting the market supply curve rightward (from S_1 to S_2) and driving down the market price. The process will stop only when the market price has reached the level at which farms using the new technology earn zero economic profit. In Figure 11, this occurs at a price of \$2.00 per bushel.⁵

From this example, we draw two conclusions about technological change under perfect competition. First, what will happen to a farmer who is reluctant to change his technology? As other farms make the change, and the market price falls from \$3.00 to \$2.00, the reluctant farmer will find himself suffering an economic loss, since his average cost will remain at \$3.00. His competitors will leave him to twist in the wind, and if he refuses to shape up, he will be forced to exit the industry. In the end, all farms in the market must use the new technology.

Second, who benefits from the new technology in the long run? Not the farmers who adopt it. Some farmers—the earliest adopters—may enjoy short-run profit before the price adjusts completely, but in the long run, all farmers will be right back where they started—earning zero economic profit. The gainers are consumers of corn, since they benefit from the lower price.

Although some of the data in this example are hypothetical, the story is not. The average American farmer today feeds 129 people, double the amount fed only a few years ago. And as our example suggests, there are powerful forces leading farmers to

⁵ In this example, we assume that the price of the new technology remains the same as it is adopted throughout the industry. If the price of the new technology were to rise, then—in the long run—the typical firm's ATC curve could still shift downward, but not as far as ATC_2 ; the market supply curve would then shift rightward, but not as far as S_2 ; and the price would drop, but not all the way to \$2.

adopt new productivity-enhancing technology. Between 1995 and 1999, the fraction of U.S. corn acreage planted with the new seeds increased from zero to about one-half.

More generally, we can summarize the impact of technological change as follows:

Under perfect competition, a technological advance leads to a rightward shift of the market supply curve, decreasing market price. In the short run, early adopters may enjoy economic profit, but in the long run, all adopters will earn zero economic profit. Firms that refuse to use the new technology will not survive.

Technological advances in many competitive industries—mining, lumber, and farming, for example—have indeed spread quickly, shifting market supply curves rapidly and steadily rightward over the past 100 years. Consumers have reaped huge rewards from these advances, but it has not always been easy on individual firms.

This may explain, at least in part, why many small farmers have lobbied for government limits on new agricultural techniques, such as the genetically altered seeds of our example. Small farmers know they will be the last to obtain these new seeds and so will suffer losses in the short run as other, larger farmers leap ahead of them. And in the long run, the most the small farmer can hope for anyway is a return to zero economic profit. Technological change is, indeed, hard on the small farmer, but it has also enabled the industry as a whole to feed a growing world population at steadily declining prices.

S U M M A R Y

Perfect competition is a market structure in which (1) there are large numbers of buyers and sellers and each buys or sells only a tiny fraction of the total market quantity; (2) sellers offer a standardized product; and (3) sellers can easily enter or exit from the market. While few real markets satisfy these conditions precisely, the model is still useful in a wide variety of cases.

Each perfectly competitive firm faces a horizontal demand curve; it can sell as much as it wishes at the market price. The firm chooses its profit-maximizing output level by setting marginal cost equal to the market price. Its *short-run supply curve* is that part of its *MC* curve that lies above average variable cost. Total profit is profit per unit ($P - ATC$) times the profit-maximizing quantity.

In the short run, market price is determined where the market supply curve—the horizontal sum of all firms' supply curves—crosses the market demand curve. In short-run equilibrium, existing firms can earn a profit (in which case new firms will enter) or suffer a loss (in which case existing firms will exit). Entry or exit will continue until, in the long run, each firm is earning zero economic profit. At each competitive firm in long-run equilibrium, price = marginal cost = minimum average total cost = minimum long-run average total cost.

When demand curves shift, prices change more in the short run than in the long run. The temporary, exaggerated price movements act as market signals, ensuring that output expands and contracts in each industry to match the pattern of consumer preferences.

K E Y T E R M S

market structure
perfect competition
price taker

shutdown price
firm's supply curve
market supply curve

normal profit
long-run supply curve
increasing cost industry

constant cost industry
decreasing cost industry
market signals

R E V I E W Q U E S T I O N S

1. What are the three characteristics that typify a perfectly competitive market? Explain the importance of each characteristic.
2. How do economists justify using the perfectly competitive model to analyze markets that clearly do not satisfy one or more of the assumptions of that model?

3. On a scale of 1 to 5, with 5 being full satisfaction and 1 being no satisfaction at all, rank the following markets in terms of their satisfaction of the three characteristics of the perfectly competitive model. Assign a score for each characteristic and justify your assignment.
 - a. Clothing stores
 - b. Restaurants
 - c. Book publishing
 - d. Home video game production
 - e. Jet aircraft production
4. Why is the demand curve facing a perfectly competitive firm infinitely elastic?
5. “To maximize profit, a perfectly competitive firm should produce the level of output at which marginal cost is equal to price.” True, false, or uncertain? Explain.
6. To calculate profit (or loss) for a perfectly competitive firm, we look at the difference between P and ATC , but to determine the profit-maximizing (or loss-minimizing) level of output, we focus on price and marginal cost. Why?
7. Discuss the following statement: “Economists need to pay more attention to the real business world. Their model of perfect competition predicts that firms in a market will end up earning no profit—nothing above costs. As any accountant can tell you, if you look at the balance sheets of most businesses in any industry, their revenue exceeds their costs; they do, in fact, make a profit.”
8. True, false, or uncertain? Explain your answer.
 - a. A perfectly competitive firm is profitable when price exceeds minimum AVC .
 - b. A competitive firm’s supply curve is just its MC curve.
9. What is the fundamental characteristic that distinguishes the short run from the long run in the analysis of a competitive market?
10. True or false? In a perfectly competitive market, an increase in output requires a high price in the short run, but not in the long run. Justify your answer.

P R O B L E M S A N D E X E R C I S E S

1. In 1999, (1) sales of sport utility vehicles (SUVs) skyrocketed, and (2) the price of gasoline rose. Because SUVs get lower gasoline mileage than the automobiles they replaced, their owners ended up buying more gasoline even as the price per gallon rose. Is this a violation of the law of demand?
2. Suppose that a perfectly competitive firm has the following total variable costs (TVC):

Quantity:	0	1	2	3	4	5	6
TVC :	0	6	11	15	18	22	28
3. Assume that the market for cardboard is perfectly competitive (if not very exciting). In each of the following scenarios, should a typical firm continue to produce or should it shut down in the short run? Draw a diagram that illustrates the firm’s situation in each case.
 - a. Minimum $ATC = \$2.00$
Minimum $AVC = \$1.50$
Market price = $\$1.75$
 - b. $MR = \$1.00$
Minimum $AVC = \$1.50$
Minimum $ATC = \$2.00$
4. The following table gives quantity supplied and quantity demanded at various prices in the perfectly competitive meat packing market:

Price (per lb.)	Q_S (in millions of lbs.)	Q_D
\$1.00	10	100
\$1.25	15	90
\$1.50	25	75
\$1.75	40	63
\$2.00	55	55
\$2.25	65	40

Assume that each firm in the meat-packing industry faces the following cost structure:

Pounds	TC
60,000	\$110,000
61,000	\$111,000
62,000	\$112,000
63,000	\$115,000

3. Assume that the market for cardboard is perfectly competitive (if not very exciting). In each of the following scenarios, should a typical firm continue to produce or should it shut down in the short run? Draw a diagram that illustrates the firm’s situation in each case.
 - a. Minimum $ATC = \$2.00$
Minimum $AVC = \$1.50$
Market price = $\$1.75$
4. What is the profit-maximizing output level for the typical firm? (*Hint*: Calculate MC for each change in

- output, then find the equilibrium price, and calculate MR for each change in output.)
- Is this market in long-run equilibrium? Why or why not? (*Hint*: Calculate ATC .)
 - What do you expect to happen to the number of meat-packing firms over the long run? Why?
- Assume that the kitty litter industry is perfectly competitive and is presently in long-run equilibrium:
 - Draw diagrams for both the market and a typical firm, showing equilibrium price and quantity for the market, and MC , ATC , AVC , MR , and the demand curve for the firm.
 - Your friend has always had a passion to get into the kitty litter business. If the market is in long-run equilibrium, will it be profitable for him to jump in head-first (so to speak)? Why or why not?
 - In a perfect competitive, increasing cost industry, is the long-run supply curve always flatter than the short-run market supply curve? Explain.
 - With a 4-panel diagram similar to Figure 10, show what happens in an *increasing-cost* industry when the market demand curve shifts *leftward*.

CHALLENGE QUESTIONS

- Figure 10 in the chapter shows the long-run adjustment process after an increase in demand. The figure assumes that input prices *rise* as industry output expands. However, in some industries, input prices might *fall* as output expands. This would occur if firms that produce inputs enjoy economies of scale (see Chapter 6). In this case, an increase in the production of inputs would actually lower their cost per unit, and ultimately lower the price of inputs.
 - Redraw Figure 10 under the assumption that input prices *fall* as industry output expands. Illustrate what happens in the short run and in the long run after the market demand curve shifts rightward.
 - Trace out the long-run supply curve for this industry. How does it differ from the long-run supply curve in Figure 10?
- How might your new figure help explain why the prices of personal computers have fallen steadily over the past two decades?
- In rare cases, existing technologies are found to be polluting or physically dangerous, and are banned. Review the “Using the Theory” section of this chapter. Then, show graphically the effects of banning a technology that is in common use in a competitive industry. (*Hint*: After the technology is banned, what will happen to the average cost curve?)

EXPERIENTIAL EXERCISES

- Economics America has an interesting module on the economics of Internet access (<http://www.economicsamerica.org/econedlink/newsline/internet/index.html>). Is provision of Internet access a competitive industry? How would you use supply-and-demand tools to model recent developments in Internet pricing?



- Rent the movie *Trading Places*, starring Eddie Murphy and Dan Ackroyd. Enjoy the movie, but pay special attention to the scene near the end where Billy Ray and Louis participate in an auction of orange juice futures. How does the arrival of new information affect the price of futures contracts? Try to model the situation using supply-and-demand curves.

MONOPOLY

CHAPTER

9

“**M**onopoly” is as close as economics comes to a dirty word. It is often associated with thoughts of extraordinary power, unfairly high prices, and exploitation. Even in the board game *Monopoly*, when you take over a neighborhood by buying up adjacent properties, you exploit other players by charging them higher rent.

The negative reputation is partly deserved—there are, indeed, negative aspects to monopoly. But a mythology has developed around the behavior of monopolies, one full of exaggerations, half-truths, and falsehoods. Many monopolies are socially harmful, but in some instances a monopoly may be the best way to organize production.

This chapter deals with monopolies in several respects: what they are, how they arise, and how they behave in different environments. With a few exceptions, we will focus on *understanding* rather than *assessment*, postponing a full discussion of what is good and bad about monopoly until Chapters 14 and 15.

WHAT IS A MONOPOLY?

In most of your purchases—a haircut, a meal at a restaurant, a car, a college education—more than one seller is competing for your dollars, and you can choose which one to buy from. But in some markets, you have no choice at all. If you want to mail a letter for normal delivery, you must use the U.S. Postal Service. If you want cable television service, you must use the one cable television company in your area. Many cities have only a single local newspaper. And if you live in a small town, you may have just one doctor, one gas station, or one movie theater to select from. These are all examples of *monopolies*:

A monopoly firm is the only seller of a good or service with no close substitutes. The market in which the monopoly firm operates is called a monopoly market.

A key concept in the definition of monopoly is the notion of *substitutability*. There is usually more than one way to satisfy a desire, and a single seller of a good or service is *not* considered a monopoly if other firms sell products—close

CHAPTER OUTLINE

What Is a Monopoly?

The Sources of Monopoly

- Economies of Scale
- Control of Scarce Inputs
- Government-Enforced Barriers

Monopoly Goals and Constraints

- Monopoly Price or Output Decision
- Profit and Loss

Equilibrium in Monopoly Markets

- Short-Run Equilibrium
- Long-Run Equilibrium
- Comparing Monopoly to Perfect Competition
- Why Monopolies Often Earn Zero Economic Profit

What Happens When Things Change?

Price Discrimination

- Requirements for Price Discrimination
- Effects of Price Discrimination

The Decline of Monopoly

Using the Theory: Price Discrimination at Colleges and Universities

Monopoly firm The only seller of a good or service that has no close substitutes.

Monopoly market The market in which a monopoly firm operates.



Characterize the Market

substitutes—that satisfy that same desire. For example, only one firm in the country—Kellogg—sells Kellogg’s Corn Flakes. But other cereal companies sell their own brands of cornflakes, which are close substitutes for Kellogg’s. And many other types of flaky cereals—wheat flakes or oat flakes—are also very close substitutes for Kellogg’s Corn Flakes. This is why we do not consider Kellogg to be a monopoly firm.

The definition of a monopoly firm or market may seem precise. But in the real world, the definition is not always so clear-cut, because it depends on how broadly or narrowly we define the market we’re analyzing. In general, in deciding whether a market is or is not a monopoly, we should include in the market all products that are *close substitutes* for the product in question. But how close must a substitute for a product be before we include it in the market?

Consider, for example, cable television service in the United States. A couple of paragraphs ago, it was one of our examples of a monopoly. But you can also get movies and other entertainment on broadcast television, at your local video store, and even over the Internet. If we include each of these products as part of a broadly defined market for entertainment services, then there are several sellers, and your cable company is *not* a monopoly. But are these other sources of entertainment *close* substitutes? For some people and some purposes, yes. But for many purposes, a cable service has no close substitutes: Using the Internet to download videos requires a special high-speed connection and a home theater that takes computer files; video stores only rent movies; and broadcast television is aimed at a broader audience than most cable channels, which tend to target specific viewers. If we define the market more narrowly as that for “reliable, in-home, specialized entertainment services,” the local cable company looks like a monopoly again.

Or consider the market for horror novels. Surely, we might think, this market is not a monopoly. After all, any reasonable definition of the market would include the products of several different publishers, all of them competing for your dollars when you browse the aisles of your local bookstore. But if you really want to read Stephen King’s latest novel, then as far as you are concerned, there is only one seller: Scribner, which publishes all of Stephen King’s books.

Because we all have different tastes and characteristics, we can have different opinions about what is, and what is not, a “close” substitute. As a result, we can have different ideas about how broadly or how narrowly we should define a market when trying to decide if it is a monopoly. It makes sense, then, to view monopoly as a spectrum rather than a strict category. On one end of this spectrum is *pure monopoly*, where there is just one seller of a good for which very few buyers could find a substitute. The only doctor, attorney, or food market in a small town comes very close to being a pure monopoly. Farther along the spectrum, we reach firms that sell a good for which reasonable substitutes do exist—at least for some buyers and for some purposes—but they are not very *close* substitutes for most buyers or most purposes. The local cable company is an example of this middle ground, and most economists would extend the label “monopoly” to this part of the spectrum. But as we go farther along the spectrum, we find goods for which so many buyers can find close substitutes that the term *monopoly* no longer makes sense. Scribner is an example of this kind of firm. If you are a devoted Stephen King fan and will accept no substitutes, then for you personally, Scribner might seem like a monopolist. But for most people, other authors will substitute reasonably well. This is why we do not consider Scribner to be a monopoly.

THE SOURCES OF MONOPOLY

In a perfectly competitive market, there are *no* significant barriers to entry by new firms. Monopoly, by contrast, arises *because* of barriers to entry. In this section, we consider the three most common barriers responsible for creating and maintaining monopoly markets: economies of scale, control of a scarce input, and barriers created by government.

ECONOMIES OF SCALE

Recall from Chapter 6 that economies of scale in production cause a firm's long-run average cost curve to slope downward. That is, the more output the firm produces, the lower will be its cost per unit. If economies of scale persist to the point where a single firm is producing for the entire market, we call the market a *natural monopoly*:

A natural monopoly exists when, due to economies of scale, one firm can produce at a lower cost per unit than can two or more firms.

The monopoly firm, or the market in which it operates, is called a *natural monopoly* for a good reason: Unless the government steps in, only one seller would survive—the market would *naturally* evolve into a monopoly. Why is this? Because, once a firm is already established, a new entrant would have to charge a lower price than the existing firm in order to attract customers. The existing firm would then lower *its* price in order to hang onto its customers. But in this battle of prices, the existing firm has a strong advantage: As the sole seller in the market, it already produces more output than the new entrant could hope to produce. Thus, its cost per unit is already lower than that of the entrant. The existing firm can lower its price to just a shade above its low cost per unit and still earn a small profit. But if the new entrant—with higher cost per unit—tries to match this price, it will suffer a loss. Anticipating this result, potential entrants will stay away.

Small local monopolies are almost always *natural* monopolies. Think of the sole gas station in a small town. Since it needs a minimum set of fixed inputs no matter how little gas it sells (a pump for each type of gas, space for cars to pull up, a zoning permit), each additional gallon sold lowers the station's cost per unit. By producing for the entire (small town) market, it has achieved the smallest possible cost per unit. Under these circumstances, a potential new entrant would have to think very hard about coming into this market, since it would not be able to survive a price war with the firm already in the market. The same logic can explain the monopoly position held by the sole movie theater, food market, or dentist in a small town. These are all natural monopolies, because they continue to enjoy economies of scale up to the point at which they are serving the entire market.¹

CONTROL OF SCARCE INPUTS

Some firms maintain their monopoly status by controlling a scarce input needed to produce a good. For example, from 1893 until the 1940s, Alcoa (the Aluminum Company of America) was the sole seller of aluminum in the United States because it owned virtually all of the country's deposits of bauxite—a natural

Natural monopoly A market in which, due to economies of scale, one firm can operate at lower average cost than can two or more firms.



De Beers Consolidated Mines is one of the most famous examples of monopoly. For more information, click on <http://www.edata.co.za/DeBeers>.

¹ Are you wondering what determines the size of the “entire market”? You’ll find out in the next chapter, in which we revisit the subject of natural monopolies, and also discuss natural oligopolies.

resource needed to produce aluminum. Similarly, since the 1880s, De Beers, a South African company, has enjoyed a near monopoly on the sale of finished diamonds by buying up most of the world's diamond mines or the raw diamonds that come from them.

GOVERNMENT-ENFORCED BARRIERS

Sometimes, the public interest is best served by having a single seller in a market. In these cases, government usually steps in and creates barriers to entry, ensuring that the market will remain a monopoly. In the United States, monopolies have been created by all levels of government—federal, state, and local. The two main methods of creating a monopoly are (1) the protection of intellectual property through patents, trademarks, and copyrights and (2) exclusive government franchises.

Protection of Intellectual Property. The words you are reading right now are an example of *intellectual property*, which includes literary, artistic, and musical works, as well as scientific inventions. Most markets for a specific intellectual property are monopolies: One firm or individual owns the property and is the sole seller of the rights to use it. There is both good and bad in this. As you will learn in this chapter, prices tend to be higher under monopoly than under perfect competition, and monopolies often earn economic profit as a consequence. This is good for the monopoly and bad for everyone else. On the other hand, it is just this promise of monopoly profit that encourages the creation of original products and ideas, which certainly benefits the rest of us. The Palm Pilot personal organizer, the Visex laser for reshaping the eye's cornea, and second-generation Internet search engines such as Google, About.com, and Direct Hit were all launched by innovators who bore considerable costs and risks with an expectation of future profits. The same is true of every compact disc you listen to, every novel you read, and every movie you see.

In dealing with intellectual property, government strikes a compromise: It allows the creators of intellectual property to enjoy a monopoly and earn economic profit, but only for a limited period of time. Once the time is up, other sellers are allowed to enter the market, and it is hoped that competition among them will bring down prices.

The two most important kinds of legal protection for intellectual property are *patents* and *copyrights*. New scientific discoveries and the products that result from them are protected by a **patent** obtained from the federal government. The patent prevents anyone else from selling the same discovery or product for about 20 years. The Eli Lilly Company, for example, holds a patent on the chemical fluoxetine, the active ingredient in Prozac—the first antidepressant drug without severe side effects. Lilly has made an enormous profit from Prozac, selling almost \$2.5 billion worth of the drug in 1999 alone. Other pharmaceutical companies, forced to work around Lilly's patent, took much longer to develop their own, similar drugs. In the meantime, Lilly was the sole seller of a product with no close substitutes.

Literary, musical, and artistic works are protected by a **copyright**, which grants exclusive rights over the material for at least 50 years. For example, the copyright on this book is owned by South-Western College Publishing. No other company or individual can print copies and sell them to the public, and no one can quote the book at length without obtaining South-Western's permission.

Patent A temporary grant of monopoly rights over a new product or scientific discovery.

Copyright A grant of exclusive rights to sell a literary, musical, or artistic work.

Copyrights and patents are often sold to another person or firm, but this does not change the monopoly status of the market, since there is still just one seller. For example, Paul McCartney has purchased the copyright to hundreds of songs he did not compose, including the song “Happy Birthday.” While you are free to sing this song at a private birthday party, anyone who wants to sing it on radio or television—that is, anyone who wants to profit from the song—must obtain a license from McCartney and pay him a small royalty.

Government Franchise. The large firms we usually think of as monopolies—utility, telephone, and cable television companies—have their monopoly status guaranteed through **government franchise**—a grant of exclusive rights over a product. Here, the barrier to entry is quite simple: Any other firm that enters the market will be prosecuted!

Governments usually grant franchises when they think the market is a *natural monopoly*. In this case, a single large firm—enjoying economies of scale—would have a lower cost per unit than multiple smaller firms, so government tries to serve the public interest by *ensuring* that there are no competitors. In exchange for its monopoly status, the seller must submit to either government ownership and control or else government regulation over its prices and profits.

This is the logic behind the monopoly status of the U.S. Postal Service. No matter how many letters it delivers, a postal firm must have enough letter carriers to reach every house every day. Two postal companies would need many more carriers to deliver the same total number of letters, raising costs and, ultimately, the price of mailing a letter. Thus, mail delivery is a natural monopoly, one that the federal government has chosen to own and control rather than merely regulate. Federal law prohibits any other firm from offering normal letter-delivery service.

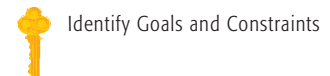
Local governments, too, create monopolies by granting exclusive franchises in a variety of industries believed to be natural monopolies. These include utility companies that provide electricity, gas, and water, as well as garbage collection services.



Since ordinary letter delivery is a natural monopoly, the U.S. Postal Service has been granted an exclusive government franchise to deliver the mail.

Government franchise A government-granted right to be the sole seller of a product or service.

MONOPOLY GOALS AND CONSTRAINTS



The goal of a monopoly—like that of any firm—is to earn the highest profit possible. And, like other firms, a monopolist faces constraints.

Reread that last sentence because it is important. It is tempting to think that a monopolist—because it faces no direct competitors in its market—is free of constraints. Or that its constraints are special ones, unlike those of any other firm. For example, many people think that the only force preventing a monopolist from charging outrageously high prices is public outrage. In this view, your cable company would charge \$200, \$500, or even \$10,000 per month if only it could “get away with it.”

But with a little reflection, it is easy to see that a monopolist faces purely *economic* constraints that limit its behavior—constraints that are similar to those faced by other, non-monopoly firms. What are these constraints?

First, there is a constraint on the monopoly’s *costs*: For any level of output the monopolist might produce, it must pay some total cost to produce it. This cost constraint is determined by the monopolist’s production technology—which tells it how much output it can produce with different combinations of inputs—and also by the prices it must pay for those inputs. In other words, the constraints on the monopolist’s costs are the same as for any other type of firm, such as the perfectly competitive firm we studied in the previous chapter.

There is also a *demand constraint*. The monopolist's demand curve—which is also the market demand curve—tells us the maximum price the monopolist can charge to sell any given quantity of output.²

To sum up:

A monopolist, like any firm, strives to maximize profit. And, like any firm, it faces constraints. For any level of output it might produce, total cost is determined by (1) its technology of production and (2) the prices it must pay for its inputs. And for any level of output it might produce, the maximum price it can charge is determined by the market demand curve for its product.

MONOPOLY PRICE OR OUTPUT DECISION

Notice that the title of this section reads “price *or* output decision,” not “price *and* output decision.” The reason is that noncompetitive firms—such as monopolies—do *not* make two separate decisions about price and quantity, but rather *one* decision. More specifically, once the firm determines its output level, it has also determined its price (the maximum price it can charge and still sell that output level). Similarly, once the firm determines its price, it has also determined its output level (the maximum output the firm can sell at that price). To keep things simple, we'll focus on the firm's *output* decision, and then determine the maximum price that will enable the firm to sell the output level on which it has decided.

Suppose a monopolist is considering selling more output. Then, since it faces a downward-sloping demand curve, it must lower its price. However, the new, lower price must be charged *not* just on the new, additional units it wants to sell, but on *all* units of output, including those it was previously selling at a higher price. For example, if your local cable television company wants more subscribers, it will have to lower its rates for everyone, including those who already subscribe at the current rate. Thus, lowering price and increasing output has two offsetting effects on total revenue: On the one hand, more output is sold, tending to *increase* total revenue; on the other hand, *all* units now go for a lower price, tending to *decrease* total revenue. The net effect may be a rise or fall in total revenue, or—another way to say the same thing—the firm's *marginal revenue* may be positive or negative.³

Figure 1 illustrates the demand and marginal revenue curves for Zillion-Channel Cable—a monopoly that sells cable television services to the residents of a town. We will assume that Zillion-Channel is free from government regulation. In the figure, the demand curve shows the number of subscribers at each monthly price for cable. The demand curve is both a *market* demand curve and the demand curve *facing the firm*, since Zillion-Channel is the only firm in its market.

Let's see what happens to Zillion-Channel's revenue as we move from point *A* to point *B* along its demand curve. At point *A*, the firm charges a monthly price of \$50

² The demand constraint can be simple or complex, depending on whether the monopolist must charge the same price on every unit of output it sells, or can charge different prices to different customers or on different units. For now, we're considering the case of the *single-price* monopolist—one that must charge the same, single price on all units. Later in this chapter, we'll discuss what happens when a monopolist can charge several different prices simultaneously.

³ This should sound familiar to you. In Chapter 7, Ned's Beds faced a downward-sloping demand curve and had to lower its price on all of its bed frames in order to sell more of them. Although Ned was not necessarily the only seller in his market, his total and marginal revenue behaved in much the same way as we are describing here.

DEMAND AND MARGINAL REVENUE

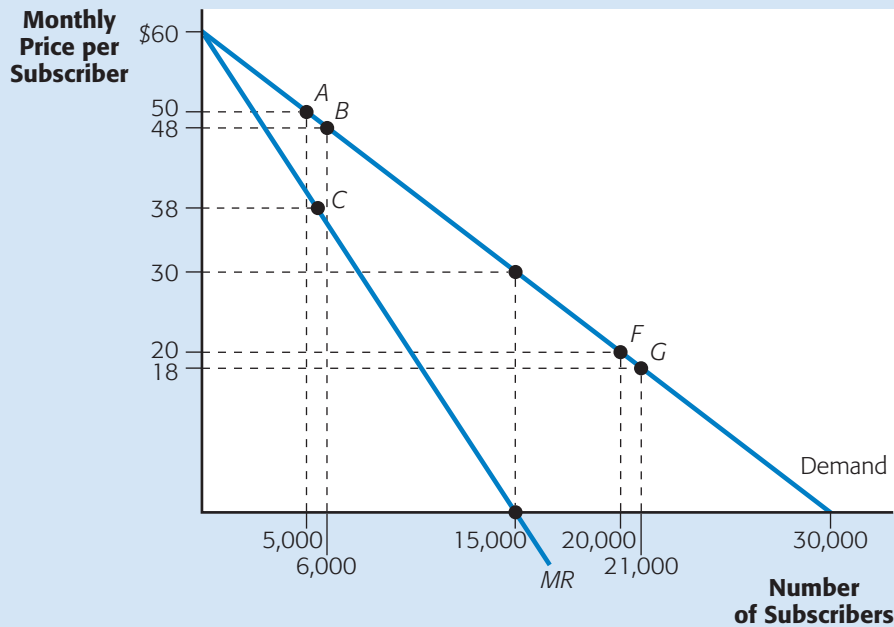


FIGURE 1

A monopoly faces a downward-sloping market demand curve. To sell additional output, the firm must lower its price. Marginal revenue (MR) shows the change in total revenue that results from a one-unit increase in output. MR is less than price; to sell an additional unit, the monopoly must lower the price on that unit *and* on previous units.

and attracts 5,000 paid subscribers, for a total revenue of $5,000 \times \$50 = \$250,000$. If it lowers its price to \$48, moving to point B , 1,000 more people will subscribe, for a total revenue of $6,000 \times \$48 = \$288,000$. Thus, in moving from point A to point B , total revenue rises by \$38,000. The marginal revenue for this move—which tells us the increase in revenue per additional unit of output—can be calculated as follows:

$$MR = \frac{\Delta TR}{\Delta Q} = \frac{(\$288,000 - \$250,000)}{(6,000 - 5,000)} = \frac{\$38,000}{1,000} = \$38$$

This value for marginal revenue—\$38—is plotted at point C , which is midway between points A and B . Notice that MR is *less* than the new price of output, \$48. This follows from the two competing effects just discussed: On the one hand, the monopoly is selling more output and getting \$48 on each additional unit; on the other hand, it has to charge a lower price on the previous 5,000 units of output it was selling. In the move from A to B , it turns out that total revenue rises, and marginal revenue is positive, but less than \$48.

When any firm—including a monopoly—faces a downward-sloping demand curve, marginal revenue is less than the price of output. Therefore, the marginal revenue curve will lie below the demand curve.

For other moves along the demand curve, total revenue may decline, so marginal revenue will be negative. (Verify this for the move from point F to point G .) For such changes, the marginal revenue curve lies below the horizontal axis.

The marginal revenue curve alone tells us something about the monopoly's output decision:

A monopoly will never produce a level of output at which its marginal revenue is negative.

We can be sure of this principle by simple logic. If marginal revenue is negative, then producing more output will *decrease* the firm's total revenue. But producing more output will also *increase* the firm's total cost. Since revenue will fall and cost will rise, profit will decrease. Thus, a firm operating in a range of negative marginal revenue will reduce its profit by producing more. Conversely, it can always increase profit by producing less.⁴ Therefore, it will never want to maintain production in a range over which marginal revenue is negative. Zillion-Channel, for example, will never want to charge less than \$30 or have more than 15,000 subscribers.

Knowing that a monopoly will produce only where marginal revenue is positive narrows down the possibilities somewhat . . . but not enough. Which of the many output levels smaller than 15,000 units will Zillion-Channel choose? To answer this, we return to the rule (from Chapter 7) that allows *any* firm to find its profit-maximizing output level:

To maximize profit, the firm should produce the level of output where $MC = MR$ and the MC curve crosses the MR curve from below.

Figure 2 adds Zillion-Channel's marginal cost curve to the demand and marginal revenue curves of Figure 1. The greatest profit possible occurs at an output level of 10,000, where the MC curve crosses the MR curve from below. In order to sell this level of output, the firm will charge a price of \$40, locating at point E on its demand curve. You can see that for a monopoly, *price and output are not independent decisions, but different ways of expressing the same decision*. Once Zillion-Channel determines its profit-maximizing output level (10,000 units), it has also determined its profit-maximizing price (\$40), and vice versa.

PROFIT AND LOSS

In Figure 2, we can determine Zillion-Channel's price and output level, but we cannot see whether the firm is making an economic profit or loss. This will require one

more addition to the diagram—the average cost curve. Remember that

$$\text{Profit per Unit} = P - ATC.$$

Now, the price, P , at any output level is read off the demand curve. Profit per unit, then, is just the distance between the firm's demand curve and its ATC curve.



A question may have occurred to you: Where is the monopoly's *supply curve*? The answer is that *there is no supply curve for a monopoly*. A firm's supply curve tells us how much output a firm will want to produce and sell when it is *presented* with different prices. This makes sense for a perfectly competitive firm that takes the market price as given and responds by deciding how much output to produce. A monopoly, by contrast, is *not* a price taker; it *chooses* its price. Since the monopolist is free to choose any price it wants—and it will always choose the *profit-maximizing* price and no other—the notion of a supply curve does not apply to a monopoly.

⁴ We can also state this result in terms of the price elasticity of demand: A single-price monopoly should never produce at an output level where demand is inelastic. Why is this? In Chapter 4, you learned that a price hike will increase total expenditure on a good when demand is inelastic. Since a monopoly is the only seller in its market, the total expenditure of consumers is the total revenue of the monopoly. Thus, when demand is inelastic, a price hike will increase the monopoly's total revenue. At the same time, it will decrease output and therefore total cost, so total profit must rise. We conclude that when demand is inelastic, the monopoly can always increase its profit by raising its price and producing less.

MONOPOLY PRICE AND OUTPUT DETERMINATION

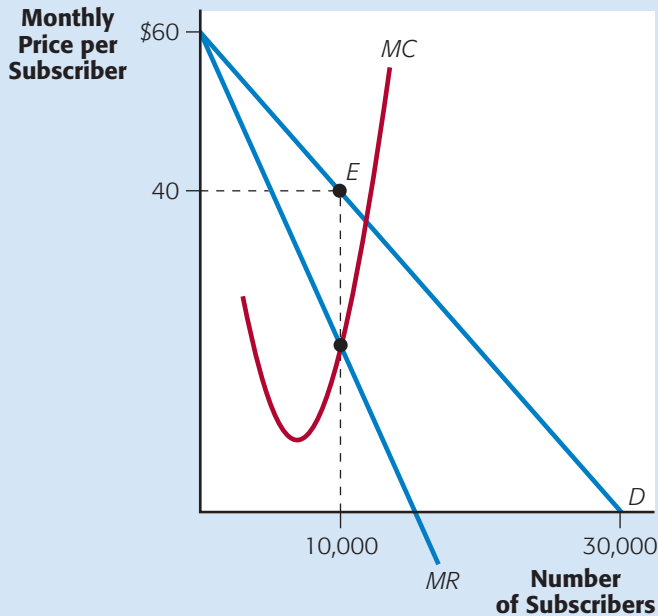


FIGURE 2

Like any firm, the monopolist maximizes profit by producing where MC equals MR . Here, that quantity is 10,000 units. The price charged (\$40) is read off the demand curve. It is the highest price at which the monopolist can sell that level of output.

Figure 3(a) is just like Figure 2 but adds Zillion-Channel's ATC curve. As you can see, at the profit-maximizing output level of 10,000, price is \$40 and average total cost is \$32, so profit per unit is \$8.

Now look at the blue rectangle in the figure. The height of this rectangle is profit per unit (\$8), and the width is the number of units produced (10,000). The *area* of the rectangle—height \times width—equals Zillion-Channel's total profit, or $\$8 \times 10,000 = \$80,000$.

A monopoly earns a profit whenever $P > ATC$. Its total profit at the best output level equals the area of a rectangle with height equal to the distance between P and ATC and width equal to the level of output.

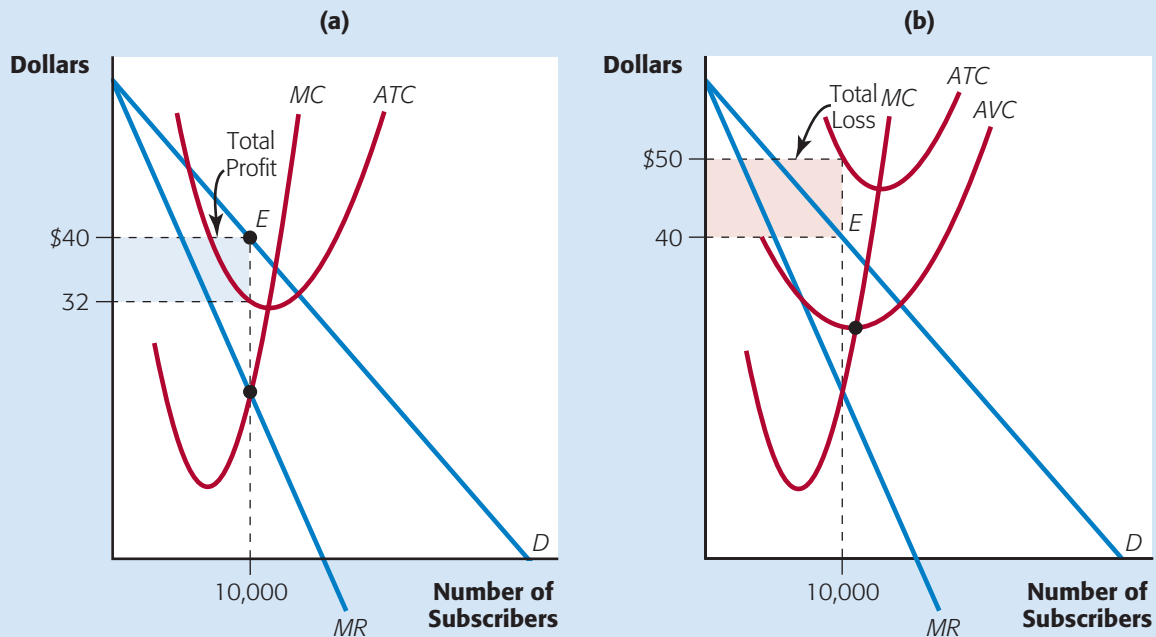
This should sound familiar: It is exactly how we represented the profit of a perfectly competitive firm (compare with Figure 3(a) in Chapter 8). The diagram looked different under perfect competition because the firm's demand curve was horizontal, whereas for a monopoly it is downward sloping.

Figure 3(b) illustrates the case of a monopoly suffering a loss. Here, costs are higher than in panel (a), and the ATC curve lies everywhere above the demand curve, so the firm will suffer a loss at any level of output. At the best output level—10,000— ATC is \$50, so the loss per unit is \$10. The total loss (\$100,000) is the area of the red rectangle, whose height is the loss per unit (\$10) and width is the best output level (10,000). Being a monopolist is no guarantee of profit. If costs are too high, or demand is insufficient, a monopolist may break even or suffer a loss.

A monopoly suffers a loss whenever $P < ATC$. Its total loss at the best output level equals the area of a rectangle with height equal to the distance between ATC and P and width equal to the level of output.

FIGURE 3

MONOPOLY PROFIT AND LOSS



In panel (a), the monopolist's profit is the difference between price and average total cost (ATC) multiplied by the number of units sold. The blue area indicates a profit of \$80,000. Panel (b) shows a monopolist suffering a loss. At the best level of output, ATC exceeds price. The red rectangle shows a loss of \$100,000.

EQUILIBRIUM IN MONOPOLY MARKETS

A monopoly market is in equilibrium when the only firm in the market—the monopoly firm—is maximizing its profit. After all, once the firm is producing the profit-maximizing quantity—and charging the highest price that will enable it to sell that quantity—it has no incentive to change either price or quantity, unless something in the market changes (which we'll explore later).

But for monopoly—as for perfect competition—we have different expectations about equilibrium in the short run and equilibrium in the long run.

Find the Equilibrium



SHORT-RUN EQUILIBRIUM

In the short run, a monopoly may earn an economic profit or suffer an economic loss. (It may, of course, break even as well; see if you can draw this case on your own.) A monopoly that is earning an economic profit will continue to operate in the short run, charging the price and producing the output level at which $MR = MC$, as in Figure 3(a).

But what if a monopoly suffers a loss in the short run? Then it will have to make the same decision as any other firm: to shut down or not to shut down. The rule you learned in Chapter 7—that a firm should shut down if $TR < TVC$ at the output level where marginal revenue and marginal cost are equal—applies to any firm, including a monopoly. As you learned in Chapter 8, we can also express the shut-down rule in terms of price and average variable cost:

Any firm—including a monopoly—should shut down if $P < AVC$ at the output level where $MR = MC$.

That is, if the firm's price per unit cannot cover its variable or operating costs per unit at its best output level (where $MR = MC$), then the firm should shut down.

In Figure 3(b), Zillion-Channel is suffering a loss, but since $P = \$40$ and AVC is less than $\$40$, we have $P > AVC$, and the firm should keep operating. On your own, draw in an alternative AVC curve in panel (b) that would cause Zillion-Channel to shut down. (*Hint: It will be higher than the existing AVC curve.*)

In some cases, the shutdown rule will accurately and realistically predict when a monopoly will shut down in the short run. In other cases, it will not. Many monopolies produce a vital service, such as transportation or communications, and these monopolies typically operate under government regulation. Suppose the monopoly experiences a temporary upward shift in its AVC curve—say, because the price of a variable input rises. Or suppose it experiences a temporary leftward shift of its demand curve—say, because household income decreases. In either case, if the monopoly suddenly finds that $P < AVC$, government will usually not allow it to shut down, but instead use tax revenue to make up for the firm's losses.

LONG-RUN EQUILIBRIUM

One of the most important insights of the previous chapter was that perfectly competitive firms *cannot* earn a profit in long-run equilibrium. Profit attracts new firms into the market, and market production increases. This, in turn, causes the market price to fall, eliminating any temporary profit earned by a competitive firm.

But there is no such process at work in a monopoly market, where barriers *prevent* the entry of other firms into the market. Outsiders will *want* to enter an industry when a monopoly is earning above economic profit, but they will be *unable to do so*. Thus, the market provides no mechanism to eliminate monopoly profit, and

unlike perfectly competitive firms, monopolies may earn economic profit in the long run.

What about economic loss? If a monopoly is a government franchise, and it faces the prospect of long-run loss, the government may decide to subsidize it in order to keep it running—especially if it provides a vital service like mail delivery or mass transit. But if the monopoly is privately owned and controlled, we do not expect to see long-run losses. A monopoly suffering an economic loss that it expects to continue indefinitely should always exit the industry, just like any other firm.

A privately owned monopoly suffering an economic loss in the long run will exit the industry, just as would any other business firm. In the long run, therefore, we should not find privately owned monopolies suffering economic losses.

COMPARING MONOPOLY TO PERFECT COMPETITION

We have already seen one important difference between monopoly and perfectly competitive markets: In perfect competition, economic profit is relentlessly reduced to zero by the entry of other firms; in monopoly, economic profit can continue indefinitely.



Macrosoft is a monopoly simulation written by Peter Wilcoxon of the University of Texas. Try it at <http://www.eco.utexas.edu/faculty/Wilcoxon/games/macsoft/index.htm>.



Find the Equilibrium

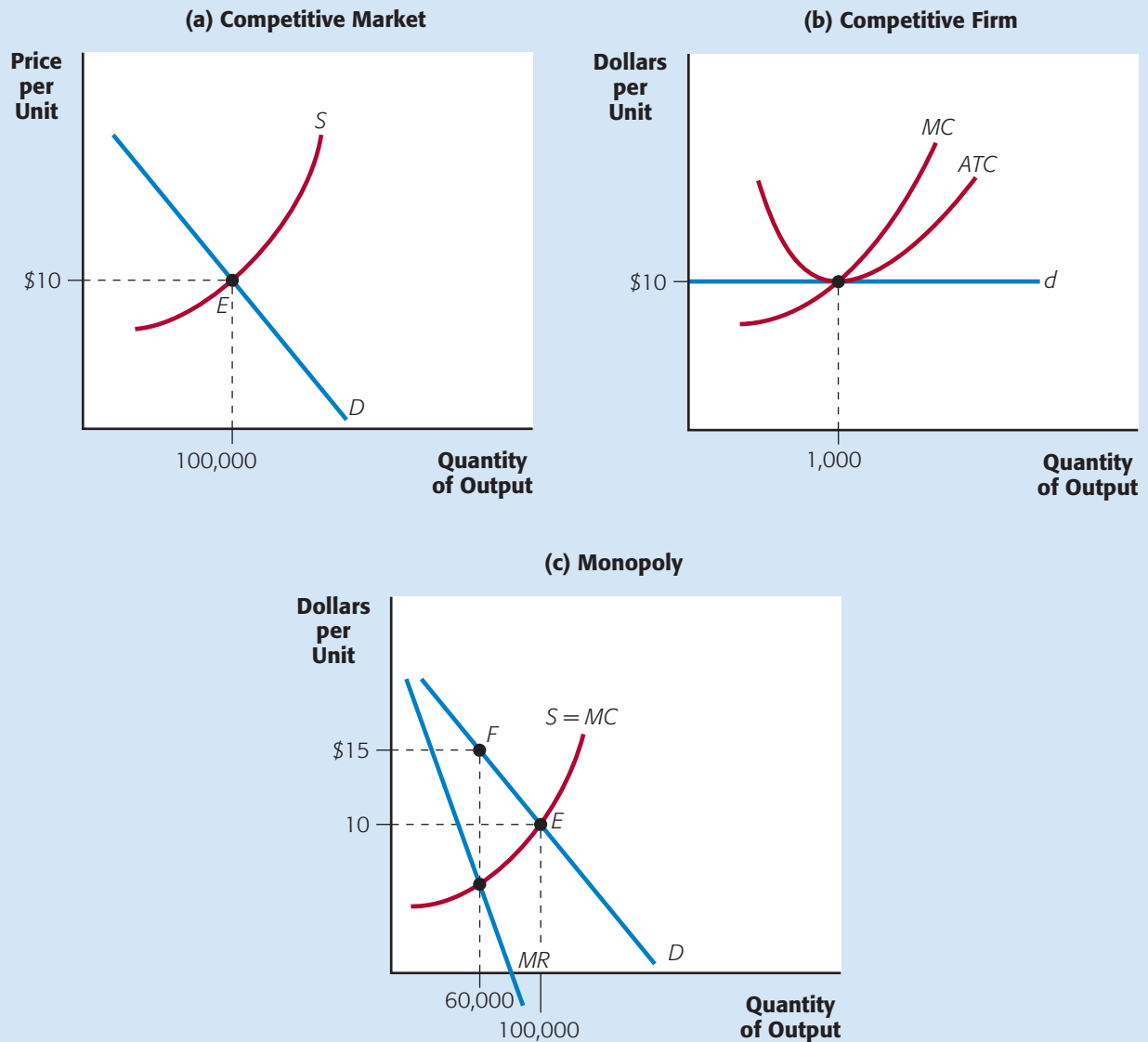
But monopoly also differs from perfect competition in another way:

We can expect a monopoly market to have a higher price and lower output than an otherwise similar perfectly competitive market.

To see why this is so, let's explore what would happen if a single firm took over a perfectly competitive market, changing the market to a monopoly. Figure 4 illustrates a competitive market consisting of 100 identical firms. The market is in long-

FIGURE 4

COMPARING MONOPOLY AND PERFECT COMPETITION



Panel (a) shows a competitive market with 100 identical firms. The market price is \$10 per unit; at that price, each firm (panel b) sells 1,000 units and earns zero economic profit. A monopolist that buys up all these firms will face the market demand curve D in panel (c). It will produce 60,000 units—where $MR = MC$. The monopolist produces less than the competitive firms did and charges a higher price (\$15 rather than \$10).

run equilibrium at point E , with a market price of \$10 and market output of 100,000 units. In panel (b), the typical firm faces a horizontal demand curve at \$10, produces output of 1,000 units, and earns zero economic profit.

Now, imagine that a single company buys all 100 firms, to form a monopoly. The new monopoly market is illustrated in panel (c). Under monopoly, the horizontal demand curve facing each firm becomes irrelevant. Instead, the demand curve the monopoly cares about is the downward-sloping *market* demand curve D , which is the same as the market demand curve in panel (a). Since the demand curve slopes downward, marginal revenue will be less than price, and the MR curve will lie everywhere below the demand curve. To maximize profit, the monopoly will want to find the output level at which $MC = MR$. But what is the new monopoly's MC curve?

We'll assume that the monopoly doesn't change the way output is produced: Each previously competitive firm will continue to produce its output with the same technology as before, only now it operates as one of 100 different plants that the monopoly controls. With this assumption, *the monopoly's marginal cost curve will be the same as the market supply curve in panel (a)*. Why? First, remember that the market supply curve is obtained by adding up each individual firm's supply curve—that is, each individual firm's marginal cost curve. Therefore, the market supply curve tells us the marginal cost—at *some* firm—of producing another unit of output for the market. When the monopoly takes over each of these individual firms, the market supply curve tells us how much it will cost the *monopoly* to produce another unit of output at one of its plants. For example, point E on the market supply curve tells us that, when total supply is 100,000, with each plant producing 1,000, increasing output by one more unit will cost the monopoly \$10, because that is the marginal cost at each of its plants. The same is true at every other point along the competitive market supply curve: It will always tell us the monopoly's cost of producing one more unit at one of the plants it now owns. In other words, the upward-sloping curve in panel (c), which is the market supply curve when the market is competitive, becomes the marginal cost curve for a single firm when the market is monopolized. This is why the curve is labeled both S (market supply) and MC (the marginal cost of the monopolist).

Now we have all the information we need to find the monopoly's choice of price and quantity. In panel (c), the monopoly's MC curve crosses the MR curve from below at 60,000 units of output. This will be the monopoly's profit-maximizing output level. To sell this much output, the monopoly will charge \$15 per unit—point F on its demand curve.

Notice what has happened in our example: After the monopoly takes over, the price rises from \$10 to \$15, and market quantity drops from 100,000 to 60,000. The monopoly, compared to a competitive market, would *charge more and produce less*.

Why? Because the monopoly—unlike each competitive firm—faces a downward-sloping demand curve. As a result, for the monopoly, marginal revenue is less than price. While a competitive firm can sell another unit of output and gain the price as additional revenue, when a monopolist sells another unit, it gains *less* than the price as additional revenue. Therefore, the monopoly will stop increasing its production at a lower level of output than would an industry of perfectly competitive firms. Of course, since the monopoly wants to sell a lower market quantity, it will charge a higher market price.

Now let's see who gains and who loses from the takeover. By raising price and restricting output, the new monopoly earns economic profit. We know this because at a price of \$10—the competitive price—each of its plants would break even, so at \$15—the profit-maximizing price—it must do better than break even.

Consumers, however, lose in two ways: They pay more for the output they buy, and—due to higher prices—they buy less output. The changeover from perfect competition to monopoly thus benefits the owners of the monopoly and harms consumers of the product.

Keep in mind, though, an important proviso concerning this result: Comparing monopoly and perfect competition, we see that price is higher and output is lower under monopoly *if all else is equal*. In particular, we have assumed that after the market is monopolized, the technology of production remains unchanged at each previously competitive firm.

But a monopoly may be able to *change* the technology of production, so that all else would *not* remain equal. For example, a monopoly may have each of its new plants *specialize* in some part of the production process, or it may be able to achieve efficiencies in product planning, employee supervision, bookkeeping, or customer relations. If these cost savings enable the monopoly to use a less costly input mix for any given output level, then the monopoly's marginal cost curve in panel (c) would be *lower* than the competitive market supply curve in panel (a). If you add another, lower *MC* curve to panel (c), you'll see that this tends to *decrease* the monopoly's price and *increase* its output level—exactly the reverse of the effects discussed earlier. If the cost savings are great enough, and the *MC* curve drops low enough, a profit-maximizing monopoly could even charge a lower price and produce more output than would a competitive market. (See if you can draw a diagram to demonstrate this case.) The general conclusion we can draw is this:

The monopolization of a competitive industry leads to two opposing effects. First, for any given technology of production, monopolization leads to higher prices and lower output. Second, changes in the technology of production made possible under monopoly may lead to lower prices and higher output. The ultimate effect on price and quantity depends on the relative strengths of these two effects.

WHY MONOPOLIES OFTEN EARN ZERO ECONOMIC PROFIT

The title of this section might puzzle you. We've just seen that in the long run a monopoly can earn economic profit and should never stay in business if it suffers a loss. Then how can it be that monopolies often earn zero profit in the real world? Is it just a coincidence? The answer is no. There are two forces tending to cut monopoly profits.

1. *Government regulation.* As discussed earlier, in many cases of natural monopoly, a firm is granted a government franchise to be the sole seller in a market. This has been true of monopolies that provide water service, electricity, and natural gas. In exchange for its franchise, the monopoly must accept government regulation, often including the requirement that it submit its prices to a public commission for approval. The government will want to keep prices high enough to keep the monopoly in business, but no higher. Since the monopoly will stay in business unless it suffers a long-run loss, the ideal pricing strategy for the regulatory commission would be to keep the monopoly's economic profit at zero. Remember, though, that economic profit includes the opportunity cost of the funds invested by the monopoly's owners. If the public commission succeeds, the monopoly's *accounting* profit will be just enough to match what the owners could

earn by investing their funds elsewhere—that is, the monopoly will earn zero economic profit. Government regulation of monopoly will be discussed further in Chapter 15, on market failures.

2. *Rent-seeking activity.* Another factor that reduces a monopoly's profit comes from the interplay between politics and economics. As we've seen, many monopolies achieve and maintain their monopoly status due to government barriers to entry. Even when a monopoly is regulated by government, the regulation may be imperfect, resulting in a higher-than-ideal price. More importantly, many monopolies created through government barriers are completely unregulated. For example, a movie theater or miniature golf course may enjoy a monopoly in an area because zoning regulations prevent entry by competitors. Or, in less developed countries, a single firm may be granted the exclusive right to sell or produce a particular good. In all of these cases, the monopoly is left free to set its price as it wishes. When regulation is imperfect or when a monopoly is free of regulation, don't we expect it to earn economic profit for its owners?

Typically, no. Government barriers to entry—for example, zoning laws—are often controversial because, as you've learned, a monopoly may charge a higher price and produce less output than would a competitive market. Thus, government will be tempted to pull the plug on a monopoly's exclusive status and allow competitors into the market. The monopoly, in turn, will often take action to *preserve* government barriers to entry. Economists call such actions *rent-seeking activity*.

*Any costly action a firm undertakes to establish or maintain its monopoly status is called **rent-seeking activity**.*

Rent-seeking activity Any costly action a firm undertakes to establish or maintain its monopoly status.

In countries with corrupt bureaucracies, rent-seeking activity includes bribes to government officials; in less corrupt governments, it includes the time and money spent lobbying legislators and the public for favorable policies. For example, in 1999, AT&T acquired several cable companies, giving the firm a monopoly on cable television and cable Internet service in millions of homes across the United States. Except for one problem: City governments—before they would approve AT&T as the new operator of their cities' cable service—were demanding that AT&T permit competing Internet service providers, such as AOL and Bell Atlantic, to use their new cable lines. This, of course, would have cut into AT&T's monopoly profits in these cities. As we would predict, AT&T launched a war against these “open access” policies. It spent millions of dollars on lawyers, lobbyists, and public relations firms. It even tried to sway public opinion by helping to fund a lobbying group, “Hands Off the Internet.” In Miami for example, AT&T was able to convince 11 of 13 city council members to change their minds and vote against “open access.” But as you might guess, AT&T's expenses cut into its monopoly profit.⁵

What is the maximum amount of rent-seeking expenditure a monopoly would be willing to undertake? The answer, as you might guess, is an amount equal to the profit the firm is trying to protect. For example, if a firm can preserve \$100 million in profit through the passage of a pending bill, it would be willing to pay *up to* \$100 million in lobbying expenses. Of course, it may or may not be necessary to pay this much, depending on the nature of the bill and the difficulty in persuading legislators. But we can say this:

⁵ Source: *Wall Street Journal*, “ATT Used Carrot and Stick Lobbying Efforts in Local Debates over Access to Cable TV Lines,” November 24, 1999, p. A20.

Rent-seeking activity that helps establish or maintain a firm's monopoly position is part of the firm's costs. As a result, rent-seeking activity tends to reduce the economic profit of the firm and may even reduce it to zero.

What Happens When Things Change?



WHAT HAPPENS WHEN THINGS CHANGE?

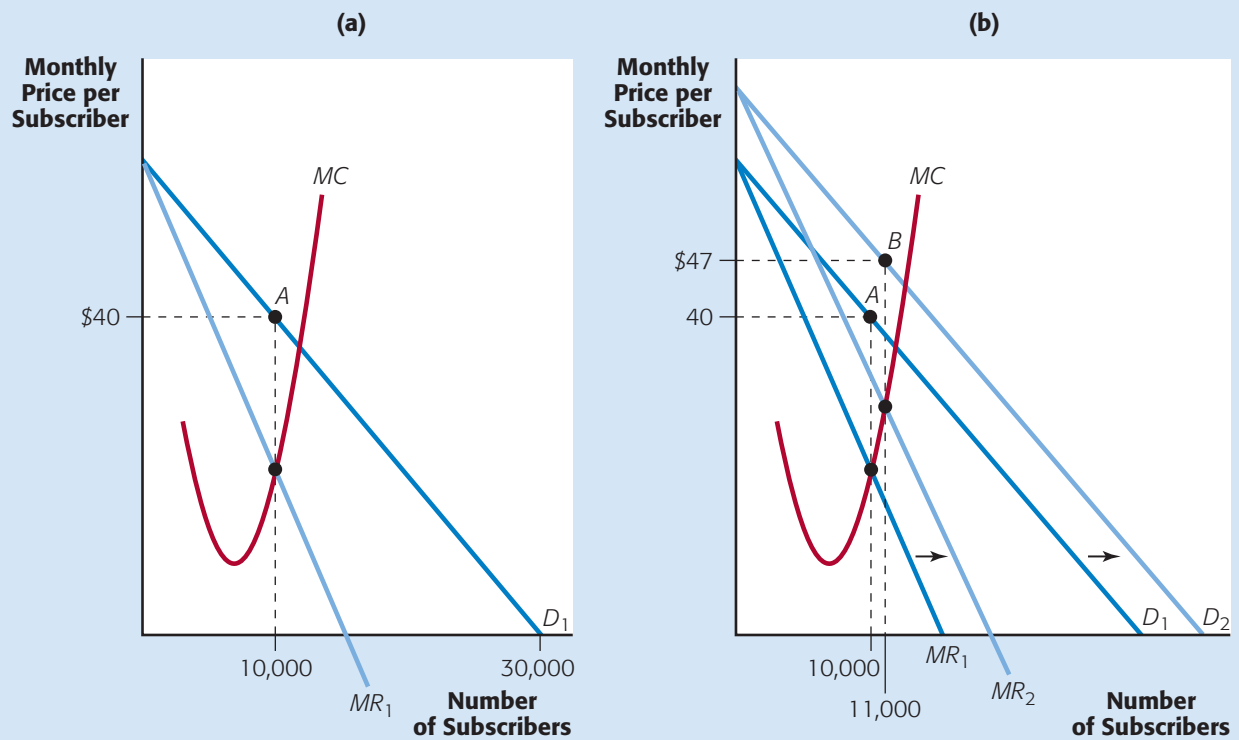
Once a monopoly is maximizing profit, it has no incentive to change either its quantity of output or its price . . . *unless* something that affects these decisions changes. In this section, we'll consider how a change in demand for the monopolist's product affects the equilibrium in a monopoly market.

Back in Chapter 8, we saw how a competitive market adjusted to a change in demand. In particular, we saw that an increase in demand caused an increase in both market price and market quantity. Does the same general conclusion hold for a monopolist? Let's see.

Panel (a) of Figure 5 shows Zillion-Channel Cable earning a positive profit in the short run. As before, it is producing 10,000 units per month, charging \$40 per unit,

FIGURE 5

A CHANGE IN DEMAND



Panel (a) shows Zillion-Channel in equilibrium. It is providing 10,000 units of cable TV service at a price of \$40 per month and earning a monthly profit of \$80,000. Panel (b) shows the same firm following an increase in demand from D_1 to D_2 . With the increased demand, MR is higher at each level of output. In the new equilibrium, Zillion-Channel is charging a higher price (\$47), providing more TV service (11,000 units), and earning a larger profit.

and earning a monthly profit of \$80,000 (not shown). The fact that Zillion-Channel is a monopolist, however, does not mean that it is immune to shifts in demand.

What might cause a monopolist to experience a shift in demand? The list of possible causes is the same as for perfect competition. If you need a reminder of these causes, look back at Figure 3 in Chapter 8. For example, an increase in consumer tastes for the monopolist's good will shift its demand curve rightward, just as it shifts the market demand curve rightward in a competitive market.

Suppose that the demand for local cable service increases because a sitcom shown on one of Zillion-Channel's premium services attracts an enthusiastic following—an increase in tastes for cable services. In panel (b) of Figure 5, this is shown by a rightward shift of the demand curve from D_1 to D_2 . Notice that the marginal revenue curve shifts as well—from MR_1 to MR_2 . With an unchanged cost structure, the new short-run equilibrium will occur where MR_2 intersects the unchanged MC curve. As you can see, the result is an increase in quantity from 10,000 to 11,000, and a higher price—\$47 per month rather than the original \$40. In this sense, monopoly markets behave very much like competitive markets (although the *extent* of the rise in price and quantity will generally *not* be the same as in a competitive market. What about the monopolist's profit, though? With both price and quantity now higher, total revenue has clearly increased. But cost is higher as well. So it seems as if profit could either rise or fall.

It turns out, however, that profit *must* be higher in the new equilibrium at point B . We know that because Zillion-Channel has the option of continuing to sell its original quantity, 10,000, at a price higher than before. If, as we assume, it started out earning a profit at that output level, then the higher price would certainly give it an even *higher* profit. But the logic of $MR = MC$ tells us that the greatest profit of all occurs at 11,000 units. We can conclude that:

A monopolist will react to an increase in demand by producing more output, charging a higher price, and earning a larger profit. It will react to a decrease in demand by reducing output, lowering price, and suffering a reduction in profit.

PRICE DISCRIMINATION

So far, we've analyzed the decisions of a **single-price monopoly**—one that charges the same price on every unit that it sells. But not all monopolies operate this way. For example, local utilities typically charge different rates per kilowatt-hour, depending on whether the energy is used in a home or business. Telephone companies charge different rates for calls made by people on different calling plans. Some Philadelphia customers' plans permit them to call Trenton, New Jersey, for free, while other customers pay 9 cents per minute for the same calls. Nor is this multi-price policy limited to monopolies: Movie theaters charge lower prices to senior citizens, airlines charge lower prices to those who book their flights in advance, and supermarkets and food companies charge lower prices to customers who clip coupons from their local newspaper.

In some cases, the different prices are due to differences in the firm's costs of production. For example, it may be more expensive to deliver a product a great distance from the factory, so a firm may charge a higher price to customers in outlying areas. But in other cases, the different prices arise not from cost differences, but from the firm's recognition that *some customers are willing to pay more than others*:

Single-price monopoly A monopoly firm that is limited to charging the same price for each unit of output sold.

Price discrimination Charging different prices to different customers for reasons other than differences in cost.

Price discrimination occurs when a firm charges different prices to different customers for reasons other than differences in costs.

The term *discrimination* in this context requires some getting used to. In everyday language, *discrimination* carries a negative connotation: We think immediately of discrimination against someone because of his or her race, sex, or age. But a price-discriminating monopoly does not discriminate based on prejudice, stereotypes, or ill will toward any person or group; rather, it divides its customers into different categories based on their *willingness to pay* for the good—nothing more and nothing less. By doing so, a monopoly can squeeze even more profit out of the market. Why, then, doesn't *every* firm practice price discrimination?

REQUIREMENTS FOR PRICE DISCRIMINATION

Although every firm would *like* to practice price discrimination, not all of them can. To successfully price discriminate, three conditions must be satisfied:

1. *There must be a downward-sloping demand curve for the firm's output.* In order to price discriminate, a firm must be able to raise its price to at least *some* customers without losing their business. A competitive firm cannot price discriminate: If it were to raise its price even slightly to some customers, they would simply buy the identical output from some other firm that is selling at the market price. This is one reason why there is no price discrimination in perfectly competitive markets like those for wheat, soybeans, and silver.

When a firm faces a downward-sloping demand curve, however, we know that some customers will continue to buy even when the price increases. Monopolies—which face downward-sloping demand curves—always satisfy the downward-sloping demand requirement.

2. *The firm must be able to identify consumers willing to pay more.* In order to determine which prices to charge to which customers, a firm must identify how much different customers are willing to pay. But this is often difficult. Suppose your barber or hairstylist wanted to price discriminate. How would he determine how much you are willing to pay for a haircut? He could *ask* you, but . . . let's be real: You wouldn't tell him the truth, since you know he would only use the information to charge you more than you've been paying. Price-discriminating firms—in most cases—must be a bit sneaky, relying on more indirect methods to gauge their customers' willingness to pay.

For example, airlines know that business travelers, who must get to their destination quickly, are willing to pay a higher price for air travel than are tourists or vacationers, who can more easily travel by train, bus, or car. Of course, if airlines merely *announced* a higher price for business travel, then no one would admit to being a business traveler when buying a ticket. So the airlines must find some way to identify business travelers without actually asking. Their method is crude but reasonably effective: Business travelers typically plan their trips at the last minute and don't stay over Saturday night, while tourists and vacationers generally plan long in advance and do stay over Saturday. Thus, the airlines give a discount to any customer who books a flight several weeks in advance and stays over, and they charge a higher price to those who book at the last minute and don't stay over. Of course, some business travelers may be able to do advance planning and pay the lower price, and some personal travelers who cannot plan

in advance might be priced out of the market. But on the whole, the airlines are able to charge a higher price to a group of people—business travelers—who are willing to pay more.⁶

Catalog retailers—such as Victoria’s Secret—have an easily available clue for determining who is willing to pay more: the customer’s address. People who live in high-income zip codes are mailed catalogues with higher prices than people who live in lower-income areas. And Internet retailers—such as CDnow—have another alternative: They use software to track customers’ past purchases and gauge whether each is a free spender or a careful shopper. Only the latter will get the low prices. (CDnow gives careful shoppers a secret Web site address with special discounts; all other customers pay full price.)⁷

3. *The firm must be able to prevent low-price customers from reselling to high-price customers.* Preventing a product from being resold by low-price customers can be a vexing problem for a would-be discriminator. For example, when airlines began price discriminating, a resale market developed: Business travelers could buy tickets at the last minute from intermediaries, who had booked in advance at the lower price and then advertised their tickets for sale. To counter this, the airlines imposed the additional requirement of a Saturday stayover in order to buy at the lower price. By adding this restriction, the airlines were able to substantially reduce the reselling of low-price tickets to business travelers.

It is easier to prevent resale of a *service* because of its personal nature. A hair-stylist can charge different prices to different customers without fearing that one customer will sell her haircut to another. The same is true of the services provided by physicians, attorneys, and music teachers.

Resale of *goods*, however, is much harder to prevent, since goods can be easily transferred from person to person without losing their usefulness. An interesting example of how far a company might have to go to prevent resale of a good is the case of Rohm and Haas, a chemical firm. In the 1940s, Rohm and Haas sold methyl methacrylate powder—used to make durable plastic—at two prices. Industrial users, who had many other options, paid 85 cents per pound; dental laboratories, which had no other choice of material for making dentures and were willing to pay more, were charged \$22 per pound. In spite of Rohm and Haas’s diligent efforts to prevent it, this price differential led to a flourishing resale market, in which industrial users were buying methyl methacrylate at 85 cents per pound and selling it for substantially more to dental laboratories. Internal memos at Rohm and Haas revealed that the company, desperate for a solution, considered (but did not finally follow) a plan to put lead or even arsenic (!) in all powder sold at the lower price so that dental laboratories would be unable to use it.⁸

⁶ It is sometimes argued that airlines’ pricing behavior is based entirely on a cost difference to the airline. For example, it is probably more costly for an airline to keep seats available until the last minute, because there is a risk that they will go unsold. The higher price for last-minute bookings would then compensate the airline for the unsold seats. (See, for example, the article by John R. Lott, Jr., and Russell D. Roberts in *Economic Inquiry*, January 1991.) But we know that cost differences are not the only reason for the price differential, or else the airlines would not have added the Saturday stayover requirement, which has nothing to do with their costs.

⁷ Woolley, Scott, “I Got it Cheaper than You,” *Forbes*, November 2, 1998.

⁸ From George W. Stocking and Myron W. Watkins, *Cartels in Action: Case Studies in International Business Diplomacy* (New York: The Twentieth Century Fund, 1946), p. 403.

EFFECTS OF PRICE DISCRIMINATION

Price discrimination always benefits the owners of a firm: When the firm can charge different prices to different consumers, it can use this ability to increase its profit. But the effects on consumers can vary. To understand how price discrimination affects the firm and the consumers of its product, let's take a simple example. Imagine that only one company—No-Choice Airlines—offers direct, small-plane flights between Omaha and Salina, Kansas. (What barrier to entry might explain No-Choice's monopoly on this route? If you're stumped, look again at the section on the sources of monopoly in this chapter.)

Figure 6(a) illustrates what No-Choice would do if it could *not* price discriminate and had to operate as a single-price monopoly. Since $MR = MC$ at 30 round-trip tickets per day, No-Choice's profit-maximizing price would be \$120 per ticket. The firm's average total cost for 30 round trips is \$80, so its profit per ticket would be $\$120 - \$80 = \$40$. Total profit is $\$40 \times 30 = \$1,200$, equal to the area of the shaded rectangle.

Price Discrimination That Harms Consumers. Now suppose that No-Choice discovers that on an average day, 10 of the 30 people buying tickets are business travelers who are willing to pay more, and it can identify them by their *unwillingness* to book in advance and stay over on Saturday night. No-Choice could price discriminate by offering two prices: \$120 for those who book in advance and stay over on Saturday, and \$160 to all others. In effect, No-Choice is raising the price from \$120 to \$160 for its 10 business customers.

Let's calculate the impact on No-Choice's profit. Since it continues to sell the same 30 round-trip tickets, there is no impact on its costs. Its revenue, however, will rise: It charges \$40 more than before on 10 of its round-trip tickets. Thus, No-Choice will earn an additional daily profit of $\$40 \times 10 = \400 . This *increase* in profit is identified as the shaded rectangle in Figure 6(b). Total profit is now the sum of two numbers: the profit No-Choice earned *before* price discrimination (\$1,200, the area of the shaded rectangle in panel (a)) and the *increase* in profit due to price discrimination (\$400, the area of the shaded rectangle in panel (b)). By price discriminating, No-Choice has raised its total profit from \$1,200 to \$1,600 per day.

What about consumers? Since 10 customers each pay \$40 more than before, they lose $10 \times \$40 = \400 from paying the higher price. Other travelers, who continue to pay \$120 for their tickets, are unaffected by the higher price.

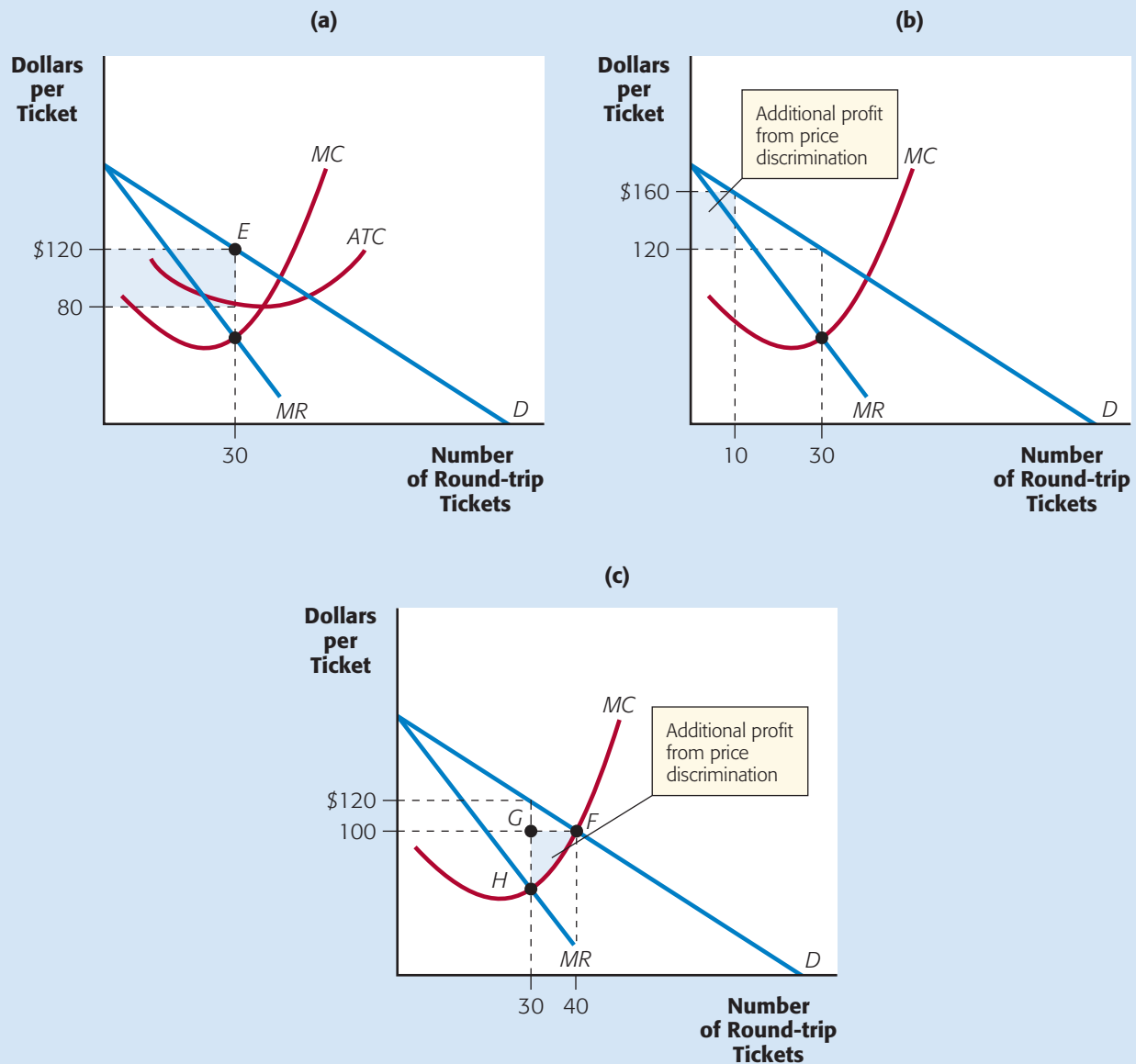
Summing up, in this case the impact of price discrimination—compared to a single-price policy—is a direct transfer of funds from consumers to the firm. The increase in the firm's profit is equal to the additional payments by consumers. This conclusion applies more generally as well:

When price discrimination raises the price for some consumers above the price they would pay under a single-price policy, it harms consumers. The additional profit for the firm is equal to the monetary loss of consumers.

Price Discrimination That Benefits Consumers. Let's go back to the initial situation facing No-Choice and suppose that, instead of charging a higher price to business travelers, it decides to price discriminate in a different way. No-Choice discovers that students who travel to college in Salina are going by train, because it is cheaper. However, at a price of \$100, the airline could sell an average of 10 round-

PRICE DISCRIMINATION

FIGURE 6



Panel (a) shows a single-price monopoly airline selling 30 round-trip tickets at \$120 each and earning a profit of \$1,200. Panel (b) shows the same airline if it can charge a higher price to its business travelers. The shaded rectangle shows the *additional* profit the airline earns by price discriminating; total profit is now \$1,600. Panel (c) shows an alternative strategy. In addition to selling 30 regular tickets at \$120 each, the airline attracts an additional 10 passengers at a lower student fare of \$100. So profit rises by the area of the shaded region.

trip tickets per day to the students. No-Choice's new policy is this: \$120 for a round-trip ticket, but a special price of \$100 for students who show their ID cards. The result is shown in panel (c). Although the decision to sell an additional 10 tickets pushes No-Choice beyond the output level at which $MC = MR$, this is no problem.

The MR curve was drawn under the assumption that No-Choice charges a single price and must lower the price on all tickets in order to sell more. But this is no longer the case. With price discrimination, the MR curve no longer tells us what will happen to No-Choice's revenue when output increases. As you are about to see, the firm will be able to increase its profit by selling the additional tickets.

The reasoning is as follows: No-Choice is now selling 10 *additional* round-trip tickets, so in this case both its cost and its revenue will change. Each additional ticket adds \$100 to the firm's revenue—this is the new marginal revenue. Each additional ticket adds an amount to costs given by the firm's MC curve. Thus, the distance between \$100 and the MC curve gives the *additional profit* earned on each additional ticket, and the total additional profit is the shaded area HGF in panel (c) of Figure 6.

What about consumers? The original 30 consumers are unaffected, since their ticket price has not changed. But the new customers—the 10 students—come out ahead: Each is able to take the flight rather than the longer train trip. In this case, price discrimination benefits the monopoly at the same time as it benefits a group of consumers—the students who were not buying the service before, but who *will* buy it at a lower price and gain some benefits by doing so. Since no one's price is raised, no one is harmed by this policy:

When price discrimination lowers the price for some consumers below what they would pay under a single-price policy, it benefits consumers as well as the firm.

Of course, it is possible for a firm to combine *both* types of price discrimination, raising the price above what it would charge as a single-price monopoly for some consumers and lowering it for others. This kind of price discrimination would increase the firm's profit, while benefiting some consumers and harming others. (For practice, draw a diagram showing the change in total profit if No-Choice were to charge three prices: a basic price of \$120, a price of \$160 for business travelers, and a price of \$100 for students. Who would gain and who would lose?)

Perfect Price Discrimination. Suppose a firm could somehow find out the maximum price customers would be willing to pay for *each* unit of output it sells. Then it could increase its profits even further by practicing *perfect price discrimination*:

Under perfect price discrimination, a firm charges each customer the most the customer would be willing to pay for each unit he or she buys.

Perfect price discrimination Charging each customer the most he or she would be willing to pay for each unit purchased.

Perfect price discrimination is very difficult to practice in the real world, since it would require the firm to read its customers' minds. However, many real-world situations come rather close to perfect price discrimination. Used car dealers routinely post a sticker price far higher than the price they think they can actually get and then size up each customer to determine the discount needed to complete the sale. The dealer may look at the customer's clothes and the car the customer is currently driving, inquire about the customer's job, and observe how sophisticated the customer is about cars, all with the aim of determining the maximum price he or she would be willing to pay. A similar sizing up takes place in flea markets, yard sales, and many other situations in which the final price is *negotiated* rather than fixed in advance.

To see how perfect price discrimination works, consider Nancy, who sells Elvis dolls at flea markets. To make our analysis simpler, we'll assume that Nancy has no

fixed costs of doing business and that each doll costs her \$10 to make, regardless of how many she produces. Thus, Nancy's cost per doll (ATC) is \$10 at every output level. Further, since each *additional* doll costs \$10 to make, her marginal cost (MC) is also \$10 at any output level. This is why, in Figure 7, both the MC and ATC curves are the same horizontal line at \$10.

Let's first suppose that Nancy is a single-price monopolist, charging a pre-announced price on every doll she sells. The figure shows the demand curve she would face on a typical day: At a price of \$30 she could sell 20 dolls, at a price of \$25 she could sell 30, and so on. Nancy would earn maximum profit by selling 30 dolls per day (why?) and charging \$25 each. Her profit per unit would be $\$25 - \$10 = \$15$ —the vertical distance between the ATC curve and the demand curve at 30 units. Her total profit would be $\$15 \times 30 \text{ dolls} = \450 per day, which is equal to the area of the shaded rectangle.

Now, suppose that Nancy becomes especially good at sizing up her customers. She learns how to distinguish true Elvis fanatics (a white, sequined jumpsuit is a dead giveaway) from people who merely want the doll as a gag gift. Moreover, by observing the way people handle the doll and listening to their conversations with their companions, Nancy can discern the exact maximum price each customer would pay. In effect, she knows exactly where on the demand curve each customer would be located. With her new skills, Nancy can increase her profit by becoming a *perfect price discriminator*—for each additional unit along the horizontal axis, she will charge the price indicated by the vertical height of the demand curve.

But how many dolls should Nancy sell now? To answer this question, we need to find the new output level at which $MR = MC$. But the MR curve in the figure is no longer valid: It was based on the assumption that Nancy had to lower the price on *all* units each time she wanted to sell another one. As a perfect price discriminator, she needs to lower the price only on the *additional* unit she sells, and her revenue will rise by the price of that additional unit. For example, if she is currently selling 30 dolls and wants to sell 31, she would lower the price on the additional doll just a tiny bit—say, to \$24.50—and in that case, her revenue would rise by \$24.50.

PERFECT PRICE DISCRIMINATION

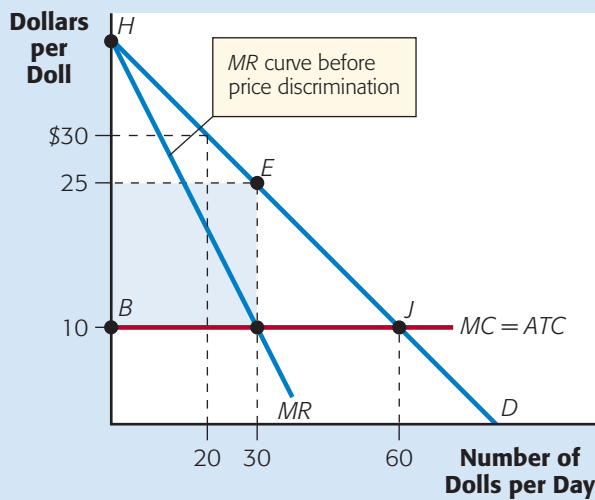


FIGURE 7

The single-price monopolist sells 30 dolls per day at \$25 each. With a constant ATC of \$10, she earns a profit of \$450 per day, as shown by the blue rectangle. However, if she can charge each customer the maximum the customer is willing to pay—shown by the height of the demand curve—she should sell 60 dolls, where $MC = P$ at point J . Her profit would increase to the area of triangle BHJ .

For a perfect price discriminator, marginal revenue is equal to the price of the additional unit sold. Thus, the firm's MR curve is the same as its demand curve.

Now it is easy to see what Nancy should do: Since our requirement for profit maximization is that $MC = MR$, and for a perfect price discriminator, MR is the same as price (P), Nancy should produce where $MC = P$. In Figure 7, this occurs at point J , where the MC curve intersects the demand curve—at 60 units of output. At that point, the only way to increase sales would be to lower the price on an additional doll below \$10, but since the marginal cost of a doll is always \$10, we would have $P < MC$, and Nancy's profit would decline.

(Think for a moment: What is Nancy's profit-maximizing price? Sorry, that's a trick question: There *is* no profit-maximizing price. As a perfect price discriminator, Nancy earns the highest profit by charging *different* prices to different customers.)

What about Nancy's total profit? On each unit of output, she charges a price given by the demand curve and bears a cost of \$10. Adding up the profit on *all* units gives us the area under the demand curve and above \$10, or the area of triangle BHJ .

Now we can determine who gains and who loses when Nancy transforms herself from a single-price monopolist to a perfect price discriminator. Nancy clearly gains: Her profit increases, from the shaded rectangle to the larger triangle BHJ . Consumers of the product are the clear losers: Since they all pay the most they would willingly pay, no one gets to buy a doll at a price he or she would regard as a "good deal."

A perfect price discriminator increases profit at the expense of consumers, charging each customer the most he or she would willingly pay for the product.

Interestingly, the E-commerce company Priceline.com provides a way for airlines to move closer to perfect price discrimination in the sale of last-minute tickets.

Priceline—which acts as a middleman between the airlines and the consumer—has found a new way to determine how much travelers are willing to pay for a ticket: *It asks them*. But how does it get an honest answer? In order to buy a ticket on Priceline.com, you must enter a bid. But—ingeniously—Priceline only allows you one bid each day. Thus, if you need to buy your tickets quickly, you have a strong incentive to bid closer to the true maximum you are willing to pay. The end result: Everyone who buys a ticket on Priceline.com pays a different price, and those who are *willing* to pay more generally end up *actually* paying more.

THE DECLINE OF MONOPOLY

The past century was not kind to monopolies. In the first half of the century, vigorous antitrust legislation and enforcement broke up many long-standing monopolies, such as Standard Oil in 1911, and Alcoa in 1945. For the rest of the century, many monopolies and would-be monopolies came under the scrutiny of government regulators, and were unable to fully maximize profit. Today, monopolies face a different threat: the relentless advance of technology.

Consider, for example, the natural monopoly of local phone service. The service—which currently takes place over local telephone wires—is characterized by economies of scale: A single company can produce at a lower cost per unit than

could several competitors. But soon, cable television companies will have the technology to offer local telephone service over *cable* wires. When this technology is put in place, every household will have two suppliers from which to choose: the existing local phone company *and* the local cable company. At this point, the monopoly status of local telephone companies will come to an end.

Even the old standard of monopolies—the post office—is being threatened by technology. Computerized inventory tracking and fuel-efficient jets have enabled companies such as Federal Express and DHL to offer low-cost overnight letter delivery services, while e-mail and bill paying by phone are cutting into the volume of old-fashioned letters. It is not hard to imagine a time in the future when you will receive all your mail on the Internet, and the notion of *hand-delivered* letters will become a thing of the past, a quaint practice you can tell your children about.

This is not to say that all monopolies are taking their last breaths. For one thing, some new technologies may be creating new barriers to entry, and may even be laying the groundwork for new monopolies. (We'll explore this further in Chapter 15). Other monopolies—like those created by patents and copyrights—will continue to make sense, because they are needed to reward those who bear the costs and risks of innovation. And some small-town monopolies—especially those that provide hands-on personal services such as medical care or haircuts—may remain immune to the technological threat. But it is safe to say that the world of monopoly, as we know it, is shrinking.

PRICE DISCRIMINATION AT COLLEGES AND UNIVERSITIES

Most colleges and universities give some kind of financial aid to a large proportion of their students. A typical aid package might include outright grants to help pay tuition and room and board, a low-interest loan, and a work-study job on campus. Colleges have many motives for this policy, such as having a more diverse student body and helping to create a better society by making educational services accessible to many who might not otherwise afford them. But increasingly, financial aid has been used as an effective method of price discrimination, designed to increase the revenue of the college.

How does a college price discriminate? By offering different levels of assistance to different students, financial aid permits the college to charge different *prices* to each one. For example, if full tuition is \$12,000 per year, then a student who receives a yearly \$5,000 grant pays only \$7,000 per year, a student who receives an \$8,000 grant pays only \$4,000 per year, and so on.

Colleges have long been in an especially good position to benefit from price discrimination, because they satisfy all three requirements:

1. *Colleges face downward-sloping demand curves.* Although colleges are not monopolies (other, similar institutions are close substitutes), they are not perfect competitors either. Each college is unique in some ways—location, reputation, living conditions, social life, and more. For this reason, colleges face downward-sloping demand curves for their services. A college can raise its price and lose only *some*—rather than all—of its enrollment applicants. Similarly, any college that wants to increase enrollment can do so by lowering its price and attracting applications from those who would not attend at the higher price.

2. *Colleges are able to identify consumers willing to pay more.* Colleges have long been in an excellent position to discover how much their customers

Using the
THEORY



would be willing to pay for their product. Applicants for financial aid have had to submit data on their families' income and wealth. Admissions officials know that students from poor families are less likely to attend their institutions, unless they are offered a relatively low price, while students from wealthier families are more likely to attend even at higher prices. In recent years, however, colleges have gone even further in their attempts to identify willingness to pay. (See below.)

3. *Colleges are able to prevent low-price customers from reselling to high-price customers.* A college education is much like other personal services: Once you pay for it, you cannot resell it to another person.

While most colleges have been active price discriminators for decades, many have stepped up their efforts since 1992. That year, Congress changed the formula used to determine financial need, making most students eligible for assistance. This allowed colleges to allocate financial aid dollars among a wider pool of students. Suddenly, price discrimination could be used even more extensively. But this required new methods of identifying willingness to pay among different students.

The market responded. Specialized consultants, using computer models to predict the likelihood that students would attend college at different prices, began offering their services. One consultant's pamphlet asked admissions officials, "Did you overspend to get students who would have matriculated with lesser aid? Did you underspend and lose students who would have come with more support?"⁹

As a result, the traditional role of financial aid as assistance for those in need has changed. Some colleges have shifted aid dollars toward top-ranked applicants—regardless of financial need—because those students have more options and are less likely to attend any college without financial aid. Drexel University in Philadelphia shifted aid toward those who applied as business majors after a computer model predicted that these students' enrollment decisions were more sensitive to price. Johns Hopkins University shifted aid dollars to humanities majors with SAT scores above 1,200 and relatively low financial need, based on a similar model. Carnegie-Mellon has gone even further. It determines the effect on student "yield" (the percentage of students with certain characteristics who will actually enroll) of shifting aid dollars from one group to another. In addition, the school asks students who have been admitted to fax the school any better financial aid offers they might receive, so it can decide whether to match the offer from a special "reaction fund."

Many financial aid consultants have even recommended that colleges shift aid money away from students who come for on-campus interviews, since by doing so, those students reveal a strong desire to attend college. There are rumors that some institutions have begun following this advice, but no college has admitted to the practice.

More effective price discrimination at colleges and universities is certainly changing the traditional view of financial assistance as a program designed primarily to help those in need. And while it has benefited some groups of students, it has harmed others. Under the newer systems, those who can signal a lower willingness to pay have benefited from reduced prices, while those signaling greater willingness to pay have suffered a price increase.

But fully assessing the effects of price discrimination at colleges is complicated by one important fact: Most educational institutions are not private firms striving to maximize profits for their owners. Rather, they are *nonprofit* institutions, *with-*

⁹ "Colleges Manipulate Financial-Aid Offers, Shortchanging Many," *Wall Street Journal*, April 1, 1996, p. 1. The specific examples of price discrimination in the discussion also come from this article.

out private owners. Thus, any additional revenue they gain through price discrimination is likely to be used for educational purposes: to attract better faculty by raising salaries, to improve living conditions for students, to keep tuition lower than it otherwise would be, and even to provide increased aid for more students in the future. Each of these alternatives has value to the college and its students, suggesting that increased price discrimination at colleges, like so many other economic issues, is a matter of tradeoffs.

S U M M A R Y

A *monopoly firm* is the only seller of a good or service with no close substitutes. Monopoly arises because of some barrier to entry: economies of scale, control of a scarce input, or a government-created barrier. As the only seller, the monopoly faces the market demand curve and must decide what price (or prices) to charge in order to maximize profit.

Like other firms, a single-price monopolist will produce where $MR = MC$ and set that maximum price consumers are willing to pay for that quantity. Monopoly profit ($P - ATC$ multiplied by the quantity produced) can persist in the long

run because of barriers to entry. However, there are reasons why monopolies often earn zero long-run profit. These reasons include government regulation and rent seeking.

Some monopolies can practice *price discrimination* by charging different prices to different customers. Doing so requires the ability to identify customers who are willing to pay more and to prevent low-price customers from reselling to high-price customers. Price discrimination always benefits the monopolist (otherwise, it would change a single price), but it *may* sometimes benefit some consumers.

K E Y T E R M S

monopoly firm
monopoly market
natural monopoly

patent
copyright
government franchise

rent-seeking activity
single-price monopoly
price discrimination

perfect price discrimination

R E V I E W Q U E S T I O N S

- Why is it sometimes difficult to decide whether a particular firm is a monopoly? Which U.S. markets are often considered to exemplify monopoly?
- Why do monopolies arise? Discuss the most common factors that explain the existence of a monopoly.
- How can the government create a monopoly? Why might the government want to do this?
- Drunk with power, the CEO of Monolith, Inc., a single-price monopoly, assumes that he can set any price he wants and sell as many units as he wants at that price. Is he correct? Why or why not?
- True or False? "A firm's marginal cost curve is always its supply curve." Explain.
- Why might the decision to shut down be different for a monopoly than for a perfectly competitive firm?
- Why might a monopoly earn an economic profit in the long run? How does this differ from the situation faced by a perfectly competitive firm?
- Explain why, if a monopoly takes over all the firms in a perfectly competitive industry, its marginal cost curve will be the same as the perfectly competitive industry's supply curve.
- Firm A maximizes profit at an output of 1,000 units, where $Price = 50$ and $MC = 50$. Firm B maximizes profit at an output of 2,000 units, where $Price = 5$ and $MC = 3$. Which firm is likely to be a monopoly and which perfectly competitive? Explain your reasoning.
- How do output and price for a monopoly compare with output and price if the same market were perfectly competitive?
- In the long run, a monopoly can earn positive economic profit; in the real world, monopolies often don't. Explain this apparent paradox.

12. Explain the difference between a single-price monopoly and a price-discriminating monopoly. What conditions must be present in order for a monopoly to price discriminate? Explain why each condition is necessary.
13. True or False? “Price discrimination by a monopoly always harms consumers.” Explain.

P R O B L E M S A N D E X E R C I S E S

- Draw the demand curve for a perfectly competitive firm and for a monopoly, showing the *MR* curve, as well as the demand curve on each graph.
 - In each case, what is the relationship between demand in the market as a whole and demand for an individual firm’s output?
 - For both graphs, explain the position of the *MR* curve in relation to the demand curve.
- In a certain large city, hot dog vendors are *perfectly competitive*, and face a market price of \$1.00 per hot dog. Each hot dog vendor has the following total cost schedule:

Number of Hot Dogs per Day	Total Cost
0	\$63
25	73
50	78
75	88
100	103
125	125
150	153
175	188
200	233

- Add a *marginal cost* column to the right of the total cost column. (*Hint*: Don’t forget to divide by the *change* in quantity when calculating *MC*.)
- What is the profit-maximizing quantity of hot dogs for the typical vendor, and what profit (loss) will he earn (suffer)? Give your answer to the nearest 25 hot dogs.

One day, Zeke, a typical vendor, figures out that if he were the only seller in town, he would no longer have to sell his hot dogs at the market price of \$1.00. Instead, he’d face the following demand schedule:

Price per Hot Dog	Number of Hot Dogs per Day
> \$6.00	0
6.00	25
5.50	50
4.00	75
3.25	100
2.75	125
2.25	150
1.75	175
1.25	200

- Add *total revenue* and *marginal revenue* columns to the table above. (*Hint*: Once again, don’t forget to divide by the *change* in quantity when calculating *MR*.)
 - As a monopolist with the cost schedule given in the first table, how many hot dogs would Zeke choose to sell each day? What price would he charge?
 - A lobbyist has approached Zeke, proposing to form a new organization called “Citizens to Eliminate Chaos in Hot Dog Sales.” The organization will lobby the city council to grant Zeke the only hot dog license in town, and it is guaranteed to succeed. The only problem is, the lobbyist is asking for a payment that amounts to \$200 per business day as long as Zeke stays in business. On purely economic grounds, should Zeke go for it? (*Hint*: If you’re stumped, re-read the section on rent-seeking activity.)
- Draw demand, *MR*, and *ATC* curves that show a monopoly that is just breaking even.
 - Below is demand and cost information for Warmfuzzy Press, which holds the copyright on the new best-seller, *Burping Your Inner Child*.

<i>Q</i> (No. of Copies)	<i>P</i> (per Book)	<i>ATC</i> (per Book)
100,000	\$100	\$20
200,000	\$ 80	\$15
300,000	\$ 60	\$16 ² / ₃
400,000	\$ 40	\$22 ¹ / ₂
500,000	\$ 20	\$31

- Determine what quantity of the book Warmfuzzy should print, and what price it should charge in order to maximize profit.
 - What is Warmfuzzy’s maximum profit?
 - Prior to publication, the book’s author renegotiates his contract with Warmfuzzy. He will receive a great big hug from the CEO, along with a one-time bonus of \$1,000,000, payable when the book is published. This payment was not part of Warmfuzzy’s original cost calculations.
How many copies should Warmfuzzy publish now? Explain your reasoning.
- Draw the *MR* and demand curves for a perfect price discriminator. How does the *MR* curve for a perfect price discriminator differ from that for a single-price monopoly?

- Look at Figure 6(c). Clearly, $MR = MC$ at point H . But when the airline sells discount tickets to college students, it is at point F , apparently violating the rule that $MR = MC$. Does this mean that for a price-discriminating monopoly, $MR = MC$ doesn't hold? Explain.
- Suppose that price discrimination were made illegal—across the board. Who would benefit and who would be harmed? Choose an example with which you are familiar and try to determine both the short-run and long-run effects of banning price discrimination in that case.

CHALLENGE QUESTIONS

- Are there any circumstances under which a monopoly will sell at the same price as would a perfectly competitive firm selling the same product? Explain.
- Let a single-price monopoly's demand curve be given by $P = 20 - 4Q$, where P is price and Q is quantity demanded. Marginal revenue is $MR = 20 - 8Q$. Marginal cost is $MC = Q^2$. How much should this firm produce in order to maximize profit?
- In the short run, a monopoly uses both fixed and variable inputs to produce its output. Draw a diagram to illustrate why, if the price of using a fixed input rises, there will be no change in the monopoly's equilibrium price or quantity. (*Hint*: Which curves shift if the price of a fixed input rises? Which curves remain unaffected?)

EXPERIENTIAL EXERCISES

- Here is a challenge. Use either Infotrac or the *Wall Street Journal* to find a case of a monopoly suffering a loss. When you find one, try to determine what kinds of changes caused the monopoly's profits to evaporate. Then model those changes using a standard monopoly diagram.
- In many large U.S. cities, monopoly owners of sports franchises have been lobbying local governments for new, publicly financed stadiums and arenas. Is this a form of rent seeking? For some background evidence, check the Anti-Stadium site at http://www.resonator.com/stad/s_issues.htm. Is there convincing evidence of rent seeking?





CHAPTER

10

MONOPOLISTIC COMPETITION AND OLIGOPOLY

CHAPTER OUTLINE

The Concept of Imperfect Competition

Monopolistic Competition

Monopolistic Competition in the Short Run

Monopolistic Competition in the Long Run

Excess Capacity Under Monopolistic Competition

Nonprice Competition

Oligopoly

Oligopoly in the Real World

Why Oligopolies Exist

Oligopoly Behavior

Cooperative Behavior in Oligopoly

The Limits to Oligopoly

Using the Theory: Advertising in Monopolistic Competition and Oligopoly

Advertising and Market Equilibrium Under Monopolistic Competition

Advertising and Collusion in Oligopoly

The Four Market Structures: A Postscript

On any given day, you are probably exposed to hundreds of advertisements. The morning newspaper announces special sales on clothes, computers, and paper towels. On the way to class, you might see numerous billboards competing for your attention, suggesting that you stay at the Holiday Inn, eat at Burger King, or organize your life with a Palm Pilot. You will likely spend more time watching advertisements for breakfast cereals on television than you will spend eating them. And as you search for information on the World Wide Web, ads for home shopping services, “webzines,” and Internet access providers flash before your eyes. No doubt about it: Advertising is everywhere in the economy.

Yet, so far in this book, not much has been said about advertising. There is a good reason for this: In the two market structures we have studied so far—perfect competition and monopoly—firms do little, if any, advertising. Indeed, perfectly competitive firms *never* advertise, since there is no point to it. Each firm in a competitive industry produces the same product as any other, so what would they advertise? And in any case, each firm can sell all it wants at the market price, so advertising would only raise costs without any benefit to the firm. Monopolists *sometimes* advertise, but—as the only seller of a good with no close substitutes—they are under no pressure to do so.

Where, then, is all the advertising coming from? To answer this question, we must look beyond the market structures we’ve studied so far and consider firms that are neither perfect competitors nor monopolists. That is what we will do in this chapter. While advertising is one interesting feature we will explore, there are many others as well.

THE CONCEPT OF IMPERFECT COMPETITION

In perfect competition, there are so many firms selling the same product that none of them can affect the market price. In monopoly, there is just *one* seller in the market, so it sets the price as it wishes. Most markets for goods and services, however, are neither perfectly competitive nor perfectly monopolistic. Instead, they lie some-

where *between* these two extremes, with more than one firm, but not enough firms to qualify for perfect competition. We call such markets *imperfectly competitive*:

Imperfect competition refers to market structures between perfect competition and monopoly. In imperfectly competitive markets, there is more than one seller, but too few to create a perfectly competitive market. In addition, imperfectly competitive markets often violate other conditions of perfect competition, such as the requirement of a standardized product or free entry and exit.

Consider the market for automobiles in the United States. It is certainly not a monopoly, since more than a dozen companies sell cars here: General Motors, Ford, DaimlerChrysler, Mazda, Toyota, Honda, Volvo, Nissan, and several more. But neither is this market perfectly competitive: Each of these firms supplies a relatively large part of the market, so each can affect the market price. Moreover, the product of each firm is different from the products of the others: A Toyota is not a Ford, and a Ford is not a Jeep. The market for automobiles, then, falls somewhere between the extremes of monopoly and perfect competition.

Or consider restaurants. Even a modest-size city such as Cincinnati has more than 3,000 different restaurants. This is certainly a large number of competitors, but they are not *perfect* competitors, since each one sells a product that is differentiated in important ways—in the type of food served, the recipes used, the atmosphere, the location, and even the friendliness of the staff.

In this chapter, we study two types of imperfectly competitive markets: *monopolistic competition* and *oligopoly*.

MONOPOLISTIC COMPETITION

What do hotels, food markets, and exterminators have in common? All three sell their products under conditions of **monopolistic competition**.

A monopolistically competitive market has three fundamental characteristics:

1. many buyers and sellers;
2. no significant barriers to entry or exit; and
3. differentiated products.

Note that monopolistic competition combines some features of both pure competition and monopoly—hence its name. Like perfect competition, there are many buyers and sellers and easy entry and exit. Restaurants, photocopy shops, dry cleaners, and virtually all retail stores, such as clothing stores or food markets, are almost always monopolistic competitors. In each case, there are many sellers in the market, and it is easy to set up a business or to exit if things don't go well. But unlike perfect competitors, each seller produces a somewhat different product from the others. No two coffeehouses, photocopy shops, or food markets are exactly the same. For this reason, a monopolistic competitor can raise its price (up to a point) and lose only *some* of its customers. The others will stay with the firm because they like its product, even when it charges somewhat more than its competitors. Thus, a monopolistic competitor faces a *downward-sloping demand curve* and, in this sense, is more like a monopolist than a perfect competitor:



Characterize the Market

Monopolistic competition A market structure in which there are many firms selling products that are differentiated, yet are still close substitutes, and in which there is free entry and exit.

Because it produces a differentiated product, a monopolistic competitor faces a downward-sloping demand curve: When it raises its price a modest amount, quantity demanded will decline (but not all the way to zero).

What makes a product differentiated? Sometimes, it is the *quality* of the product. By many objective standards—longevity, performance, frequency of repair—a Toyota is a better car than a Volkswagen. Similarly, based on room size and service, the Hilton has better hotel rooms than Motel 6. In other cases, the difference is a matter of taste rather than quality. In terms of measurable characteristics, Colgate toothpaste is probably neither better nor worse than Crest, but each has its own flavor and texture, and each appeals to different people.

Another type of differentiation arises from differences in *location*. Two bookstores may be identical in every respect—range of selection, atmosphere, service—but you will prefer the one closer to your home or office.

Ultimately, though, product differentiation is a subjective matter: A product is different whenever people *think* that it is, whether their perception is accurate or not. You may know, for example, that all bottles of bleach have identical ingredients—5.25 percent sodium hypochlorite and 94.75 percent water. But if *some* buyers think that Clorox bleach is different and would pay a bit more for it, then Clorox bleach is a differentiated product. Thus, whenever a firm that is not a monopolist faces a downward-sloping demand curve, we can assume that it produces a differentiated product. The *reason* for the downward slope may be a difference in product quality, consumer tastes, or location, or it may be entirely illusory, but the economic implications are always the same: The firm can raise its price without losing all of its business.

MONOPOLISTIC COMPETITION IN THE SHORT RUN

Identify Goals and Constraints



The individual monopolistic competitor behaves very much like a monopoly. Its constraints are its given technology of production, the prices it must pay for its inputs, and the downward-sloping demand curve that it faces. And, like any other firm, its goal is to maximize profit by producing where $MR = MC$. The result may be economic profit or loss in the short run.

The key difference is this: While a monopoly is the *only* seller in its market, a monopolistic competitor is one of many sellers. When a *monopoly* raises its price, its customers must pay up or consume less of the good. When a *monopolistic competitor* raises its price, its customers have one additional option: They can buy a similar good from some other firm. Thus, all else equal, the demand curve facing a firm should be flatter under monopolistic competition than under monopoly. That is, since closer substitutes are available under monopolistic competition than under monopoly, a given rise in price should cause a greater fall in quantity demanded.

Figure 1 illustrates the situation of a monopolistic competitor—Kafka Exterminators. The figure shows the demand curve— d_1 —that the firm faces, as well as the marginal revenue, marginal cost, and average total cost curves. As a monopolistic competitor, Kafka Exterminators competes with many other extermination services in its local area. Thus, if it raises its price, it will lose some of its customers to the competition. If Kafka had a *monopoly* on the local extermination business, we would expect the same rise in price to cause a smaller drop in quantity demanded, since customers would have to buy from Kafka or else get rid of their bugs on their own.

Like any other firm, Kafka Exterminators will produce where $MR = MC$. As you can see in Figure 1, when Kafka faces demand curve d_1 and the associated

A MONOPOLISTICALLY COMPETITIVE FIRM IN THE SHORT RUN

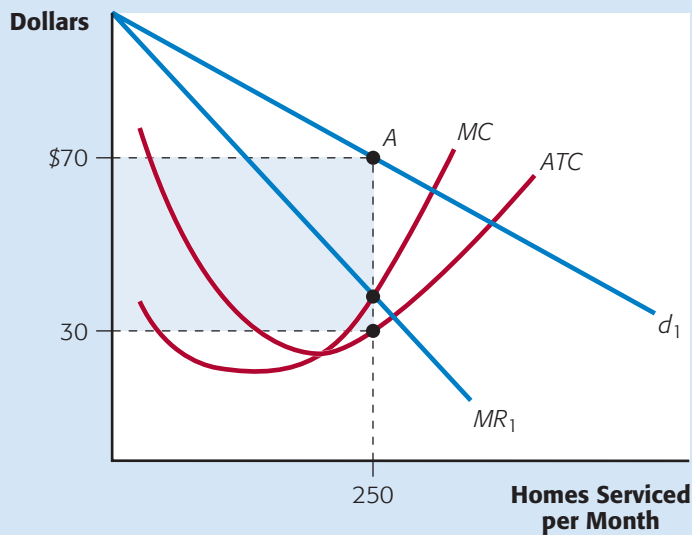


FIGURE 1

Kafka Exterminators, a monopolistic competitor, faces downward-sloping demand curve d_1 and marginal revenue curve MR_1 . It services 250 homes per month (where $MR = MC$), charges \$70 per home, and earns a short-run profit of \$10,000, shown by the blue rectangle.

marginal revenue curve MR_1 , its profit-maximizing output level is 250 homes served per month, and its profit-maximizing price is \$70 per home. In the short run, the firm may earn an economic profit or an economic loss, or it may break even. In the figure, Kafka is earning an economic profit: Profit per unit is $P - ATC = \$70 - \$30 = \$40$, and total monthly profit—the area of the blue rectangle—is $\$40 \times 250 = \$10,000$.



Find the Equilibrium

MONOPOLISTIC COMPETITION IN THE LONG RUN

If Kafka Exterminators were a monopoly, Figure 1 might be the end of our story. The firm would continue to earn economic profit forever, since barriers to entry would keep out any potential competitors. But under monopolistic competition—in which there are no barriers to entry and exit—the firm will not enjoy its profit for long. As new sellers enter the market, attracted by the profits that can be earned there, some of Kafka's customers will sign up with the new entrants. At any given price, Kafka will find itself servicing fewer homes than before, and the demand curve it faces will shift leftward. Entry will continue to occur, and the demand curve will continue to shift leftward, until Kafka and other firms are earning zero economic profit.¹ This is shown in Figure 2. Zero profit requires that the profit-maximizing price—\$40—be equal to the average total cost of production.

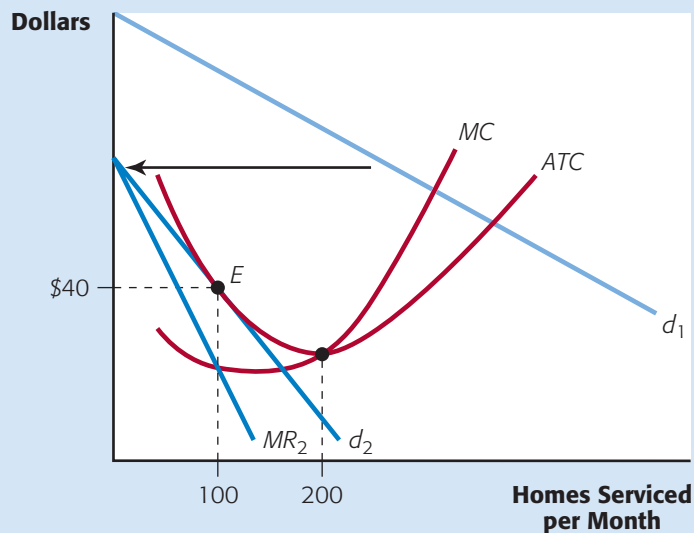
Notice that the new demand curve, d_2 , lies to the left of the original demand curve d_1 from Figure 1. The slope of the demand curve may change as well. Since the demand curve has shifted leftward, so has the MR curve, to MR_2 . Producing one more unit of output will add less to total revenue than it did before the shift,

¹ Other things may also happen as the industry expands. For example, the increased demand for inputs may raise or lower the typical firm's ATC and MC curves, depending on whether we are dealing with an increasing- or decreasing-cost industry. (See Chapter 8.) This does not change our result, however: Entry into the market will continue until the typical firm earns zero economic profit, even if its MC and ATC curves have shifted.

FIGURE 2

If existing firms are earning short-run profits, entry will occur. The demand curve facing Kafka Exterminators will shift left. Long-run equilibrium occurs at point E , the quantity at which the new marginal revenue curve, MR_2 , intersects MC . The price of \$40 equals ATC , so the firm earns zero economic profit.

A MONOPOLISTICALLY COMPETITIVE FIRM IN THE LONG RUN



Find the Equilibrium



since the price consumers are willing to pay at each output level is now lower. The new profit-maximizing output level—where $MR = MC$ —is 100 units, and the profit-maximizing price is \$40 per unit. Since ATC is also \$40, the firm is earning zero economic profit.

We can also reverse these steps. If the typical firm is suffering an economic loss (draw this diagram on your own), *exit* will occur. With fewer competitors, those firms that remain in the market will gain customers, so their demand curves will shift *rightward*. Exit will cease only when the typical firm is earning zero economic profit—as in Figure 2. Thus, Figure 2 represents the long-run equilibrium of the typical firm whether we start from a position of economic profit or economic loss:

Under monopolistic competition, firms can earn positive or negative economic profit in the short run. But in the long run, free entry and exit will ensure that each firm earns zero economic profit, just as under perfect competition.

Is this prediction of our model realistic? Indeed it is: In the real world, monopolistic competitors often earn economic profit or loss in the short run, but—given enough time—profits attract new entrants, and losses result in an industry shake-out, until firms are earning zero economic profit. In the long run, restaurants, retail stores, hair salons, and other monopolistically competitive firms earn zero economic profit for their owners. That is, there is just enough accounting profit to cover the implicit costs of doing business, which is just enough to keep the owners from shifting their time and money to some alternative enterprise.

Think of your own city or town. Has a certain kind of business been springing up everywhere? Is another type of business gradually disappearing? If you look around, you will see entry and exit occurring right before your eyes, as monopolistically competitive markets adjust from short-run to long-run equilibrium.

EXCESS CAPACITY UNDER MONOPOLISTIC COMPETITION

Take another look at Figure 2. When the typical firm earns zero economic profit, its demand curve touches—but *does not cross*—its *ATC* curve. This will always be the case in the long run. To see why, draw a diagram right now (you can use the margin of this page) that shows the demand curve actually *crossing* the *ATC* curve. If you do this correctly, you will find that at *some* output levels, price is greater than *ATC*, and the firm can earn economic profit by producing there. But such profit attracts entry, so we are not yet in long-run equilibrium.

Notice, too, that at point *E*, where the two curves touch, the *ATC* curve has the same slope as the demand curve—a *negative* slope. Thus, in the long run, a monopolistic competitor always produces on the *downward-sloping* portion of its *ATC* curve and therefore *never produces at minimum average cost*. Indeed, its output level is always *too small* to minimize cost per unit. The firm operates with *excess capacity*. (The output level at which cost per unit is minimized is often called capacity output.) In Figure 2, Kafka Exterminators would reach minimum cost per unit by servicing 200 homes per month—the firm’s capacity output—but in the long run, it will service only 100 homes per month.

In the long run, a monopolistic competitor will operate with excess capacity—that is, it will produce too little output to achieve minimum cost per unit.

To see why a monopolistic competitor *cannot* minimize average cost in the long run, imagine that Kafka Exterminators wanted to do so, by servicing 200 homes per month. With its current demand curve, it would suffer a loss, since $P < ATC$ at that output level. But there might be another way for Kafka: Perhaps it can buy out one of its competitors and take over its business. Then Kafka’s demand and marginal revenue curves would lie farther to the right (you may want to draw this), and the company—by producing more—might be able to achieve minimum (or at least lower) per-unit costs. But while this might work in the short run, it cannot work in the long run. With its new demand curve and its lower cost per unit, Kafka would earn a profit. In the long run, profit would attract entry, and Kafka would be back where it started.

This example gives us another way to view excess capacity: In the long run, there are *too many* firms under monopolistic competition, each one producing too little output, to achieve minimum cost per unit. If there were fewer firms, then each could reduce its *ATC*, but the situation would not last. Each firm would earn a profit, profit would cause entry, entry would force each firm to reduce its output, and in the long run, there would once again be too many firms producing too little output to minimize average cost.

Excess capacity is easy to observe. Think of the suburban shopping mall with a dozen or more clothing stores. Much of the time, one or more of the stores has no customers. Or think of all the restaurants in your town. How often are they all fully occupied? In each case, serving more customers would bring down cost per unit, but this would require fewer firms, which—as we know—there cannot be in the long run.

Excess capacity suggests that monopolistic competition is costly to consumers, and indeed it is. Recall that under perfect competition, $P = \text{minimum } ATC$ in the long run. (Look back at Figure 9 in Chapter 8, on p. 235.) Under monopolistic competition, we have $P > \text{minimum } ATC$ in the long run. Thus, if the *ATC* curves were the same, price would always be greater under monopolistic competition.



For a fascinating analysis of trends in product differentiation see, The Federal Reserve Bank of Dallas's 1998 Annual Report, "The Right Stuff: America's Move to Mass Customization." You can find it at the Bank's Web site (<http://www.dallasfed.org>).

Nonprice competition Any action a firm takes to increase the demand for its product, other than cutting its price.

This reasoning may tempt you to leap to a conclusion: Consumers are better off under perfect competition. But don't leap so fast: Remember that in order to get the beneficial results of perfect competition, all firms must produce identical output. It is precisely because monopolistic competitors produce *differentiated* output—and therefore have downward-sloping demand curves—that $P >$ minimum *ATC* in the long run. And consumers usually *benefit* from product differentiation. (If you don't think so, imagine how you would feel if every restaurant in your town served an identical menu, or if everyone had to wear the same type of clothing, or if every rock group in the country performed the same tunes in exactly the same way.) Seen in this light, we can regard the higher costs and prices under monopolistic competition as the price we pay for product variety. Some may argue that there is too much variety in a market economy—how many different brands of toothpaste do we really need?—but few would want to transform all monopolistically competitive industries into perfectly competitive ones.

NONPRICE COMPETITION

If a monopolistic competitor wants to increase its output, one way is to cut its price—that is, it can move *along* its demand curve. But a price cut is not the only way to increase output. Since the firm produces a differentiated product, it can sell more by convincing people that its own output is better than that of other firms. Such efforts, if successful, will *shift* the firm's demand curve rightward.

Any action a firm takes to increase the demand for its output—other than cutting its price—is called nonprice competition.

Better service, product guarantees, free home delivery, more attractive packaging—as well as advertising to inform customers about these things—are all examples of nonprice competition. Fast-food restaurants are notorious for nonprice competition. When Burger King says, "Have it your way," the company is saying, "Our hamburgers are better than those at McDonald's because *we* make them to order." When McDonald's responds with an attractive, fresh-faced young woman behind the counter, smiling broadly when you order a Happy Meal, it is saying, in effect, "So what if we don't make your burgers to order; our staff is better looking and more upbeat than Burger King's."

Nonprice competition is another reason why monopolistic competitors earn zero economic profit in the long run. If an innovative firm discovers a way to shift its demand curve rightward—say, by offering better service or more clever advertising—then in the short run, it may be able to earn a profit. This means that other, less innovative firms will experience a leftward shift in *their* demand curves, as they lose sales to their more innovative rival.

But not for long. Eventually, *all* firms will imitate the actions of the most successful among them. If product guarantees are enabling some firms to earn economic profit, then *all* firms will offer product guarantees. If advertising is doing the trick, then *all* firms will start ad campaigns. In the long run, we can expect *all* monopolistic competitors to run advertisements, to be concerned about service, and to take whatever actions have proven profitable for other firms in the industry. All this nonprice competition is costly—one must *pay* for advertising, for product guarantees, for better staff training—and these costs must be included in each firm's *ATC* curve, shifting it upward. But this does not change any of our conclusions about monopolistic competition in the long run.

Indeed, nonprice competition strengthens our conclusions. In the short run, a firm may earn profit because it has relatively few competitors or because it has discovered a new way to attract customers. But in the long run, the profitable firm will find its demand curve shifting leftward due to the entry of new firms, or the imitation of its successful nonprice competition, or both. In the end, each firm will find itself back in the situation depicted in Figure 2. Because of the costs of nonprice competition, each firm's *ATC* curve will be higher than it would otherwise be. However, it will still touch, but not cross, the demand curve, and the firm will still earn zero economic profit. We will take a closer look at one form of nonprice competition—advertising—in the “Using the Theory” section at the end of the chapter.

OLIGOPOLY

A monopolistic competitor enjoys a certain amount of independence. There are so many *other* firms selling in the market—each one such a small fish in such a large pond—that each of them can make decisions about price and quantity without worrying about how the others will react. For example, if a single pharmacy in a large city cuts its prices, it can safely assume that any other pharmacy that could benefit from price cutting has already done so, or will shortly do so, regardless of its own actions. Thus, there is no reason for the price-cutting pharmacy to take the reactions of other pharmacies into account when making its own pricing decisions.

But in some markets, most of the output is sold by just a few firms. These markets are not monopolies (there is more than one seller), but they are not monopolistically competitive either. There are so few firms that the actions taken by any one will *very much* affect the others and will likely generate a response. For example, more than 60 percent of the automobiles sold in the United States are made by one of the “Big Three”: General Motors, Ford, and DaimlerChrysler. If GM were to lower its price in order to increase its output, then Ford and Chrysler would suffer a significant drop in their own sales. They would not be happy about this and would probably respond with price cuts of their own. GM's output, in turn, would be affected by the price cuts at Ford and Chrysler.

When just a few large firms dominate a market, so that the actions of each one have an important impact on the others, it would be foolish for any one firm to ignore its competitors' reactions. On the contrary, in such a market, each firm recognizes its *strategic interdependence* with the others. Before the management team makes a decision, it must reason as follows: “If we take action *A*, our competitors will do *B*, and then we would do *C*, and they would respond with *D . . .*,” and so on. This kind of thinking is the hallmark of the market structure we call *oligopoly*:

An oligopoly is a market dominated by a small number of strategically interdependent firms.



Characterize the Market

There are many different types of oligopolies. The output may be more or less identical among firms—such as copper wire—or differentiated—such as laptop computers. An oligopoly market may be international, as in the market for automobile tires; national, as in the market for breakfast cereals; or local, as in the market for daily newspapers. There may be one dominant firm whose share of the market far exceeds all the others, such as Microsoft in the market for personal

Oligopoly A market structure in which a small number of firms are strategically interdependent.

computer software. Or there may be several large firms of roughly similar size, like Boeing and Airbus in the global market for large passenger aircraft. You can see that oligopoly markets can have different characteristics, but in all cases, *a small number of strategically interdependent firms produce the dominant share of output in the market.*

OLIGOPOLY IN THE REAL WORLD

When we apply our definition of oligopoly to the real world, things begin to look a little ambiguous. First, we must decide how to define our market. If we use a very narrow definition (for example, the market for movies within 10 blocks of your house), we will find very few firms in the market offering similar products. But as we broaden to more typical definitions (such as the market for movies in Austin, Texas), the number of competitors increases, and many of these markets will be seen more accurately as monopolistically competitive. In practice, in deciding whether a market is an oligopoly, we define the market broadly enough to include all reasonably close substitutes. Thus, we refer to the market for breakfast cereal, rather than the market for food, which would be too broad, or the market for cornflakes, which would be too narrow. We refer to the market for steel, rather than the market for metal—too broad—or the market for six-inch steel ingots—too narrow.

A thornier problem in identifying real-world oligopolies is the meaning of the phrase *few firms*. How many firms can there be before the market is no longer an oligopoly? In theory, we require a number small enough so that each firm needs to consider the reactions of its rivals when making decisions—that is, a number small enough for strategic interdependence to occur. But strategic interdependence, itself, is a matter of degree. A market with just 4 large firms will display significant interdependence. As we consider markets with 6, or 10, or 15 firms, interdependence will diminish, and we may choose to apply the nonstrategic model of monopolistic competition instead.

Finally, “market domination” by a few firms is not a precise concept. In order for firms to be strategically interdependent—the key requirement of an oligopoly—the top few firms must produce a large share of the market output. If the three largest firms together had a market share of, say, 5 percent, decisions by any one would have very little impact on the others. With a 90 percent share, we would all agree that they dominate that market, that they will be strategically interdependent, and that the market is therefore an oligopoly. But if the combined market share is only, say, 30 percent, strategic interaction may be weak enough to ignore. Again, the monopolistic competition model might be more appropriate.

You can see that oligopoly is a matter of degree, not an absolute classification. We can imagine a spectrum: At one end are industries in which a very small number of firms produce a large share of the output. In these industries, there is strong strategic interdependence among firms, so our ideas about oligopoly will fit very closely. As we proceed along the spectrum, market domination by the largest firms decreases, strategic interdependence declines, and oligopoly analysis has less to contribute to our understanding of firm and market behavior.

WHY OLIGOPOLIES EXIST

Oligopoly firms do not always earn economic profit in the long run, but even when they do, entry into the market does not occur—a few large firms continue to dominate the industry. Thus, our search for the origin of oligopolies is really a search for

the specific *barriers to entry* that keep out competitors and maintain the dominance of just a few firms. What are these barriers?

Economies of Scale: Natural Oligopolies. Economies of scale (see Chapter 6) can explain why some industries remain oligopolies. The output level at which economies of scale are exhausted—and the firm’s *LRATC* curve bottoms out—is called the firm’s **minimum efficient scale (MES)**. A firm’s MES depends on its production technology and the prices it must pay for its inputs.

Figure 3 illustrates three different possibilities for the MES of a typical firm in an industry. In all three cases, the minimum long-run average cost is assumed to be \$50, so this is the lowest price firms could charge without suffering a long-run loss. The demand curve in such a market tells us that, if price *were* \$50, quantity demanded would be 100,000 units. Since \$50 is the lowest price firms could charge in the long run, 100,000 units is the greatest possible long-run quantity sold in this market.

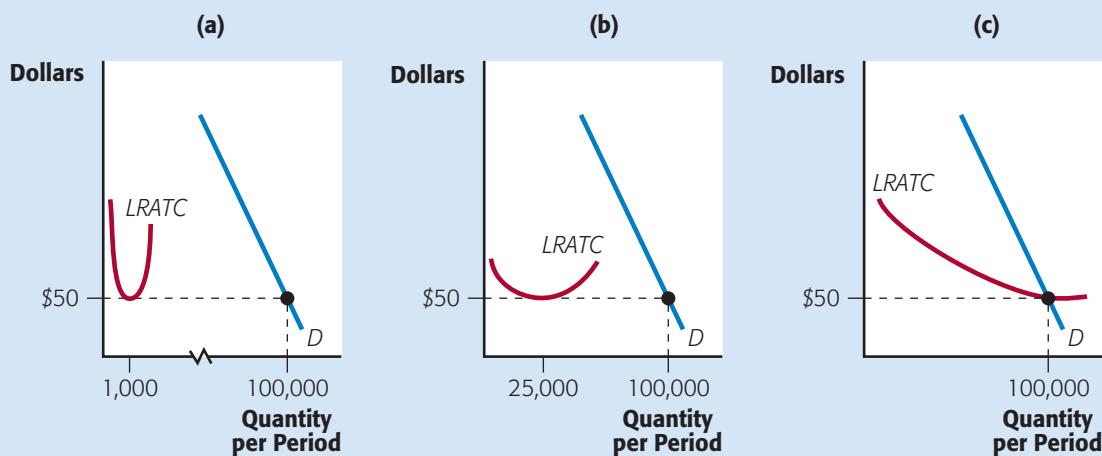
In panel (a), the MES occurs at 1,000 units. A small firm producing 1,000 units would have a cost advantage over a large firm producing, say, 10,000 units. If there are no barriers to entry, we would expect to see many small firms in this market (but no more than 100 firms—can you explain why?) Thus, we would expect the market to be either perfectly competitive or monopolistically competitive.

Panel (c) illustrates the case of *natural monopoly*, discussed in Chapter 9. Here, economies of scale continue over such a wide range of output that the lowest cost per unit is achieved when a single firm produces for the entire market. Once a single firm is established in this industry, it could prevent entry by underpricing any potential competitor, since—by producing for the entire market—it achieves the lowest possible average cost.

Minimum efficient scale (MES) The level of output at which economies of scale are exhausted and minimum *LRATC* is achieved.

MINIMUM EFFICIENT SCALE AND MARKET STRUCTURE

FIGURE 3



If minimum *LRATC* is \$50, the lowest price that could be charged in the long run is \$50. Given the demand curve *D*, 100,000 units is the maximum long-run quantity that could be sold. In panel (a), MES occurs at 1,000 units. With no barriers to entry, a large number of small firms should populate the market. Panel (c) illustrates a natural monopoly; the lowest average cost is achieved when a single firm supplies the entire market. Panel (b) shows a natural oligopoly. MES occurs at 25,000 units, so there should be no more than four firms.

Finally, panel (b) illustrates the intermediate case, where economies of scale extend over a large range of output, but there is still room for more than one competitor. This is the case of *natural oligopoly*. In the figure, the MES is achieved at 25,000 units, and we would expect this industry to have no more than four firms. Why? Because once four firms are established, they could easily underprice any new entrant by temporarily producing 25,000 units each, achieving the lowest possible cost per unit, and charging a price of \$50. (Why couldn't *five* firms each produce 25,000 units in this market?) Moreover, a new entrant must often start out small and then attempt to gradually gain customers from the preexisting firms. But by starting small, the new entrant will have higher costs per unit than any of those firms and will have a hard time gaining customers from them.

Reputation as a Barrier. A new entrant may suffer just from being new. Established oligopolists are likely to have favorable reputations. In many oligopolies—like the markets for soft drinks and breakfast cereals—heavy advertising expenditure has also helped to build and maintain brand loyalty. A new entrant might be able to catch up to those already in the industry, but this may require a substantial period of high advertising costs and low revenues. In some cases, where the potential profits are great, investors may decide it is worth the risk and accept the initial losses in order to enter the industry. Ted Turner took such a risk and sustained several years of losses before his cable ventures (Cable News Network, Turner Network Television, and Turner Broadcasting System) earned a profit. But in other industries, the initial losses may be too great and the probability of success too low for investors to risk their money starting a new firm. And even if they do, they may prefer a low-cost/low-risk strategy, setting up a small firm that does not challenge the dominance of those already in the market. Tom's of Maine—a small upstart toothpaste company—seems to be following this approach. Next time you are in the supermarket, look at the shelf space devoted to Tom's of Maine compared to that given over to dominant brands such as Colgate and Crest.

Strategic Barriers. Oligopoly firms often pursue strategies *designed* to keep out potential competitors. They can maintain excess production capacity as a signal to a potential entrant that, with little advance notice, they could easily saturate the market and leave the new entrant with little or no revenue. They can make special deals with distributors to receive the best shelf space in retail stores or make long-term arrangements with customers to ensure that their products are not displaced quickly by those of a new entrant. And they can spend large amounts on advertising to make it difficult for a new entrant to differentiate its product.

Government-Created Barriers. Like monopolies, oligopolies are not shy about lobbying the government to preserve their market domination. One of the easiest targets is foreign competition. U.S. steel companies are relentless in their efforts to limit the amount of foreign—especially Japanese—steel sold in the U.S. market. In the past, they have succeeded in getting special taxes on imported steel and financial penalties imposed upon successful foreign steel companies. Other U.S. industries—including automobiles, textiles, and computer memory chips—have had similar successes.

Government barriers can operate against domestic entrants, too. Zoning regulations may prohibit the building of a new supermarket, movie theater, or auto repair shop in a local market, preserving the oligopoly status of the few firms

already established there. Lobbying by established firms is often the source of these restrictive practices.²

OLIGOPOLY BEHAVIOR

Of the market structures you have studied in this book, oligopoly presents the greatest challenge to economists. In the other types of markets—perfect competition, monopoly, and monopolistic competition—each firm acts independently, without worrying about the reactions of other firms. Its task is a simple one: to select an output level along its demand curve that gives it maximum profit.

But this approach doesn't describe an oligopolist. The essence of oligopoly, remember, is *strategic interdependence*, wherein each firm anticipates the actions of its rivals when making decisions. Thus, we cannot analyze one firm's decisions in isolation from other firms. In order to understand and predict behavior in oligopoly markets, economists have had to modify the tools used to analyze the other market structures and to develop entirely new tools as well.

Let's look at this idea of strategic interdependence more closely and see why the simple approach used in other markets will not work in an oligopoly. Imagine that Kafka Exterminators, instead of being a monopolistic competitor, was one of just three exterminators in town—an oligopolist. In order to draw Kafka's demand curve—like the one in Figure 1 (p. 277)—we would have to assume that the prices of its competitors remain unchanged as Kafka changes its own price. But what if one or both competitors *lowered* their prices? Then these competitors would lure some of Kafka's customers away, and Kafka's demand curve would shift leftward—it would have fewer customers at any given price. Similarly, if one or both rivals *raised* their price, Kafka's demand curve would shift rightward. Thus, the position of Kafka's demand curve will depend on the prices set by its rivals.

This complicates the firm's decision making. Each time Kafka considers moving along its demand curve by changing its own price, it knows its competitors will react by changing *their* prices, causing Kafka's own demand curve to shift. Thus, Kafka does not face a stable demand curve, and we cannot analyze its decision-making process with a simple $MC = MR$ rule, as we did in other types of markets. You can see why oligopoly presents such a challenge, not only to the firms themselves, but also to economists studying them.

Although great progress has been made, there is not yet a single, unified theory of oligopoly. Rather, there have been a variety of approaches, with important new discoveries continuing to deepen our understanding. The approaches that have offered the richest insights into oligopoly behavior make use of **game theory**.

Game theory An approach to modeling the strategic interaction of oligopolists in terms of moves and countermoves.

The Game Theory Approach. The word *game* applied to oligopoly decision making might seem out of place. Games—like poker, basketball, or chess—are usually played for fun, and even when money is at stake, the sums are usually small. What do games have in common with important business decisions, where hundreds of millions of dollars and thousands of jobs may be at stake?

² This kind of lobbying is often disguised. In July 1995, Home Depot, Inc., sued Rickel Home Centers for secretly forming an organization called Concerned Citizens for Community Preservation, whose sole purpose was to prevent Home Depot from opening new stores in towns where Rickel already had its own outlets (*The Wall Street Journal*, August 18, 1995, p. 1).



<http://>

The Prisoner's Dilemma game is the prototype for thinking about game theory in economics. Visit Bryn Mawr College's interactive prisoner's dilemma game at <http://serendip.brynmawr.edu/~ann/pd.html> to try it out.

Payoff matrix A table showing the payoffs to each of two players for each pair of strategies they choose.

In fact, quite a bit. In all games—except those of pure chance, such as roulette—a player's strategy must take account of the strategies followed by other players. This is precisely the situation of the oligopolist. Game theory analyzes oligopoly decisions as if they were games by looking at the rules players must follow, the payoffs they are trying to achieve, and the strategies they can use to achieve them.

The Prisoner's Dilemma. The easiest way to understand how game theory works is to start with a simple, noneconomic example—the *prisoner's dilemma*—that explains why a technique for obtaining confessions, commonly used by police, is so often successful. Imagine that two partners in crime (let's call them Rose and Colin) have committed a serious offense (say, murder) but have been arrested for a lesser offense (say, robbery). The police have enough evidence to ensure a robbery conviction, but their evidence for murder cannot be used in court. Their only hope for a murder conviction is to get one or both partners to incriminate the other.

The traditional strategy is to separate the partners and explain the following to each one: "Look, you're already facing a five-year sentence for robbery. But we'll offer you a deal: If you confess to the murder and implicate your partner, and your partner does *not* confess, we'll make sure that the D.A. goes easy on you. You'll get three years, tops. If you and your partner *both* confess, we'll send you each away for 20 years. But if your partner confesses, and you do *not*, we'll send *you* away for 30 years."

Each partner in this situation is a *player* in a *game*, and Figure 4 shows the **payoff matrix** for this game—a listing of the payoffs that each player will receive for each possible combination of strategies the two might select. The payoff matrix presents a lot of information at once, so let's take it step-by-step.

FIGURE 4

THE PRISONER'S DILEMMA

		Colin's Actions	
		Confess	Don't Confess
Rose's Actions	Confess	Rose gets 20 years Colin gets 20 years	Rose gets 3 years Colin gets 30 years
	Don't Confess	Rose gets 30 years Colin gets 3 years	Rose gets 5 years Colin gets 5 years

First, notice that each *column* represents a strategy that Colin might choose: confess or not confess. Second, each *row* represents a strategy that Rose might select: confess or not confess. Thus, each of the four boxes in the payoff matrix represents one of four possible strategy combinations that might be selected in this game:

1. Upper left box: both Rose and Colin confess.
2. Lower left box: Colin confesses and Rose doesn't.
3. Upper right box: Rose confesses and Colin doesn't.
4. Lower right box: Neither Rose nor Colin confesses.

Let's now look at the game from Colin's point of view. The entries shown in *purple* in each box are Colin's possible *payoffs*—jail sentences. (Ignore the red entries for now.) For example, the lower left square shows that when Colin confesses and Rose does not, Colin will receive just a three-year sentence.

Colin wants the best possible deal for himself, but he is not sure what his partner will do. (Remember, they are in separate rooms.) So Colin first asks himself which strategy would be best *if* his partner were to confess. The *top row* of the matrix guides us through his reasoning: "If Rose decides to confess, my best choice would be to confess, too, because then I'd get 20 years rather than 30." Next, Colin determines the best strategy if Rose does *not* confess. As the *bottom row* shows, he'll reason as follows: "If Rose does not confess, my best choice would be to confess, because then I'd get 3 years rather than 5."

Let's recap: If Rose confesses, Colin's best choice is to confess; if Rose does *not* confess, Colin's best choice is—once again—to confess. Thus, regardless of Rose's strategy, Colin's best choice is to confess. In this game, the strategy "confess" is an example of a *dominant strategy*:

A dominant strategy is a strategy that is best for a player regardless of the strategy of the other player.

If a player has a dominant strategy in a game, we can safely assume that he will follow it.

What about Rose? In another room, she is presented with the *same* set of options and payoffs as her partner—as shown by the red entries in the payoff matrix. When Rose looks down each *column*, she can see her possible payoffs for each strategy that Colin might follow. As you can see (and make sure that you can, by going through all the possibilities), Rose has the same dominant strategy as Colin—confess. We can now predict that *both* players will follow the strategy of confessing and that the outcome of the game—the upper left-hand corner—is a confession from both partners, with each receiving a 20-year sentence.

Notice that there is a better outcome for both Rose and Colin, located in the lower right corner. But this outcome—5 years in prison for each of them—requires that each of them *not* confess. And that, in turn, requires each of them to trust the other. For if Rose doesn't confess, hoping that Colin will do the same, and Rose turns out to be wrong, then Rose will get 30 years—the worst outcome of all. The same holds for Colin—he, too, will get 30 years if he doesn't confess and Rose does. So, as long as each player acts in an entirely self-interested manner, they are unable to achieve the best outcome for both of them.

Simple Oligopoly Games. The same method used to understand the behavior of Rose and Colin in the prisoner's dilemma can be applied to a simple oligopoly market. Imagine a town with just two gas stations: Gus's Gas and Filip's Fillup. This is an example of an oligopoly with just two firms, called a **duopoly**. We assume that



For centuries, police investigators have used the logic of the Prisoner's Dilemma game to outsmart partners in crime.

Dominant strategy A strategy that is best for a firm no matter what strategy its competitor chooses.

Duopoly An oligopoly market with only two sellers.

Identify Goals and Constraints



Gus and Filip, like Rose and Colin in the prisoner's dilemma, must make their decisions independently, without knowing in advance what the other will do.

Figure 5 shows the payoff matrix facing each duopolist, where each of the two must choose between a high price and a low price for his gasoline.³ The columns of the matrix represent Gus's possible strategies, while the rows represent Filip's strategies. Each square shows a possible payoff—yearly profit—for Gus (shaded purple) and Filip (shaded red). (Make sure you can see, for example, that if Gus sets a high price and Filip sets a low price, then Gus will suffer a loss of \$10,000 while Filip will enjoy a profit of \$75,000.)

The payoffs in the figure follow a logic that we find in many oligopoly markets: Each firm will make greater profit if all firms charge a higher price. But the best situation for any one firm is to have its rivals charge a high price, while *it alone* charges a low price and lures customers from the competition. The worst situation for any one firm is to charge a high price while its rivals charge a low one, for then it will lose much of its business to its rivals.

The entries in the payoff matrix in Figure 5 reflect this situation: Profits are higher (\$50,000) for both Gus and Filip when they both charge a high price and lower (\$25,000) when they both charge a low price. But when the two follow different strategies, the low-price firm gets the best possible payoff (\$75,000), while the high-price firm gets the worst possible payoff (−\$10,000).

Let's look at the game from Gus's point of view, using the purple-shaded entries in the payoff matrix. If Filip chooses a low price (the top row), then Gus should choose a low price, too, since this will get him a \$25,000 profit instead of a \$10,000 loss. If Filip selects a high price (the bottom row), then, once again, Gus should choose a low price, since this will get him a profit of \$75,000 rather than \$50,000. Thus, no matter what Filip does, Gus's best move is to charge a low price—his *dominant* strategy.

A similar analysis from Filip's point of view, using the red-shaded entries, would tell us that his dominant strategy is the same: a low price. Thus, the outcome of this game is the box in the upper left-hand corner, where both players charge a low price and each earns a profit of \$25,000.

Find the Equilibrium



The outcome of the game is also the market *equilibrium* for this oligopoly. When each decision maker is charging the low price, he is doing the best that he can do, given the actions of the other. Therefore, once they reach the upper left-hand corner, neither Gus nor Filip will have any incentive to change his price. In our simple characterization, where a low price and a high price are the only options, Gus and Filip are each doing the best that they can do, given the choice of the other. As long as each is acting independently, neither has any incentive to change his decision. The *market equilibrium* price, in this case, is the low price.

Oligopoly Games in the Real World. While our simple example helps us understand the basic ideas of game theory, real-world oligopoly situations are seldom so simple. First, there will typically be more than two strategies from which to choose (for example, a variety of different prices or several different amounts to spend on *nonprice* competition such as advertising). Also, there will usually be more than two players, so a two-dimensional payoff matrix like the one in Figure 5 would not suf-

³ In a real-world market for gasoline—even one with just two gas stations—there would be many prices from which to choose. Our assumption of just two prices is a “simplifying assumption” that makes it easier to see what is going on.

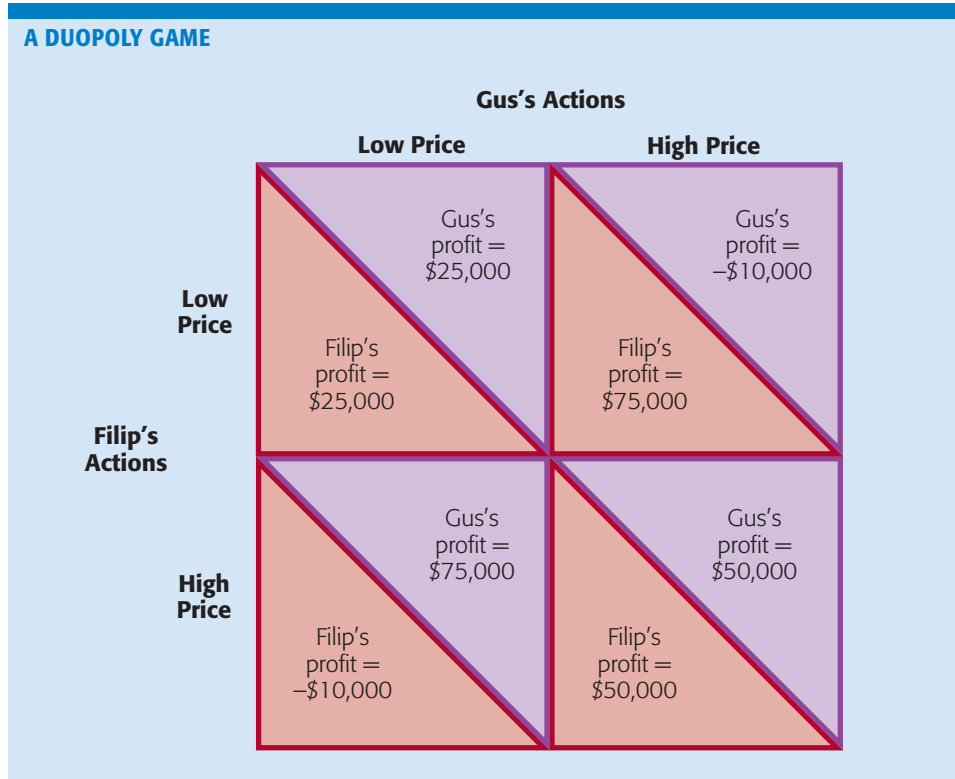


FIGURE 5

fice. Still, as long as each firm has a dominant strategy, we can predict the outcome of the game—the market equilibrium—although we might need the help of a computer in the more complex cases.

Second, in some games, one or more players may *not* have a dominant strategy. For example, if we alter just one entry in the figure—changing Gus's payoff of \$75,000 (lower left-hand box) to \$40,000—Gus would no longer have a dominant strategy. To see why, draw the revised payoff matrix. (Take a moment to do this before reading on.) If you've drawn the revised matrix correctly, you should be able to verify the following: If Filip charges a low price, Gus should charge a low price; but if Filip charges a high price, Gus should charge a high price. Thus, Gus's choice depends on Filip's choice. However, since we have not changed any of Filip's payoffs, he still has a dominant strategy—to charge a low price. Since Gus *knows* that Filip will select a low price, Gus will always select a low price, too. Thus, we can still predict the market equilibrium: a low price for both firms. This example shows us that *when one player has a dominant strategy, we can still predict the game's outcome whether the other player has a dominant strategy or not.*

But what if we *also* change Filip's payoff of \$75,000 (upper right-hand corner) to \$40,000? Then, as you can verify, *neither* player will have a dominant strategy, so neither can predict what the other will do. Moreover, we—as outside observers—will be unable to predict the outcome. *When neither player has a dominant strategy, we will need a more sophisticated analysis to predict an outcome to the game.*

Third, in our example, we've limited the players to *one* play of the game. While this might make sense in the prisoner's dilemma—where the players get only one chance to make a decision—it is not realistic for most oligopoly markets. In reality, for gas stations and almost all other oligopolies, there is **repeated play**, where

Repeated play A situation in which strategically interdependent sellers compete over many time periods.

both players select a strategy, observe the outcome of that trial, and play the game again and again, as long as they remain rivals. Repeated play can fundamentally change the way players view a game and lead to new strategies based on long-run considerations. One possible result of repeated trials is *cooperative behavior*, to which we now turn.

Find the Equilibrium



COOPERATIVE BEHAVIOR IN OLIGOPOLY

In the real world, oligopolists will usually get more than one chance to choose their prices. Pepsi and Coca-Cola have been rivals in the soft drink market for most of this century, as have Ford, DaimlerChrysler, and GM in the automobile market and Kellogg, Post (Kraft Foods), Quaker, and General Mills in the breakfast cereal market. These firms can change their prices based on the past responses of their rivals.

The equilibrium in a game with repeated plays may be very different from the equilibrium in a game played only once. Often, firms will evolve some form of *cooperation* in the long run.

For example, look again at Figure 5. If this game were played only once, we would expect each player to pursue its dominant strategy, select a low price, and end up with \$25,000 in yearly profit. But there is a better outcome for both players. If each were to charge a high price, each would make a profit of \$50,000 per year. If Gus and Filip remain competitors year after year, we would expect them to realize that by cooperating, they would both be better off. And there are many ways for the two to cooperate.

Explicit collusion Cooperation involving direct communication between competing firms about setting prices.

Cartel A group of firms that selects a common price that maximizes total industry profits.

Explicit Collusion. The simplest form of cooperation is **explicit collusion**, in which managers meet face to face to decide how to set prices. In our example, Gus and Filip might strike an agreement that each will charge a high price, moving the outcome of the game to the lower right-hand corner in Figure 5, where each earns \$50,000 in yearly profit instead of \$25,000.

One form of explicit collusion is a **cartel**, wherein the parties select a price along the market demand curve that maximizes total profits in the industry. They do this by choosing the price and quantity of output that a monopoly would charge if it owned all of the firms in the market. To maintain its monopoly profit, the cartel must ensure that the combined output of all firms equals the profit-maximizing quantity. It accomplishes this by allocating a share of the market output to each member of the cartel.

The most famous cartel in recent years has been OPEC—the Organization of Petroleum Exporting Countries—which meets periodically to set the price of oil and the amount of oil that each of its members can produce. In the mid-1970s, OPEC quadrupled its price per barrel in just two years, leading to a huge increase in profits for the cartel’s members. In the late 1990s, OPEC exerted its muscle once again, doubling the price of oil over a period of 18 months.

If explicit collusion to raise prices is such a good thing for oligopolists, why don’t all oligopolists do it? For two reasons. First, it is illegal in many countries, including the United States, and the penalties, if the oligopolists are caught, can be severe. OPEC was not considered illegal by any of the participating nations. But in most cases, explicit collusion *is* illegal and must be conducted with the utmost secrecy.

Second, it is difficult to maintain explicit collusion. In a cartel, each member can steal business from the others—and increase its profit—by selling more than its allocated share. The cartel needs to have some enforcement mechanism—some way to punish firms that produce more than their agreed-upon shares. Of course, be-

cause of its illegal status, the cartel cannot bring offenders to court. But alternative enforcement mechanisms, such as threatening to allow other members to increase their output, may lack credibility at best or destroy the cartel at worst. (See the discussion of OPEC to follow.)

Tacit Collusion. Since explicit collusion is illegal, it is rare in the United States. But other ways of cooperating have evolved among oligopolists. Any time firms cooperate *without* an explicit agreement, they are engaging in **tacit collusion**. Typically, players adopt strategies along the following lines: “In general, I will set a high price. If my rival also sets a high price, I will go on setting a high price. If my rival sets a low price this time, I will punish him by setting a low price next time.” You can see that if both players stick to this strategy, they will both always set the high price. Each is waiting for the other to go first in setting a low price, so it never happens.

An example of this type of strategy is **tit for tat**, defined as doing to the other player what he has just done to you. In our gas station duopoly, for example, Gus will pick the high price whenever Filip has set the high price in the previous play, and Gus will pick the low price if that is what Filip did in the previous play. With enough plays of the game, Filip may eventually catch on that he can get Gus to set the desired high price by setting the high price himself and that he should not exploit the situation by setting the low price, because that will cause Gus to set the low price next time. The outcome in every play will then be in the lower right-hand corner of Figure 5, with each firm earning the higher \$50,000 in profit.

Tit-for-tat strategies are prominent in the airline industry. When one major airline announces special discounted fares, its rivals almost always announce identical fares the next day. The response from the rivals not only helps them remain competitive, but also provides a signal to the price-cutting airline that it will not be able to offer discounts that are unmatched by its rivals.

However, tit-for-tat is not always effective, and the airline industry has had periods of instability. In 1992, when several airlines announced special restricted summer fares, American Airlines responded with even lower fares, cutting the price of many tickets in half. Most other airlines copied American’s move, resulting in fares that were way below the profit-maximizing level for most of the summer. Continental and Northwest, two of the airlines hit hardest by American’s fare cut, sued American on the theory that American was trying to teach them a lesson not to deviate from normal fares. The jury rejected Continental’s and Northwest’s claims and concluded that this type of conduct was not a violation of antitrust laws. Airline fares were much more stable for several years after the jury’s verdict. American’s rivals were no doubt reluctant to make fare cuts that would attract a similar tit-for-tat reaction.

In May 1999, the U.S. Justice Department accused American of a different and more lethal type of tit-for-tat strategy. According to the government, American dramatically cut prices and increased the number of flights originating in Dallas-Fort Worth in order to drive out three new entrants: Vanguard, Sun Jet International, and Western Pacific. American, as the stronger airline, knew it would be able to tolerate a price war longer than the new entrants, who did not have as many other profitable routes to keep them afloat. If the government’s charges are accurate, American’s actions were designed in part as a signal to other potential entrants. American was in effect saying, “If you’re thinking of competing with us in routes that we dominate, expect a strong, tit-for-tat reaction that will finish you off for good.”⁴

Tacit collusion Any form of oligopolistic cooperation that does not involve an explicit agreement.

Tit for tat A game-theoretic strategy of doing to another player this period what he has done to you in the previous period.

⁴ “Government Sues American Airlines, Accusing It of Predatory Pricing,” *New York Times*, May 14, 1999, and “The Problems with Proving Predatory Pricing,” *New York Times*, May 20, 1999.

Price leadership A form of tacit collusion in which one firm sets a price that other firms copy.

Another form of tacit collusion is **price leadership**, in which one firm—the *price leader*—sets its price, and other sellers copy that price. The leader may be the dominant firm in the industry (the one with the greatest market share, for example), or the position of leader may rotate from firm to firm. During the first half of this century, U.S. Steel typically acted as the price leader in the steel industry: When it changed its prices, other firms would automatically follow. More recently, Goodyear has been the acknowledged leader in the tire industry; its price increases are virtually always matched within days by Michelin, Bridgestone, and most other firms in the industry.

With price leadership, there is no formal agreement. Rather, the choice of the leader, the criteria it uses to set its price, and the willingness of other firms to follow come about because the firms realize—without formal discussion—that the system benefits all of them. To keep the price-following firms from cheating—taking large amounts of business by setting a lower price than the price leader—the leader and the firms that choose to follow it must be able to punish a cheater. They can do this by setting a low price as quickly as possible after anyone cheats. The expectation of that response will be enough to prevent the cheating in the first place.

The Limits to Collusion. It is tempting to think that collusion—whether explicit or tacit—gives oligopolies absolute power over their markets, leaving them free to jack up prices and exploit the public without limit. But oligopoly power—even with collusion—has its limits.

First, even colluding firms are constrained by the market demand curve: A rise in price will always reduce the market quantity demanded. There is a single price—the cartel monopoly price—that maximizes the total profits of all firms in the market, and it will never serve the group's interest to charge any price higher than this.

Second, collusion—even when it is tacit—may be illegal. Although it may be difficult to prove, companies that even *appear* to be colluding may find themselves facing close government scrutiny. Indeed, hardly a month goes by without the announcement of one or more new investigations of collusion by the Justice Department.⁵

Third, collusion is limited by powerful incentives to cheat on any agreement. As the next section shows, cheating is an endemic problem among colluding oligopolists and often leads to the collapse of even the most formal agreements.

The Incentive to Cheat. Let's go back to Gus and Filip for a moment. After repeated plays of the game in Figure 5, with each play ending in the upper left-hand corner (\$25,000 in profit for each player), our two gas station owners realize that they can do better with some form of collusion. One way or another—through a formal, explicit agreement, through tit-for-tat behavior, or through an understanding that one of the two will become the price leader—they arrive at the high-price cooperative solution. The outcome of the game then moves to the lower right-hand corner, where each firm earns \$50,000 in profit. Will the market stay there?

Maybe. And maybe not. The problem is, each player may conclude that he can do even better by cheating. For example, once Gus commits to a high price, Filip can make even more profit (\$75,000) by cheating and selling his gasoline at a lower price. This would reduce Gus's profit to -\$10,000, so he, too, would likely switch to the low price, and the two players would be back to the noncooperative outcome based on their dominant strategies.

⁵ Since the prices of other goods and services rose during this period as well, the increase in the *relative* price of oil was somewhat smaller—400 percent—but still quite dramatic.

You might think that in a small-town duopoly of two gas stations such cheating would never occur, since each player can so easily observe what the other is doing, and neither party wants to return to the noncooperative equilibrium. But it may be in each player's interest to cheat *occasionally*. Filip, for example, might think he can enjoy a spell of high profit before Gus has a chance to react, and then—when Gus *does* react—Filip can revert to the cooperative scheme. Gus, by contrast, may try to discourage this with tit-for-tat moves, *punishing* Filip every time he cheats by matching Filip's price or going farther and charging an even lower price (not shown on the payoff matrix). By doing so, he is telling Filip: Cheating is not in your interest. Gus, on the other hand, could then set his price still lower, informing Filip, "You better let me cheat occasionally, because punishing me is not in *your* interest."

As you can see, analyzing this sort of behavior requires some rather sophisticated game theory models, and economists are actively engaged in building them. Some of these models predict occasional price wars such as those observed in small-town markets for gasoline and fresh fruit or national markets for air travel.

When Is Cheating Likely? While no firm wants to completely destroy a collusive agreement by cheating—since this would mean a return to the noncooperative equilibrium wherein each firm earns lower profit—some firm may be willing to *risk* destroying the agreement if the benefits are great enough. In any collusive agreement, we can expect each firm to weigh the costs and benefits of cheating. On the cost side is the probability of being detected, bringing about a punitive reaction from other firms or a collapse of the agreement. On the benefit side is the additional profit from charging a lower price than other firms and gaining additional profit.

This logic suggests that cheating is most likely to occur—and collusion will be least successful—under the following conditions:

Difficulty Observing Other Firms' Prices. In markets where prices are negotiated with each customer—as in general contracting or retail auto sales—it is difficult for firms to observe the prices actually charged by their competitors. In such markets, where the probability of being caught cheating is low, we would expect little cooperation or collusion.

By contrast, when other firms' prices are easy to observe, cheating is more easily detected and therefore less likely to occur. For example, when the buyers are government agencies or public utilities, competing bids must be made public. The airline industry provides another interesting example: In most cases, the airlines can observe their rivals' prices because these are public information. However, since the late 1990s, many airlines have found a way to overcome this problem. By selling tickets to the on-line travel company Priceline.com, which in turn sells them at a variety of different prices to individuals, the airlines can cut prices on at least some of their tickets without alerting the other players. Over time, this should make price cooperation among the airlines more difficult.

Unstable Market Demand. Frequent shifts in market demand encourage cheating for a number of reasons. First, any pre-existing agreement may no longer make sense once market conditions change. A new arrangement must be established, but this often takes time. In the interim, there may be substantial benefits to cheating on the old agreement. Second, with unstable market demand, it is more difficult for firms to interpret each other's actions. If a firm lowers its price, is it cheating on the arrangement? Or is it merely responding to changing demand conditions

and perhaps trying to become a price leader itself? With signals less clear the detection of cheating is less likely, so we would expect collusion to break down.

A Large Number of Sellers. The greater the number of firms, the more cheating we expect to occur. That's because with many firms, each may reason that it can cheat—increasing its output beyond its allocated quantity—without having much of an impact on the market price. Thus, its cheating will go undetected. Even if it thinks its cheating *will* be detected, each firm may view itself as a small fish in a big pond, reasoning that its own cheating will be tolerated as a nuisance, rather than a threat.

However, if several firms behave this way—all increasing their output and hoping the other cartel members won't notice—the market price will drop significantly. Cheating will then be noticed. But in this case, the individual firm has even *more* incentive to cheat—otherwise it suffers.

The history of collusion is rife with cheating and the ultimate breakdown of cooperation among firms, suggesting that these three conditions are sufficiently satisfied in many markets. When the benefits to cheating are great and the costs low, we can expect collusive arrangements to collapse.

THE LIMITS TO OLIGOPOLY

Some people think that the U.S. and other Western economies are moving relentlessly toward oligopoly as the dominant market structure. Technological change is often cited as the reason. For example, in the early part of the century, several dozen U.S. firms manufactured passenger cars. With the development of mass-production technology, the number has steadily fallen to three. Stories like this suggest an economy in which markets are increasingly controlled and manipulated by a few players who—by colluding exploit the public for their own gain. In 1932, two economists—Adolf Berle and Gardiner Means—noted the trend toward big business and predicted that, unless something were done to stop it, the 200 largest U.S. firms would control the nation's entire economy by 1970.

These fears have proven to be unfounded. Today, there are hundreds of thousands of business firms in the United States. Moreover, the evidence shows no strong trend toward increasing concentration in U.S. industries.

We have already noted one reason why: In many industries, the minimum efficient scale of production is so small relative to the size of the market, that small firms have no cost disadvantage. And there are other, powerful forces operating to restrict and even reduce the extent of oligopoly in the economy.

Antitrust Legislation and Enforcement. Antitrust policies in the United States and many other countries are designed to protect the interests of consumers by ensuring adequate competition in the marketplace. In practice, antitrust enforcement has focused on three types of actions: (1) preventing collusive agreements among firms, such as price-fixing agreements; (2) breaking up or limiting the activities of large firms whose market dominance harms consumers; and (3) preventing mergers that would lead to harmful market domination.

The impact of antitrust actions goes far beyond the specific companies called into the courtroom. Managers of other firms considering anticompetitive moves have to think long and hard about the consequences of acts that might violate the antitrust laws. For example, many economists believe that in the late 1940s and early 1950s, General Motors would have driven Ford and Chrysler out of business or bought them out were it not for fear of antitrust action. (Antitrust law is discussed in more detail in Chapter 15.)

The Globalization of Markets. Although oligopolists often try to prevent it, they face increasingly stiff competition from foreign producers. Some economists have argued, for example, that the U.S. market for automobiles now has so many foreign sellers that it resembles monopolistic competition more than oligopoly. Similar changes have occurred in the U.S. markets for color televisions, stereo equipment, computers, beer, and wine. At the same time, the entry of U.S. producers has helped to increase competition in foreign markets for movies, television shows, clothing, household cleaning products, and prepared foods.

Technological Change. You may think that technological change invariably favors bigness and domination by a few firms. But many new technologies serve to *increase* competition by eliminating barriers to entry. The oligopoly of the three major television networks (CBS, ABC, and NBC) was due in part to the limited television broadcast spectrum. Cable television has broken through that barrier and significantly reduced the domination of the networks.

New technology can also destroy a natural oligopoly by eroding economies of scale. Recall (see Figure 3) that in a natural oligopoly, the minimum efficient scale (MES) is large relative to the size of the market. Thus, anything that decreases the MES or increases the size of the market may increase the number of firms that can effectively compete in the industry. In the home entertainment industry, both of these changes are about to occur. Wireless technology will eliminate the need to string expensive cable and thus reduce the number of subscribers needed to minimize cost per unit (the MES). At the same time, this technology will enable firms to sell in more than one locality, thus changing the home entertainment market from a local one to a national one, where dozens or even hundreds of firms will compete.

ADVERTISING IN MONOPOLISTIC COMPETITION AND OLIGOPOLY

We began this chapter by noting that perfect competitors never advertise and monopolies advertise relatively little. But advertising is almost always found under monopolistic competition and very often in oligopoly. Why? All monopolistic competitors, and many oligopolists, produce differentiated products. In these types of markets, the firm gains customers by convincing them that its product is different and better in some way than that of its competitors. Advertising, whether it merely informs customers about the product (“The new Toyota Corolla gets 45 miles per gallon on the highway”) or attempts to influence them more subtly and psychologically (“Our exotic perfume will fill your life with mystery and intrigue”), is one way to sharply differentiate a product in the minds of consumers. Since other firms will take advantage of the opportunity to advertise, any firm that *doesn’t* advertise will be lost in the shuffle. In this section, we use the tools we’ve learned in this chapter to look at some aspects of the economics of advertising.

ADVERTISING AND MARKET EQUILIBRIUM UNDER MONOPOLISTIC COMPETITION

A monopolistic competitor advertises for two reasons: to shift its demand curve rightward (greater quantity demanded at each price) and to make demand for its output *less* elastic (so it can raise price and suffer a smaller decrease in quantity demanded). Advertising costs money, so in addition to its impact on the demand

Using the
THEORY



What Happens When Things Change?

curve, it will also affect the firm's ATC curve. What is the ultimate impact of advertising on the typical firm?

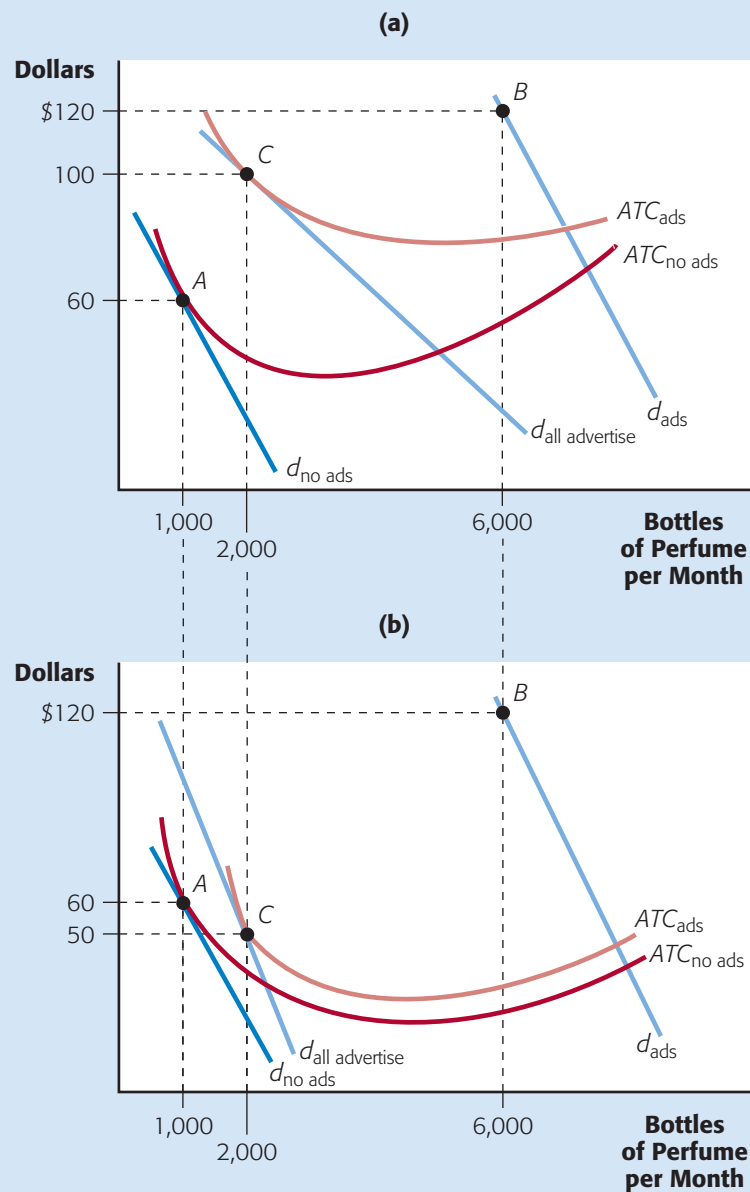
Figure 6(a) shows demand and ATC curves for a company—Narcissus Fragrance—that manufactures and sells perfume. Initially, when there is no advertising at all in the industry, Narcissus is in long-run equilibrium at point A , in panel (a), where its demand curve ($d_{no\ ads}$) and ATC curve ($ATC_{no\ ads}$) touch. The firm charges \$60 per bottle, sells 1,000 bottles each month, and earns zero economic profit.

Now suppose that Narcissus decides to run a costly television ad campaign and, for now, that no other firm advertises. Then the cost of advertising will shift the

FIGURE 6

ADVERTISING IN MONOPOLISTIC COMPETITION

Narcissus Fragrance is initially in long-run equilibrium at point A in panel (a). If it runs an advertising campaign, its average costs increase to ATC_{ads} and its demand curve shifts out to d_{ads} . As a result, it earns an economic profit per unit of $P - ATC_{ads}$. This profit leads its competitors to undertake their own advertising. As they do, Narcissus's demand curve shifts inward to $d_{all\ advertise}$. Long-run equilibrium is reestablished at point C , where Narcissus is again earning zero economic profit. In panel (b), as in panel (a), advertising shifts the ATC curve upward. In this case, though, expansion of industry output drives down the firm's costs per unit. The decline is great enough so that the long-run price is lower after advertising (\$50) than before (\$60).



company's ATC curve upward, to ATC_{ads} . Cost per unit will be greater at every output level. Notice, however, that the rise is smaller at higher output levels, where the cost of the ad is spread over a larger number of units. In addition to the shift in ATC , the ad campaign would shift the demand curve rightward and make it steeper. Since Narcissus is the *only* firm advertising, the effect on the demand curve would be substantial, shifting it all the way to d_{ads} . Notice that there are many points along this new demand curve where Narcissus could earn a profit; the greatest profit will be the output level at which its MC and MR curves (not shown) intersect. In panel (a), we assume that this occurs at 6,000 bottles per month. The firm will thus operate at point B along its new demand curve, charge a price of \$120 per bottle, and earn an economic profit, since $P > ATC$.

Narcissus will not be able to remain at point B for long. In the long run, its profit will tempt other firms to initiate ad campaigns of their own, which will shift Narcissus's demand curve leftward and make it flatter. And if, with all firms advertising, there are *still* profits in the market, then entry will occur, shifting the demand curve farther leftward. In the end, Narcissus will end up at a point like C on demand curve $d_{all\ advertising}$ where $P = ATC = \$100$, and the firm earns zero economic profit.

Notice that, once other firms are advertising, Narcissus *must* advertise as well. Why? If it chooses not to advertise, its ATC curve will return to $ATC_{no\ ads}$, but its demand curve will lie somewhere to the *left* of $d_{no\ ads}$. ($d_{no\ ads}$ was the demand curve when *no* firm was advertising. But now, with its competitors running ad campaigns, if Narcissus chooses *not* to advertise, it will sell less output at any price than it did originally.) With average costs given by $ATC_{no\ ads}$, and the demand curve somewhere to the left of $d_{no\ ads}$, Narcissus would suffer a loss at *any* output level. Thus, if it wants to stay in business in the long run, it *must* advertise.

We can summarize the impact of advertising as illustrated in panel (a) this way: The output of the typical firm has increased (from 1,000 to 2,000 units), and thus, advertising has increased the total size of the market—more perfume is being bought than before. But the individual firm does not benefit from this. Since each firm must pay the costs of advertising, and more competitors have entered the market, Narcissus and its competitors are each earning normal economic profit—just as they were originally.

But what about the price consumers will pay? We would think that costly advertising will raise the price to consumers, and in panel (a), that is what has happened: Advertising has raised the price from \$60 to \$100 in the long run.

But this is not the *only* possible result. Panel (b) illustrates the somewhat surprising case where advertising leads to *lower* costs per unit and a *lower* price for consumers. As before, we begin with Narcissus at point A with no advertising in the market, then move to point B when Narcissus is the only firm running ads, and end up at point C after imitation by other firms and entry have eliminated Narcissus's economic profit.

Notice that, in panel (b), the ultimate impact of advertising is to decrease both cost per unit and price from \$60 to \$50. How can this be? By advertising, the firm is able to produce and sell more output. This remains true even when *all* firms advertise because total market demand has increased. Since the firm was originally on the downward-sloping portion of its ATC curve, we know that its *non*advertising costs per unit will decline as output expands. If this decline is great enough—as in panel (b)—then costs per unit will drop, even when the cost of advertising is included. In other words, because you and I and everyone else is buying more perfume, each producer can operate closer to capacity output, with lower costs per unit. In the long run, entry will force each firm to pass the cost savings on to us.

Our analysis suggests the following conclusion:

Under monopolistic competition, advertising increases the size of the market. More units are sold. But in the long run, each firm earns zero economic profit, just as it would if no firm were advertising. The price to the consumer, however, may either rise or fall.

What Happens When
Things Change?



ADVERTISING AND COLLUSION IN OLIGOPOLY

In this chapter, you've learned that oligopolists have a strong incentive to engage in tacit collusion. But such collusion is difficult to detect. When one firm raises its prices and others follow, that may be evidence of price leadership, or it may be that costs in the industry have risen, and *all* firms—affected in the same way—have decided independently to raise their prices. But in some cases, such as strategic decisions about oligopoly, we can use a simple game theory model to show that collusion is almost certainly taking place.

Let's take the airline industry as an example. Polls show that passengers are very much concerned about airline safety, and any airline that could convince the public of its superior safety record would profit considerably.⁶ Yet no airline has ever run an advertisement with information about its safety record or attacked that of a competitor. ("Fly United. We'll get you there . . . alive!"). Let's see why.

Figure 7 shows some hypothetical payoffs from this sort of advertising as seen by two firms—United Airlines and American Airlines—competing on a particular route. Focus first on the top, purple-shaded entries, which show the payoffs for American. If neither firm ran safety ads, American would earn a level of profit we will call *medium*, as a benchmark. If American ran ads touting its own safety, but United did not, American's profit would certainly increase—to "high" in the payoff matrix. If both firms ran safety ads—especially negative ads that attacked their rival—the public's demand for airline tickets would certainly decline. Reminded of the dangers of flying, more consumers would choose to travel by train, bus, or car. American's profit in this case would be lower than if *neither* firm ran ads, so we have labeled it "low" in the payoff matrix. Finally, the worst possible result for American—"very low" in the figure—occurs when United touts its own safety record, but American does not.

Now look at American's possible strategies. If United decides to run the ads (the top row), American's best action is to run them as well. If United does not run the ads (bottom row), American's best action is still to run the ads. Thus, American has a dominant strategy: Regardless of what United does, it should run the safety ads.

As you can verify, United, whose payoffs are the lower, red-shaded entries, faces an entirely symmetrical situation, and it, too, has the same dominant strategy: Run the ads. Thus, when each airline acts independently, the outcome of this game is shown in the upper left-hand corner, where each airline runs ads and earns a low profit. So why don't we observe that outcome?

The answer is that the airlines are playing against each other repeatedly and reach the kind of cooperative equilibrium we discussed earlier. Each airline can punish its rival next time if it fails to cooperate this time. In the cooperative outcome, each airline plays the strategy that it will *not* run the ads as long as its rival

⁶ There are a number of ways to interpret accident statistics: number of passenger deaths or injuries per mile flown, number of crashes per mile flown, number of passenger deaths or injuries per takeoff, and so on. By searching hard enough, most airlines could come up with a measure by which they would appear the "safest."

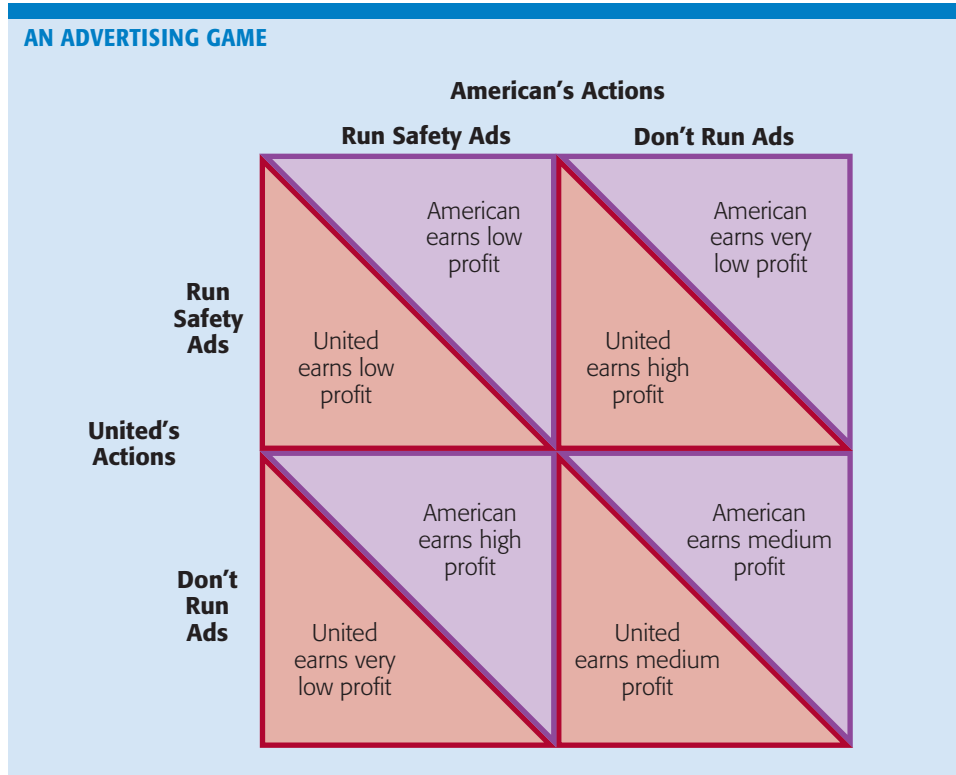


FIGURE 7

does not. The game's outcome moves to the lower right-hand corner. Here, neither firm runs ads, and each earns medium rather than low profit. This is the result we see in the airline industry.

Should we be surprised at the cooperative outcome in this case? Not really. Recall that the ability to get away with cheating is one of the chief obstacles to cooperation. But when the agreement involves *advertising*, cheating would be instantly detected and would therefore be unlikely to occur. This makes advertising a particularly good opportunity for cooperation.

Until the 1980s, a similar collusive understanding seemed to characterize the automobile industry. As long as the “Big Three” dominated auto sales in the United States, the word *safety* was never heard in their advertising. There seemed to be an understanding that all three would earn greater profits if consumers were *not* reminded of the dangers of driving. Things changed in the 1980s, however, as foreign firms' share of the U.S. market rose dramatically. One of the new players—Volvo—decided that its safety features were so far superior to its competitors that it no longer paid to play by the rules. Volvo began running television advertisements that not only stressed its own safety features, but implied that competing products were dangerous. (On a rainy night, a worried father stops his son at the door, hands him some keys, and says, “Here, son, take the Volvo.”) Once Volvo began running ads like these, the other automakers had no choice but to reciprocate. Now, automobile ads routinely mention safety features like antilock brakes and air bags.

Something similar may be about to happen in the airline travel industry. But the “Volvo” in this case is not an airline but an aircraft manufacturer. In November 1999, Airbus ran ads designed to convince the public that its four-engine A340 jets were safer for transatlantic travel than Boeing's twin-engine 777s. A print ad—taken

out in more than a dozen newspapers and magazines, including the *Economist*, *Fortune*, and the *Wall Street Journal*—shows a lone Airbus A340 flying under ominous, dark skies, with a choppy sea below. The caption reads, “If you’re over the middle of the Pacific, you want to be in the middle of four engines.” Not surprisingly, Boeing condemned the ad, declaring that “this not so subtle scare-tactic . . . is a dramatic departure from the high standards our industry has traditionally met. Airbus’s actions have, rightfully so, raised a considerable amount of displeasure in our industry.” The major airlines reacted even more strongly. The CEO of Continental Airlines, Gordon Bethune, informed Airbus that the ad “makes it more unlikely we would put our confidence in you or your products.”⁷

Has Airbus’s action permanently destroyed the “no-ads” cooperative equilibrium among aircraft manufacturers? Or will cooperation be restored? Only time will tell.

THE FOUR MARKET STRUCTURES: A POSTSCRIPT

You have now been introduced to the four different market structures: perfect competition, monopoly, monopolistic competition, and oligopoly. Each has different characteristics, and each leads to different predictions about pricing, profit, non-price competition, and firms’ responses to changes in their environments.

Table 1 summarizes some of the assumptions and predictions associated with each of the four market structures. While the table is a useful review of the *models* we have studied, it is not a how-to guide for analyzing real-world markets: We cannot simply look at the array of markets we see around us and say, “This one is perfectly competitive,” “That one is an oligopoly,” and so on. Why not? Because markets in the real world will typically have characteristics of more than one kind of market structure. A barbecue restaurant, for example, may be viewed as a monopolistic competitor in the market for restaurants in Memphis, or an oligopolist in the market for barbecue restaurants in Memphis, or a monopolist in the market for barbecue restaurants within walking distance of Graceland.

You can see how market structure models help us organize and understand the apparent chaos of real-world markets. Now, it seems, we’ve ended up with a different type of chaos: We can usually choose among two, three, or even four different models when studying a particular market.

But, as we’ve seen several times in this text, our choice of model is not really arbitrary; rather, it depends on the *questions we are trying to answer*. To explain why a *particular* barbecue restaurant with no nearby competitors earns economic profit year after year, or why it spends so much of its profit on rent-seeking activity (lobbying the local zoning board), we would most likely use the monopoly model. If we want to explain why *most* barbecue restaurants do *not* earn much economic profit, or why they pay for advertisements in the yellow pages and the local newspapers, or why there is so much excess capacity (empty tables) in the industry, we would use the model of monopolistic competition. To explain a price war among the few restaurants in a neighborhood, or to explore the possibility of explicit or tacit collusion in pricing or advertising, we would use the oligopoly model. And if we want the *simplest* possible explanations about prices, entry and exit, and profit over the short run and the long run, we would use the perfectly competitive model, which ignores the distinctions between meals at different restaurants and any barriers to entry that might exist.

⁷ “Competitor’s ‘scare tactic’ vexes Boeing,” *Herald Net*, November 6, 1999 (www.heraldnet.com); “Airlines Blast New Ads from Airbus . . .” *Wall Street Journal*, November 22, 1999.

A SUMMARY OF MARKET STRUCTURES

TABLE 1

	Perfect Competition	Monopolistic Competition	Oligopoly	Monopoly
<i>ASSUMPTIONS ABOUT:</i>				
Number of Firms	Very Many	Many	Few	One
Output of Different Firms	Identical	Differentiated	Identical or Differentiated	—
View of Pricing	Price taker	Price setter	Price setter	Price setter
Barriers to Entry or Exit?	No	No	Yes	Yes
Strategic Interdependence?	No	No	Yes	—
<i>PREDICTIONS:</i>				
Price and Output Decisions	$MC = MR$	$MC = MR$	Through strategic Interdependence	$MC = MR$
Short-Run Profit	Positive, zero, or negative	Positive, zero, or negative	Positive, zero, or negative	Positive, zero, or negative
Long-Run Profit	Zero	Zero	Positive or zero	Positive or zero
Advertising?	Never	Almost always	Yes, if differen-	Sometimes tiated product

This example should convince you that economics is as much art as science, and this, in part, is what keeps it interesting and intellectually challenging. And the questions are continually changing as well. In recent years, microeconomists have addressed a number of problems that could not have been imagined just a decade ago. How does Microsoft's dominant position in the computer software industry affect the price and quality of products at your local software outlet? How would cooperation among IBM, Apple, Compaq, and Toshiba on hardware design ultimately affect consumers? Why is e-mail so cheap, and will it stay that way? Should local and long-distance phone companies be kept out of each other's markets? Why is the pricing of airline tickets so much more complicated and unstable than the pricing of bus or rail tickets? To answer these questions requires specific knowledge about the different industries; but it also requires an understanding of the different market structures and some careful thinking about which one to use in each case.

We will come back to the four market structures again when we consider the operation of the microeconomy as a whole, the notion of economic efficiency, and the proper role of government in the economy. But first we must explore another type of market, one that, until now, we've ignored.

S U M M A R Y

Monopolistic competition is a market structure in which there are many small buyers and sellers, no significant barriers to entry or exit, and firms sell differentiated products. As in monopoly, each firm faces a downward-sloping demand curve, chooses the profit-maximizing quantity where $MR = MC$, and charges the maximum price it can for that quantity. As in perfect competition, short-run profit attracts new entrants. As firms enter the industry, the demand curves facing

existing firms shift left. Eventually, each firm earns zero economic profit and produces a level of output above minimum average cost.

Oligopoly is a market structure dominated by a small number of strategically interdependent firms. New entry is deterred by economies of scale, reputational barriers, strategic barriers, and government-created barriers to entry. Because each firm, when making decisions, must anticipate its rivals'

reactions, oligopoly behavior is hard to predict. However, one approach—*game theory*—has offered rich insights.

In game theory, a *payoff matrix* indicates the payoff to each firm for each combination of strategies adopted by that firm and its rivals. A *dominant strategy* is a strategy that is best for a particular firm regardless of what its rival does. If there is no cooperation among firms, any firm that has a dominant strategy will play it, and that helps predict the outcome of the game. If no firm has a dominant strategy, it is much harder to predict what will happen—especially for games that are played only once.

Sometimes oligopolists can cooperate to increase profits. *Explicit collusion*, in which managers meet to set prices, is illegal in the United States. As a result, other forms of *tacit collusion* have evolved. Still, cheating is a constant threat to collusion. Cheating is most likely when there is difficulty observing prices, when market demand is unstable, and when there are a large number of sellers. Government antitrust enforcement, market globalization, and technological change all threaten collusion by oligopolists.

KEY TERMS

monopolistic competition	minimum efficient scale	duopoly	tacit collusion
nonprice competition	game theory	repeated play	tit for tat
oligopoly	payoff matrix	explicit collusion	price leadership
	dominant strategy	cartel	

REVIEW QUESTIONS

- What features does a monopolistically competitive market share with a perfectly competitive market? With a monopoly market?
- True or false? “In the long run, a monopolistic competitor will produce the level of output that minimizes its average total cost.” Explain.
- True or false? “The only way for a monopolistic competitor to increase its sales is to lower its price.” Explain.
- How does oligopoly differ from monopolistic competition, perfect competition, and monopoly?
- Classify each of the following business firms as perfectly competitive, monopolistically competitive, oligopolistic, or monopolistic. Justify your answer. That is, discuss what characteristic(s) of the market designation you assign are likely to be present.
 - General Motors
 - An Iowa corn farmer
 - Kinko’s copy shop (large city)
 - Kinko’s copy shop (the only copy center within a two-mile radius of your campus)
 - De Beers Diamonds (international)
 - Ben & Jerry’s ice cream (national)
 - Daily newspaper (one of two in a medium-sized city)
 - Spanish-language newspaper (the only one in the Hispanic community of a medium-sized Southwestern city)
- What is the difference between a natural oligopoly and a natural monopoly?
- Discuss some factors that might keep new entrants out of an oligopolistic market.
- What conditions are likely to lead to cheating on a collusive arrangement? Explain why each makes cheating more probable.
- The minimum efficient scale in a certain industry is 2,300 units. Exactly what additional information do you need in order to predict whether this industry will be perfectly (or monopolistically) competitive, an oligopoly, or a monopoly?
- How does technological change limit the degree of concentration in an industry? Give some examples.
- Discuss how much advertising each of the following will be likely to do and why. In each case where a firm may advertise, explain exactly what it might be trying to accomplish with its advertising.
 - Continental Cablevision, the sole cable provider in Cambridge, Massachusetts
 - A dairy farm in upstate New York
 - Blockbuster video stores
 - Homestake, a gold-mining company
 - Dell Computer Co.

P R O B L E M S A N D E X E R C I S E S

1. Draw the relevant curves to show a monopolistic competitor suffering a loss in the short run. What will this firm do in the long run if the situation does not improve? How would this action affect *other* firms in this market?
2. Assume that the plastics business is monopolistically competitive.
 - a. Draw a graph showing the long-run equilibrium situation for a typical firm in the industry. Clearly label the demand, *MR*, *MC*, and *ATC* curves.
 - b. One of the major inputs into plastics is oil. Draw a new graph illustrating the short-run position of a plastics company after an increase in oil prices. Again, show all relevant curves.
 - c. If oil prices remain at the new, higher level, what will happen to get firms in the plastics industry back to a long-run equilibrium?
3. In a small Nevada town, Ptomaine Flats, there are only two restaurants—the Road Kill Cafe and, for Italian fare, Sal Monella’s. Each restaurant has to decide whether to clean up its act or to continue to ignore health code violations.

Each restaurant currently makes \$7,000 a year in profit. If they both tidy up a bit, they will attract more patrons but must bear the (substantial) cost of the cleanup; so they will both be left with a profit of \$5,000. However, if one cleans up and the other doesn’t, the influx of diners to the cleaner joint will more than cover the costs of the scrubbing; the more hygienic place ends up with \$12,000, and the grubbier establishment incurs a loss of \$3,000.

 - a. Write out the payoff matrix for this game, clearly labeling strategies and payoffs to each player.
 - b. What is each player’s dominant strategy?
 - c. What will be the outcome of the game? Explain your answer.
 - d. Suppose the two restaurants believe they will face the same decision repeatedly. How might the outcome differ? Why?
4. Assume that Nike and Adidas are the only sellers of athletic footwear in the United States. They are deciding how much to charge for similar shoes. The two choices are “High” (H) and “Outrageously High” (OH). The payoff matrix is as follows (Nike’s payoffs are in the lower left of each cell):
 - a. Do both companies have dominant strategies? If so, what are they?
 - b. What will be the outcome of the game?
 - c. If Nike becomes the acknowledged price leader in the industry, what will be its dominant strategy? What will be the outcome of the game? Why?
- e. Assume that if one cleans up and one stays dirty, the cleaner restaurant makes only \$6,000 in profit. All other payoffs are the same as before. What will the outcome of the game be now without collusion? With collusion?

		Adidas	
		H	OH
Nike	H	\$500,000 \$1 mil.	\$300,000 \$1.7 mil.
	OH	\$550,000 \$800,000	\$600,000 \$1.2 mil.

C H A L L E N G E Q U E S T I O N S

1. Suppose that the government has decided to tax all the firms in a monopolistically competitive industry. Specifically, suppose it levies a fixed tax on each firm—that is, the amount of the tax is the same regardless of how much output the firm produces. In the short run, how would that tax affect the price, output level, and profit of the typical firm in that industry? What would be the effect in the long run?
2. On page 304 you will find the payoff matrix for a two-player game, where each player has three possible strategies: *A*, *B*, and *C*. The payoff for player 1 is listed in the lower left portion of each cell. Assume there is no cooperation among players.
 - a. Does either player have a dominant strategy? If so, which player or players, and what is the dominant strategy?

- b. Can we predict the outcome of this game from the payoff matrix? Why or why not?
- c. Suppose that strategy C is no longer available to either player. Does either player have a dominant strategy now? Can we now predict the outcome of the game? Explain.

		Player 2		
		A	B	C
Player 1	A	4, 9	6, 2	7, 8
	B	1, 3	4, 8	3, 7
	C	7, 7	3, 6	4, 9

EXPERIENTIAL EXERCISES

1. One important way monopolistic competitors differentiate their products is by location. Read John Campbell's article, "Time to Shop: The Geography of Retailing" in *The Region* (http://www.bos.frb.org/economic/nerr/camp96_3.htm). What locational strategies are retailers using? What does the theory of monopolistic competition predict about the success of these strategies in the short run and in the long run?



2. If you look carefully, you can often find evidence of price leadership. For example, the *Wall Street Journal* frequently runs stories about airfares. Typically, one airline will change its fares—on certain routes or across the board—and other airlines will match those changes within a day or two. So, check the *Wall Street Journal* or Infotrac. When you find such a story, check back over the next few days. Did other airlines match the leader, or was the leader forced to back off the price changes?

THE LABOR MARKET

In the late 1990s, thousands of new specialists in pulmonary medicine, anesthesiology, ophthalmology, neurosurgery, and radiology were shocked and disappointed. After more than 4 years of college—and an additional 10 years of medical school and residency—their salaries ended up far lower than they had anticipated when they started down their career pathways. Why did the salaries of medical specialists fall while these students were preparing to become physicians?

Once you understand how labor markets work—the subject of this chapter—you will know how to answer this question and a host of others: Will salaries in your own field rise or fall in the future? What makes firms decide to hire more workers or lay them off? Why is it that, most of the time, the economy seems to have enough of each type of worker, without any government agency making sure? And why, occasionally, are there shortages in some professions, and what should be done about them?

FACTOR MARKETS IN GENERAL

So far in this book, we have analyzed a variety of markets—for wheat, cable TV service, household exterminators, gasoline, perfume, airline travel, and more. All of these markets had one thing in common: They were **product markets**, in which firms sell goods and services to households or other firms. Of course, products aren't made out of thin air, but rather from the economy's *resources*—labor, capital, land, and natural resources. These resources must be purchased from those who own them. Since resources are sometimes called *factors of production*, the markets in which they are traded are called **factor markets**.

In this and the next two chapters, we switch our focus from product markets to factor markets. Figure 1 illustrates what this switch entails.

Notice that in product markets, households demand the products, and firms supply them. In factor markets, these roles are typically reversed: Firms demand land, labor, and capital, and households, which own them, are the suppliers.

Why are factor markets important? First, because we cannot fully understand markets for goods and services unless we understand markets for the *resources* needed

CHAPTER OUTLINE

Factor Markets in General

Labor Markets in Particular

- Defining a Labor Market
- Competitive Labor Markets
- Firms in Labor Markets

Demand for Labor by a Single Firm

- Goals and Constraints
- The Firm's Employment Decision When Only Labor Is Variable
- The Firm's Employment Decision When Several Inputs Are Variable

The Market Demand for Labor

- Shifts in the Market Labor Demand Curve

Labor Supply

- Individual Labor Supply
- Market Labor Supply
- Shifts in the Market Labor Supply Curve
- Short-Run versus Long-Run Labor Supply

Labor Market Equilibrium

What Happens When Things Change?

- A Change in Labor Demand
- A Change in Labor Supply
- Labor Shortages and Surpluses

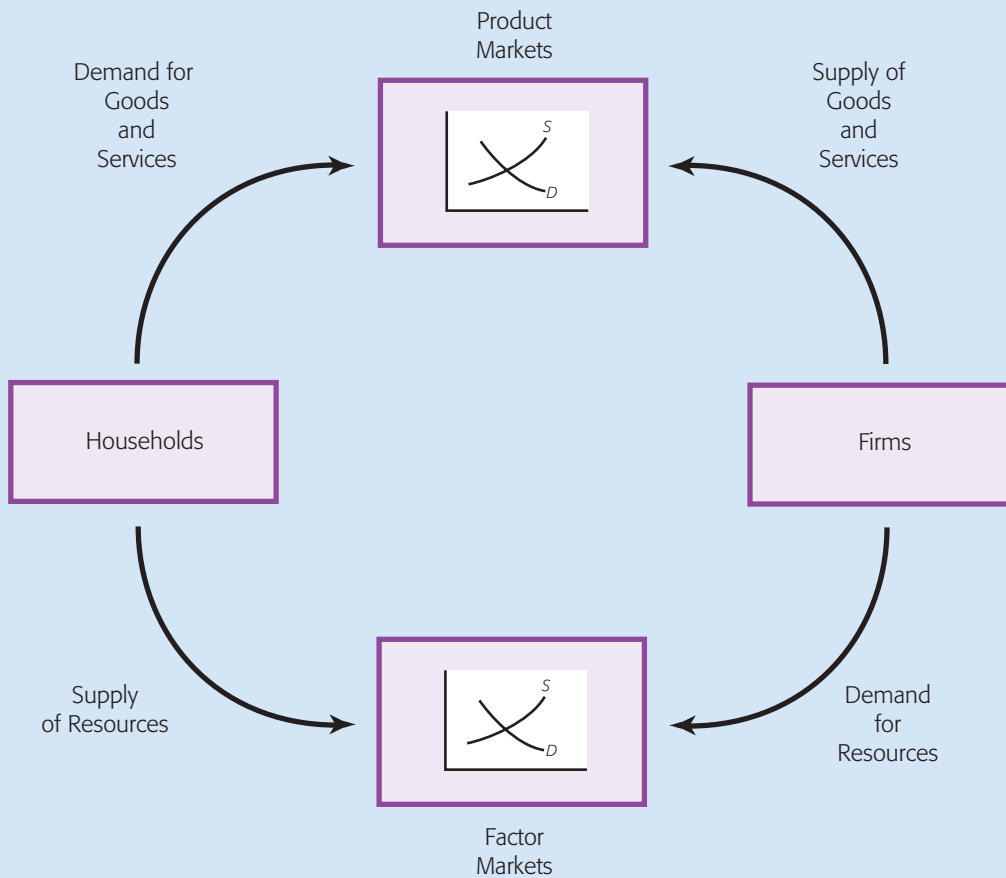
Using the Theory: Understanding the Market for College-Educated Labor

Product markets Markets in which firms sell goods and services to households or other firms.

Factor markets Markets in which resources—capital, land, labor, and natural resources—are sold to firms.

FIGURE 1

PRODUCT MARKETS AND FACTOR MARKETS



In product markets, households demand goods and services, and firms supply them. In factor markets, the roles are reversed: Firms demand labor, capital, land, and natural resources, and households supply them.

to produce them. Indeed, separating these two types of markets as we've done, while useful for learning, is highly artificial, since the decision to produce more output implies a decision to employ more resources. For example, when Ford decides to produce more automobiles, it must use more labor, more machinery, or more land for its factories, and perhaps more of all three. It will also need greater quantities of inputs produced by other firms—steel, tires, windshields—and these firms, in turn, will have to obtain more labor, capital, land, and natural resources to produce them. You can see that *products* and the *resources* used to produce them are really two sides of the same coin. What you learn in these next few chapters will help you understand how these two sides of the economy—product markets and factor markets—fit together.

LABOR MARKETS IN PARTICULAR

The basic approach we will take in studying the labor market may initially strike you as a bit heartless: We will treat labor as a commodity—something that is bought and sold in the marketplace—and regard the wage rate as the price of that

commodity. The wage rate can be defined as an hourly rate (e.g., \$20 per hour), a daily rate (\$160 per day), or for any other time unit.

As a first approximation, we explain how a worker's wage rate is determined in the same way we'd explain the price of a bushel of wheat. That is, we look at how groups of economic decision makers come together in markets in order to trade (Key Step #1), with each decision maker trying to maximize something and each facing constraints (Key Step #2). We then look for the equilibrium price determined in those markets (Key Step #3) and—eventually—explore how various changes affect that equilibrium price (Key Step #4). We do this for one simple reason: It works.

Of course, labor *is* different from other things that are traded. First, sellers of wheat do not care who buys their product, as long as they get the market price. Sellers of labor, on the other hand, care about many things besides their wage rate when they look for a job: working conditions, friendly coworkers, commuting distance, possibilities for advancement, prestige, a sense of fulfillment, and more.

A second distinct feature of labor is the special meaning of the price in this market: the wage rate. Most of the income people earn over their lifetimes will come from their jobs, and their hourly, weekly, or yearly wage will determine how well they can feed, clothe, house, and otherwise provide for themselves and their families. This adds a special moral dimension to events in the labor market.

In this chapter, we apply the basic model of supply and demand to explain how wage rates and employment are determined and what causes them to change. Toward the end of the chapter, we'll also discuss some of the special features of the labor market, and continue to explore them in the next chapter.

DEFINING A LABOR MARKET

If you are like most college students, you will be looking for a full-time job shortly after you graduate. From the economic point of view, you will become a seller in a labor market. But which labor market? As you've seen several times in this book,

how broadly or narrowly we define a market depends on the specific questions we wish to answer.

For example, suppose we are interested in explaining why college graduates, on average, earn more than those with just high school diplomas. Then we would want to define the labor market very broadly: the market for all college-educated labor in the United States. In this market, you would be one of about 35 million sellers, and your employer would be one of hundreds of thousands of buyers.

On the other hand, we might be interested in finding out how salaries in some profession (say, medicine) are determined. For this purpose, we would use a narrower definition: the market for physicians in the United States. The sellers would be all individuals with medical degrees, and the buyers would be all the hospitals, universities, and private practices that hire them. Or, we could go even narrower, and ask why the wage rates of physicians in Boston are higher than the wage rates of physicians elsewhere. Here, the buyers and sellers would be limited to those already in the Boston area or those who could move there within the period we are considering. In this chapter, we will be asking many different questions about labor markets, and will need to look at both broadly and narrowly defined markets to answer them.

COMPETITIVE LABOR MARKETS

Just as there are different types of product markets, so, too, there are different types of labor markets. The number of buyers and sellers, the presence or absence



Characterize the Market



Characterize the Market

Perfectly competitive labor market

Market with many indistinguishable sellers of labor and many buyers, and that involves no barriers to entry or exit.

of barriers to entry, whether all workers in the market are more or less the same—all of these characteristics help to define the *structure* of the labor market. This chapter focuses on **perfectly competitive labor markets**. A labor market is considered to be perfectly competitive if it satisfies three conditions:

1. There are a great many buyers (firms) and sellers (households) of labor in the market.
2. All workers in the market appear the same to firms.
3. There are no barriers to entering or leaving the labor market.

Do these conditions sound familiar? They should, since they are almost identical to the features of perfect competition in a product market (Chapter 8). The only difference is that here it is labor, rather than a good or service, that is being traded.

Few labor markets will strictly satisfy each of these requirements. You might think, then, that the competitive model can be applied only in a few, limited cases. Yet when economists look at real-world labor markets, they use the model of perfect competition more than any other model. Why?

First, the competitive model allows us to come to some powerful conclusions about labor markets and how they respond to changes in the economy using simple techniques. Other labor market models, while often valuable, are also more cumbersome and their predictions less clear cut. Second, most labor markets—while not perfectly competitive—come close enough to justify using the model. The more closely a particular labor market satisfies the conditions, the more accurate our analysis will be.

For example, consider the requirement that all workers are the same. We know that no two workers are ever precisely the same, just as no two bushels of wheat are truly identical: Close inspection would always find some differences. What matters in both cases, though, is that a potential buyer perceives no important differences.

Some labor markets fit this requirement more closely than others. A farmer hiring apple pickers will make little distinction among different job candidates, as long as they all appear to meet the minimum requirements for strength and agility. On the other hand, the manager of a large corporation hiring computer programmers might take more notice of differences in talent and skill among the applicants. Still, the manager will usually regard all graduates of an accredited training program as close substitutes for one another. And in most labor markets, the similarities between workers are more important than their differences, and our assumption that all workers are the same to firms is not too far off the mark.

We will devote most of this chapter to the competitive model, because it can be applied so broadly and because it serves as a benchmark against which other types of labor markets can be measured. When an economist is asked to analyze the market for computer programmers, attorneys, college professors, nurses, librarians, or stockbrokers, he will reach for the competitive model first, even though in each of these cases one or more of the requirements of perfect competition is not strictly satisfied. We will, however, also look at some important departures from perfect competition in the next chapter.¹

FIRMS IN LABOR MARKETS

You might think that firms that compete in the same product market also compete in the same labor market. And this is *sometimes* true. For example, the artichoke farms in Northern California all compete in the same product market (the national

¹ In particular, Chapter 12 will explore what happens in labor markets when there are *barriers to entry*, *differences in ability*, and *discrimination*.

market for artichokes) and also in the same labor market (the market for farm labor in Northern California).

But this is not always the case. First, some firms that compete in the same product market operate in entirely *different* labor markets. For example, Volvo and General Motors share several product markets around the world—such as the U.S. market for passenger cars—but they participate in entirely different labor markets. While Volvo hires most of its labor in the Swedish labor market, GM hires most of its labor in the United States.

Second, some firms that operate in entirely different product markets compete in the *same* labor market. For example, airlines operate in entirely different product markets from Internet consulting firms like Viant and USWeb/CKS. But when the airlines want to hire managers, they must go to the national market for MBAs. In that labor market, the airlines will find themselves in direct competition with the Internet firms, who are also trying to hire MBAs.

The demand side of a labor market includes all firms hiring labor in that labor market. These firms may, but do not necessarily, compete in the same product market.

DEMAND FOR LABOR BY A SINGLE FIRM

A competitive labor market has two sides: buyers and sellers. In this section, we begin our exploration of the buying side of the market—labor demand—by looking at how a typical firm in a labor market decides how much labor to employ. But before we get into the mechanics, let's step back a bit and consider what labor demand is all about.

The demand for labor is unlike the other types of demand you have studied so far in one very important respect: It is a demand for an *input* in production, not a demand for output. When consumers demand outputs—such as perfume, bed frames, or movies—they do so because these things give them pleasure or satisfaction. They are wanted in and of themselves. But Microsoft does not demand labor because its stockholders get satisfaction from employing people. Rather, in its attempt to maximize profit, it chooses to produce and sell a certain amount of software, and this *requires it* to hire a certain number of workers. Microsoft's demand for labor is therefore *derived* from the public's demand for its software:

*The demand for labor is a **derived demand**—it arises from, and will vary with, the demand for the firm's output.*

The phrase “will vary with” is important: The demand for labor by a firm will *change* whenever the demand for the firm's product changes. Over the past decade, as the demand for software has increased, the demand for labor by software manufacturers—such as Microsoft, Intuit, and Oracle—has grown along with it. By contrast, the demand for dictating machines has decreased over the last decade, as even the highest-level managers now type their own memos into their computers. The demand for labor by makers of dictating machines has fallen accordingly.

GOALS AND CONSTRAINTS

How does the firm decide how many workers to hire? As always, we view the firm as an *economic decision maker*, that is striving to *maximize profit*. However, the firm faces *constraints* as it makes its employment decision.

Derived demand The demand for an input that arises from, and varies with, the demand for the product it helps to produce.



Identify Goals and Constraints

These constraints can be simple or complex, depending on how much freedom the firm has to select its inputs. For now, we'll simplify our discussion by assuming that the firm can vary *only* its labor, and is stuck with given quantities of capital and other inputs. This assumption will fit most closely for a firm using a short-run horizon. For example, in the short run, a farm might be able to hire or fire workers, but may be stuck with a given number of tractors and a given amount of land.

What are the constraints the firm faces when labor is the only input that it can vary?

First, the firm faces some given technology, as represented by its production function. The firm's technology tells us how much output the firm can produce with each quantity of labor it might employ. For example, we know that if a farm hires more labor, it can produce more of its crop. The farm's technology tells us *how much* more it can produce each time it hires another worker.

Second, the firm faces a constraint on the price it can charge for its *output*. This constraint is determined by the demand curve faced by the firm. To keep our analysis simple, we'll assume that the firm sells its output in a perfectly competitive product market. As you learned in Chapter 8, this means that the firm faces a horizontal demand curve for its output. Equivalently, the firm is a price taker: It can sell any level of output it chooses to, but each unit must be sold at the market price, and not a nickel more. This may seem somewhat limiting, but remember that the competitive model is a useful approximation to many product markets, even those that do not strictly satisfy all of the conditions of perfect competition. (You may want to review Chapter 8 on this point.)

Finally, the firm must *pay* for its labor, just as it must pay for its other inputs. But how much must the firm pay? Remember that we are dealing with perfectly competitive labor markets in this chapter. Since there are many firms in a competitive labor market—so many that each one hires only a tiny fraction of the total labor available—no single firm's employment decision can have any perceptible impact on the market wage rate. A single restaurant, for example, could double or triple the number of waiters it employs, without having any impact on the wage rate it will have to pay—the going wage for waiters.

Wage taker Any firm that takes the market wage rate as a given when making employment decisions.

In competitive labor markets, each firm is a wage taker: It takes the market wage rate as a given.

This is an important constraint on the firm's behavior in the labor market: It cannot decide what wage rate to pay, but can only decide *how many workers to hire at the going wage* for its labor. In sum, as long as we treat the firm's product market and its labor market as perfectly competitive:

the firm faces three constraints as it decides how much labor to employ. (1) Its technology determines how much output the firm can produce with each quantity of labor; (2) the market price in its product market tells the firm how much it can sell its output for; and (3) the market wage rate in its labor market tells the firm how much it must pay each worker.

There is nothing the firm can do about these constraints. However, given these constraints, the firm can choose how many workers to hire. How does the firm make this decision? As always, we will use the principle of marginal decision making to understand the firm's behavior, examining how a particular decision changes its revenue on the one hand and its costs on the other. But instead of

looking at the decision to produce one more unit of output as we've done until now, we will look at the decision to hire one more worker. This decision will change the firm's revenue, since adding a worker will increase the firm's output, and the additional output will be sold. But it will also change the firm's costs, since the additional worker must be paid.

THE FIRM'S EMPLOYMENT DECISION WHEN ONLY LABOR IS VARIABLE

In Table 1, we return to a firm we first met in Chapter 6—Spotless Car Wash. The first two columns in the table are reproduced from Table 1 of that chapter and introduce nothing new. Column 1 shows different numbers of workers that Spotless can hire, Column 2 the quantity of output produced each day.

Column 3 shows the marginal product of labor (*MPL*)—the additional output produced when *one more* worker is hired. For example, when the firm hires the third worker, output rises from 90 to 130, so the *MPL* for this change is 40.

The marginal product of labor was discussed in Chapter 6, but it was not included in the table there. In this chapter, however, we'll be very explicit about the marginal product of labor and how it behaves. In particular, notice that in Table 1, the marginal product of labor *increases* as employment rises from 0 to 1 to 2 workers. This tells us that, from 0 to 2 workers, Spotless has *increasing returns to labor*. Beyond two workers, however, additional employment causes the marginal product of labor to *decrease*, and Spotless has *diminishing returns to labor*. (See Chapter 6 if you need a refresher on returns to labor.)

Column 4 shows the price Spotless can charge for each car wash. The price remains constant at \$4, no matter how much output is produced, telling us that

**DATA FOR SPOTLESS CAR WASH
(PERFECTLY COMPETITIVE PRODUCT AND LABOR MARKETS)**

TABLE 1

(1) Quantity of Labor	(2) Total Product (Cars Washed per Day)	(3) Marginal Product of Labor (<i>MPL</i>)	(4) Price per Car Wash	(5) Total Revenue	(6) Marginal Revenue Product (<i>MRP</i>)	(7) Wage (<i>W</i>)
0	0		\$4	\$0		
1	30	30	\$4	\$120	\$120	\$60
2	90	60	\$4	\$360	\$240	\$60
3	130	40	\$4	\$520	\$160	\$60
4	155	25	\$4	\$620	\$100	\$60
5	172	17	\$4	\$688	\$68	\$60
6	185	13	\$4	\$740	\$52	\$60
7	190	5	\$4	\$760	\$20	\$60

Spotless is a competitive firm in its product market. It can wash all the cars it wants without decreasing the price. The fifth column lists the firm's total revenue for each number of workers, found by multiplying the quantity (column 2) by the price (column 4).

Marginal Revenue Product (MRP). Now look at column 6, which introduces something new. This column lists the *increase in the firm's revenue from hiring an additional worker*. For example, when the firm hires the fourth worker, its daily revenue rises from \$520 to \$620, an increase of \$100. This increase in revenue is called the *marginal revenue product of labor*:

Marginal revenue product (MRP)

The change in revenue from hiring one more worker.

The marginal revenue product (MRP) of labor is the change in total revenue from hiring one more worker. Mathematically, MRP is calculated by dividing the change in total revenue (ΔTR) by the change in employment (ΔL): $MRP = \Delta TR / \Delta L$.

We can also think of the *MRP* in another way: When the firm hires another worker, the rise in output is given by the marginal product of labor. The increase in revenue will be the amount of money for which the additional output can be sold, or price (P) times the marginal product of labor (*MPL*). For example, when moving from 3 to 4 workers, Spotless washes 25 more cars ($MPL = 25$), at \$4 each, so revenue will rise by $MPL \times P = 25 \times \$4 = \$100$. This is the same value we obtained for *MRP* earlier, using $\Delta TR / \Delta L$.

When output is sold in a competitive product market, the MRP can be calculated by multiplying the marginal product of labor by the price of output: $MRP = MPL \times P$.



Remember that marginal productivity (and the marginal revenue product derived from it) is a characteristic of *production*, not a characteristic of an individual worker. The *MPL* tells us how much a firm's output will increase when one more worker is hired. It is easy to confuse this with an individual's *personal* productivity, which is based on skill and effort. To see the difference, consider this example: Suppose you can type 90 words a minute with no mistakes—your personal productivity as a typist is very high. If a word-processing firm hires you, by how much will its output of finished manuscripts increase? That depends. Suppose the firm has just five computers. If you were, say, the fifth worker hired, you would get your own computer, and production would increase considerably. But if you were the twentieth worker hired, you would have to share a computer with perhaps three other workers, and much of your time would be spent waiting for a machine; output would not rise much at all. Even though your own skills are the same in both cases, the output you would *add* to the firm if hired—the marginal productivity of labor at the firm—would be quite different.

This explains why the *MRP* values in the table first rise and then fall. As you've learned from Chapter 6, we generally expect increasing returns to labor (rising *MPL*) at very low levels of employment, followed eventually by diminishing returns to labor (falling *MPL*) at higher levels of employment. Since P remains constant when output is sold competitively, the behavior of $MRP = P \times MPL$ mirrors that of *MPL*, first rising and then falling.²

² If output is sold under conditions of *imperfect* competition, *MRP* will *not* equal $P \times MPL$. Although hiring another worker will still increase output by the *MPL*, the firm must drop the price in order to sell the additional output. In this case, hiring another worker will increase the firm's revenue by the additional output produced (*MPL*) times the additional revenue per unit of output (*MR*), so $MRP = MPL \times MR$.

The Cost of an Additional Worker. We have just seen how hiring an additional worker will change the firm's revenue. But how will it change the firm's cost? Remember that a competitive firm in the labor market is a wage taker. In Table 1, the market wage rate for car washers is \$60 per day, so each time Spotless hires an *additional* worker, its total cost per day will *rise* by \$60.

The Profit-Maximizing Employment Level. Now that we know how hiring another worker changes the firm's cost and its revenue, we can put this knowledge together to see how a firm decides how many workers to employ. Recall (from Chapter 7) the marginal approach to profits:

The marginal approach to profit states that a firm should take any action that adds more to its revenue than to its cost.

When we apply this approach to the firm's employment decision, the action facing the firm is *hiring one more worker*. Since that decision will add *MRP* to the firm's revenue each day, and the daily wage *W* to its cost, the firm will earn the highest possible profit by following this simple guideline:

Hire another worker when $MRP > W$, but not when $MRP < W$.

Let's apply the guideline to Spotless Car Wash. When going from 0 to 1 worker, revenue rises by \$120 ($MRP = \120) and costs rise by \$60 ($W = \60). Since revenue rises more than costs ($MRP > W$), hiring this first worker will add to the firm's profit. The same is true when the second, third, fourth, and fifth workers are hired. (Verify this on your own.) But in moving from the fifth to the sixth worker, $MRP = \$52$, while $W = \$60$. Since $MRP < W$, the firm should *not* hire the sixth worker; it should stop at the fifth. We have found the firm's profit-maximizing level of employment: five workers.

We can understand Spotless's employment decision even better by graphing the marginal data from Table 1, as we've done in Figure 2. As usual, marginal values are plotted *between* employment levels, since they tell us what happens as employment changes from one level to another. The value of *MRP* first rises and then falls as employment changes, so the *MRP curve* in the figure first slopes upward and then downward. The wage rate—the cost per day of hiring the additional worker—is always the same, as shown by the horizontal line at \$60.

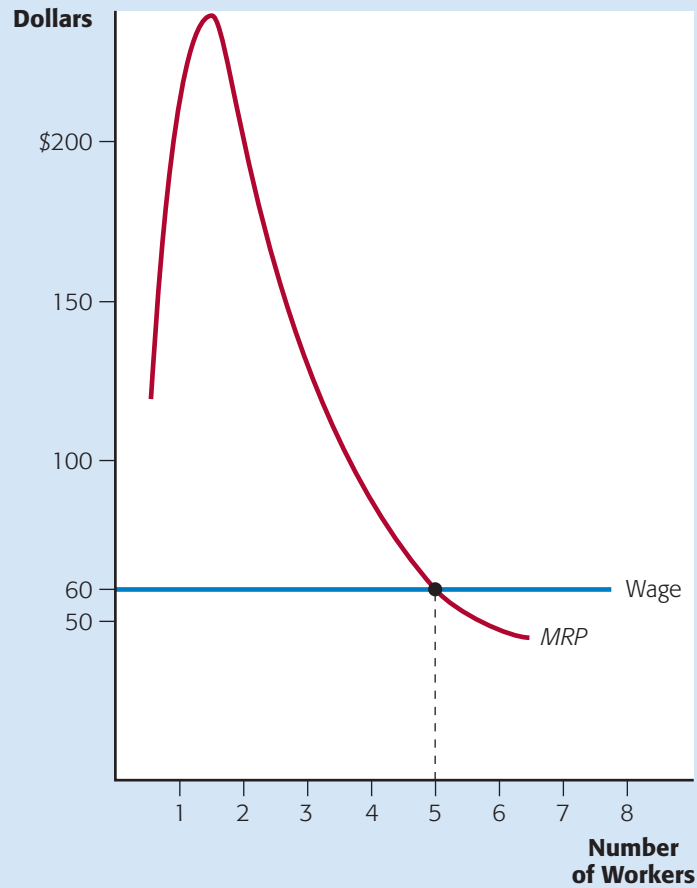
As long as employment is less than five workers, the *MRP curve* lies above the wage line ($MRP > W$), so the firm should hire another worker. But suppose the firm has hired five workers and is considering hiring a sixth. For this move, the *MRP curve* lies *below* the wage line. Since $MRP < W$, increasing employment would *decrease* the firm's profit. The same is true for every increase in employment beyond five workers: In this range, the *MRP curve* always lies below the wage line, so the firm will decrease its profits by hiring another worker. Using Figure 2, we see that the optimal employment level is five workers, just as we found earlier using Table 1.

The profit-maximizing number of workers—five—is the employment level closest to where $MRP = W$ —that is, where the *MRP curve* crosses the wage line. The reason for this is straightforward: For each change in employment that *increases* profit, the *MRP curve* will lie above the wage line. The first time that hiring a worker *decreases* profit, the *MRP curve* will cross the wage line and dip below it.

FIGURE 2

The firm should take any action that adds more to revenue than to cost. In a labor market, it should continue increasing the size of its workforce as long as the marginal revenue product of labor (*MRP*) exceeds the wage rate. The profit-maximizing level of employment for Spotless Car Wash is five workers.

THE PROFIT-MAXIMIZING EMPLOYMENT LEVEL



This observation allows us to state a simple rule for the firm's employment decision:

*To maximize profit, the firm should hire the number of workers such that $MRP = W$ —that is, where the *MRP* curve intersects the wage line.³*

The Firm's Labor Demand Curve. In Table 1, the wage rate the firm had to pay was \$60 per day. But what if the wage had been different, say, \$50 per day? As

³ There is one proviso, however: Profits are maximized only if the *MRP* curve crosses the wage line *from above*—that is, if we are on the *downward-sloping* portion of the *MRP* curve. To prove this to yourself, draw an example in which the upward-sloping portion of the *MRP* curve crosses the wage line from below. Notice that the *MRP* will always be greater than the wage to the right of the crossing point, so it will always pay for the firm to *increase* employment beyond the crossing point. From now on, the diagrams in this chapter will show only the downward-sloping part of the *MRP* curve, since this is the only part used by the firm to make its employment decision.

you can verify on your own, at this lower wage rate, the firm would have hired six workers instead of five. The optimal level of employment will always depend on the wage rate.

Figure 3 shows what happens at the typical firm as the wage rate varies. For each wage rate, the optimal level of employment, where $MRP = W$, is found by traveling horizontally over to the MRP curve and then down to the horizontal axis. For example, with a wage rate of W_1 , the firm will want to hire n_1 workers. If the wage drops to W_2 , the optimal level of employment rises to n_2 . As the wage rate drops, the firm moves along its MRP curve in deciding how many workers to hire. This is why we call the downward-sloping portion of the MRP curve the *firm's labor demand curve*:



You have now learned two different rules for the firm to follow in maximizing profit. The firm uses $MR = MC$ to find the profit-maximizing output level and $MRP = W$ to find the profit-maximizing employment level. But this suggests a potential problem: What if the $MRP = W$ rule and the $MR = MC$ rule lead to different conclusions? For example, what if the $MRP = W$ rule tells the firm to hire 5 workers, which implies a total output of 172 car washes, but the $MR = MC$ rule tells the firm to wash 185 cars?

In fact, this can never happen, because our two rules are really just two different ways of viewing the *same* firm decision. They will always lead to the same decision.

Let's see why. Remember that hiring another worker will increase the firm's output and therefore its revenue. If hiring that worker increases revenue more than it raises the firm's cost ($MRP > W$), it must be that each unit of additional *output* produced by the new worker adds more to revenue than to cost ($MR > MC$). Hence, whenever $MRP > W$, we know that $MR > MC$. Thus, whether the firm follows the profit-maximizing output rule or the profit-maximizing employment rule, it will always end up hiring the same number of workers, and producing the same level of output.

When labor is the only variable input, the downward-sloping portion of the MRP curve is the firm's labor demand curve, telling us how much labor the firm will want to employ at each wage rate.

THE FIRM'S LABOR DEMAND CURVE

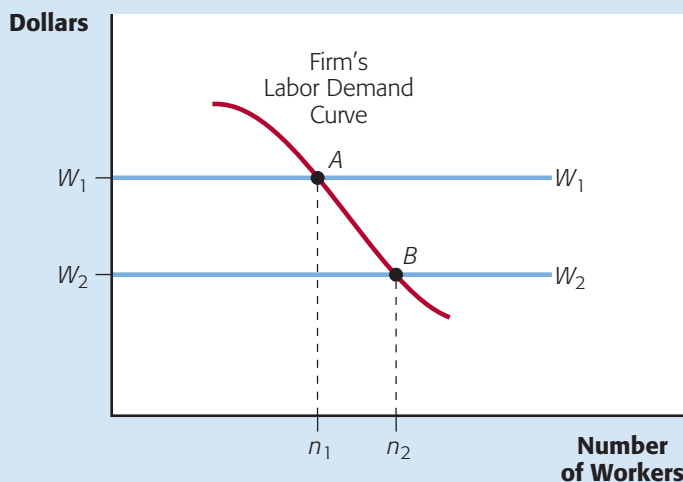


FIGURE 3

As the wage rate varies, the firm moves along its MRP curve in deciding how many workers to hire. As a result, the downward-sloping portion of the MRP curve is the firm's labor demand curve. It shows how many workers will be demanded at each wage rate.

THE FIRM'S EMPLOYMENT DECISION WHEN SEVERAL INPUTS ARE VARIABLE

So far, we've limited ourselves to a firm that can change only its employment of labor. Other inputs were assumed to be fixed in quantity. But sometimes, the firm must decide on the quantities of two or more inputs simultaneously. For example, with a long-run planning horizon, a firm will view *all* of its inputs as variable, including not just its labor, but also its capital equipment and the size of its plant. Even over the short run, a firm may be able to vary some types of capital (like hand tools), raw materials, and energy, as well as labor. When there is more than one variable input, can we still use the concepts we've developed for the single-variable case? Yes we can, but with some adjustment.

The optimal level of employment will *still* satisfy the condition that $MRP = W$; after all, if $MRP > W$, the firm can always increase its profit by hiring another worker, and it will do so. Thus, even when there is more than one variable input, the $MRP = W$ rule will still be satisfied when the firm is doing the best that it can do. What is different is this: With more than one variable input, the MPL —and therefore, the MRP of labor—will depend on the quantities of *other* inputs the firm uses. This requires us to derive the firm's labor demand curve in a slightly different way, as you are about to see.

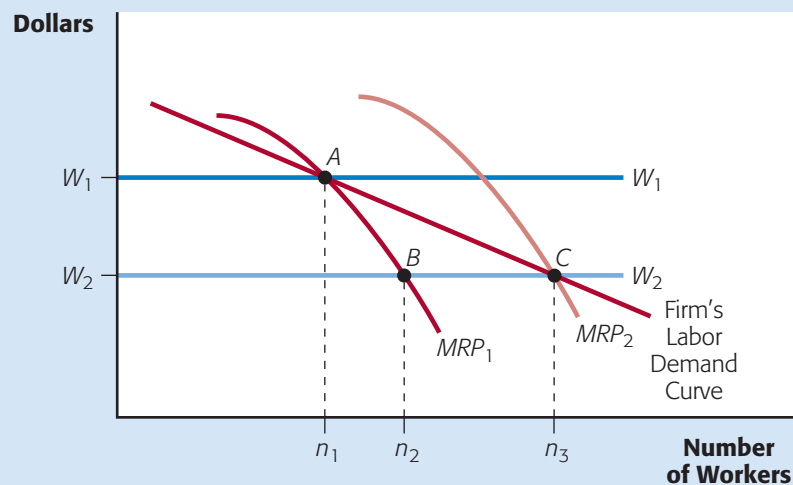
The Firm's Labor Demand Curve with More Than One Variable Input. Figure 4 shows the situation facing Spotless Car Wash when it can vary *two* inputs: labor and capital (automated car wash lines). When the wage rate is W_1 , the firm will employ n_1 workers. This puts us at point A in the figure.

Now the wage drops to W_2 . What will the firm do? If it can vary *only* its labor, it will move along the curve MRP_1 to point B, expanding employment to n_2 workers. This is because, along any MRP curve, the quantities of all other inputs are assumed to remain unchanged. At point B, the firm will use more labor to wash cars, but it will continue to use the same number of automated lines.

FIGURE 4

THE EMPLOYMENT DECISION WITH SEVERAL VARIABLE INPUTS

For a given set of nonlabor inputs, Spotless hires n_1 workers at a wage of W_1 —determined at point A along curve MRP_1 . If the wage falls to W_2 , the firm will want to hire more labor, but it will also want to add more capital. As it does so, the marginal product of labor will rise, and the MRP curve will shift to MRP_2 . At wage W_2 , it now hires n_3 workers—at point C. Connecting points such as A and C yields the firm's labor demand curve in this case.



But if the firm can vary capital along with its labor, things will be different. Once again, suppose the wage drops to W_2 . The firm, as before, will want to hire more labor and wash more cars. Now, however, it is free to choose the *least-cost* combination of inputs to produce its new, higher level of output. This combination will generally include not only more labor, but also more capital than before.

With more capital to work with, labor will be more productive. Therefore, hiring another worker will add *more* to the firm's output than it did before. That is, the marginal productivity of labor will increase, and the *MRP* curve will shift upward, to MRP_2 in the figure.⁴ As a result, when the wage drops to W_2 , the firm will locate at point *C*, where its *new MRP* curve, MRP_2 , crosses W_2 . The profit-maximizing level of employment is now n_3 workers. The firm's labor demand curve, showing its optimal employment at each wage rate, will be the line connecting points like *A* and *C*. As in the single-input case we explored earlier, the labor demand curve slopes downward—at a lower wage rate, the firm will employ more workers. But notice that now, the same change in the wage rate causes a larger rise in employment. The demand for labor is more elastic—more sensitive to changes in the wage rate.

Let's recap: When the firm can vary more than one input, a drop in the wage rate will cause it to increase the quantity of labor demanded and generally increase its usage of other inputs as well. The *MRP* curve will shift upward, and desired employment will rise by *more* than when only labor can be varied. As a result, the labor demand curve looks different in the multiple-input case—it is flatter than in the single-input case—but our two most important conclusions still hold:

Whether the firm can vary just labor, or several inputs simultaneously, the optimal level of employment will satisfy the $MRP = W$ rule, and the firm's labor demand curve will slope downward: A decrease in the wage rate will cause an increase in employment.

THE MARKET DEMAND FOR LABOR

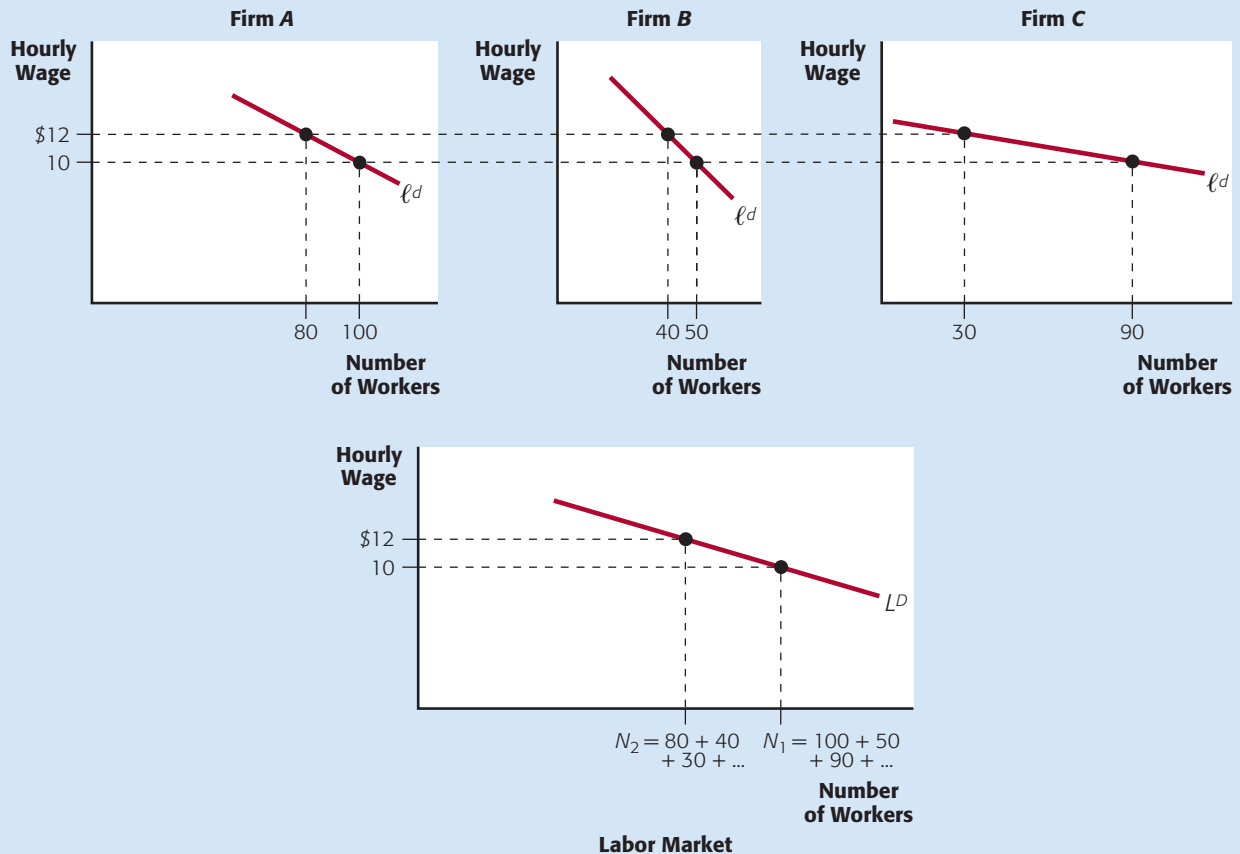
How many workers will all firms in a labor market want to employ? This question is answered by the *market* labor demand curve. Look at Figure 5, which shows the labor demand curves for three of the many firms in a labor market.⁵ At an hourly wage rate of \$10, Firm *A*'s labor demand curve, ℓ^d , tell us that it demands 100 workers, while Firm *B* demands 50 workers, Firm *C* demands 90, and so on, for all

⁴ In our example, a drop in the wage rate makes the firm decide to use *more* capital. But for some firms and some types of capital, a drop in the wage rate can cause a *decrease* in capital, as the firm decides to substitute more of the now-cheaper labor. We call this *capital-labor substitution*. Some examples of capital inputs that are *substitutable* for labor are given a bit later.

⁵ You might think that whenever the wage rate rises, and firms in the labor market respond by decreasing employment and output, the price of output will rise. But this is not necessarily so: Firms that hire in the same labor market may not sell in the same product market. (See the “Dangerous Curves” box on this topic.) In that case, a change in the wage rate will not affect the price of output. But in other cases, where a change in the wage rate affects many firms in the same product market, it may also cause the price of output to change and shift each firm's *MRP* curve. We cannot then simply sum the *MRP* curves of individual firms to obtain the market labor demand curve, since a change in wage rate (and a change in output price) will cause these curves to shift. Nevertheless, the market labor demand curve will still slope downward, and none of our important conclusions will be affected.

FIGURE 5

THE MARKET DEMAND FOR LABOR



Each firm participating in a labor market will have its own downward-sloping labor demand curve. The market demand curve is found by adding up the quantity of labor demanded by each firm at each wage rate.

of the other firms in this labor market. By adding up these numbers, we get the market quantity of labor demanded when the wage rate is \$10: $N_1 = 100 + 50 + 90 + \dots$. Now suppose the wage rate rises to \$12. Firm A will drop down to 80 workers, Firm B will drop to 40 workers, Firm C to 30 workers, and so on. With fewer workers demanded by each individual firm, the market quantity of labor demanded will shrink to $N_2 = 80 + 40 + 30 + \dots$.

Market labor demand curve Curve indicating the total number of workers all firms in a labor market want to employ at each wage rate.

The market labor demand curve tells us the total number of workers all firms in a labor market want to employ at each wage rate. It is found by horizontally summing across all firms' individual labor demand curves.

Notice that the market labor demand curve slopes downward just like the labor demand curve of each firm. If a drop in the wage rate causes each firm in the market to want to employ more workers, then total quantity demanded will increase as well.

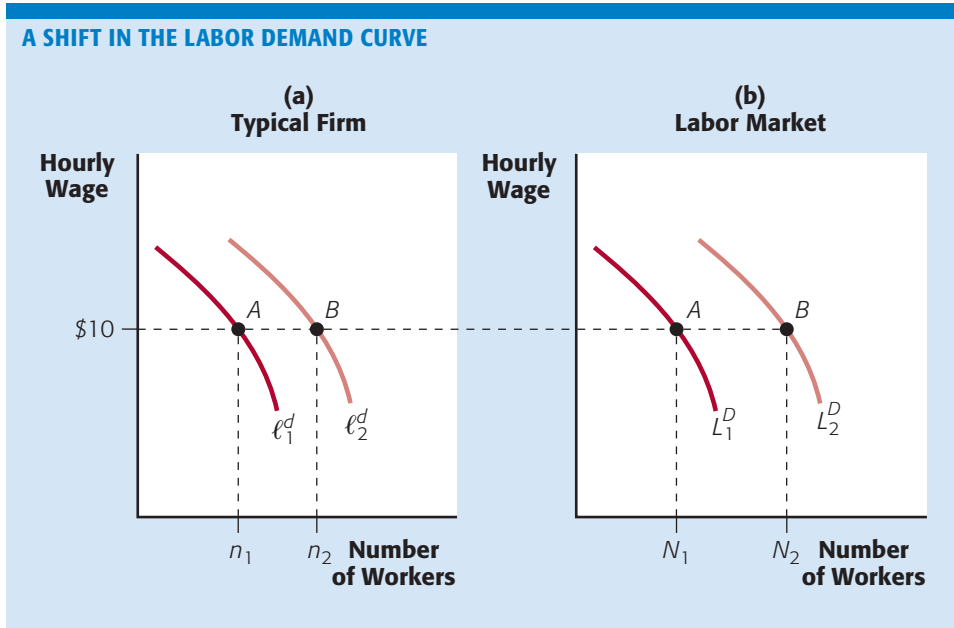


FIGURE 6

An increase in the price of output, a change in technology, or a change in the price of another input will cause the individual firm's demand for labor to increase—as from ℓ_1^d to ℓ_2^d in panel (a). In panel (b), the market labor demand curve shifts to the right as well. At any wage rate, more labor is demanded.

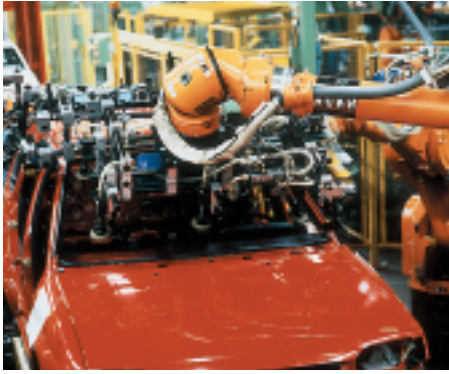
SHIFTS IN THE MARKET LABOR DEMAND CURVE

Labor markets, like other markets in the economy, are undergoing constant change, in part caused by *shifts* in labor demand curves. As we'll see in the second half of this chapter, these shifts can have dramatic effects on workers, increasing or decreasing their wage rates, or causing some to lose their jobs entirely.

We've already seen that a change in the wage rate will cause us to move *along* a labor demand curve, as in the move from point A to point B in Figure 3 (p. 315). But when something *other* than a change in the wage rate causes firms to demand more or less labor, the labor demand curve will shift. Figure 6 illustrates a general example. In panel (a), the typical firm experiences a rightward shift of its labor demand curve, from ℓ_1^d to ℓ_2^d . As a result, the market labor demand curve—the horizontal sum of all firms' labor demand curves—shifts rightward as well, from L_1^D to L_2^D in panel (b). After the shift, more labor will be demanded at any wage rate.

What factors would cause the shifts in labor demand curves, such as the ones in Figure 6?

A Change in the Price of Firms' Output. Remember that the demand for labor is a *derived* demand—it arises from demand for firms' output. Suppose demand increases in a product market, so that the price there (P) rises. Then each firm that sells output in that market will also change its employment decisions. Since $MRP = P \times MPL$, the rise in price will cause MRP to be greater at each level of employment; that is, the MRP curve of each affected firm will shift upward. Therefore, its labor demand curve will shift upward (and rightward) as well. Now, *if* many of these firms (the ones whose output price has risen) hire employees in the same labor market, then the *market* demand for labor in that labor market will increase as well. If very few firms whose price has risen hire in this labor market, there will be no perceptible change in the market labor demand curve. Thus,



Industrial robots are substitutable for less-skilled, assembly-line labor, but complementary with highly-skilled labor that programs and repairs the robots.

Complementary input An input whose utilization increases the marginal product of another input.

Substitute input An input whose utilization decreases the marginal product of another input.

the effect of a change in output price on labor demand depends on whether many firms in the labor market also share the same product market. When they do, a rise in output price will shift the market labor demand curve rightward; a fall in output price will shift the market labor demand curve leftward.

A Change in Technology. Technological progress changes the firm's production function—the relationship between its inputs and its output. One type of progress is an increase in the amount of output that can be produced with a *given* collection of inputs. For example, many firms have found that offering workers flextime—the freedom to allocate their weekly hours as they wish—makes their employees more productive. Flextime might then enable firms to produce more output with the same quantity of labor, capital, and raw materials. The *MPL* and *MRP* of labor at each employment level would increase, shifting each firm's labor demand curve—and the market labor demand curve—rightward.

Another type of technological progress occurs when an entirely new input is introduced. How will the new input affect the market demand for labor? That depends. If the new input is **complementary** with labor—*increasing* marginal product at each employment level—it will shift the typical firm's *MRP* curve (its labor demand curve) rightward, as in the shift from ℓ_1^d to ℓ_2^d in Figure 7. For example, workers in a blue-jean factory can make more jeans with sewing machines than they can by hand. If a firm brings sewing machines into its factory, the marginal product of labor will increase.

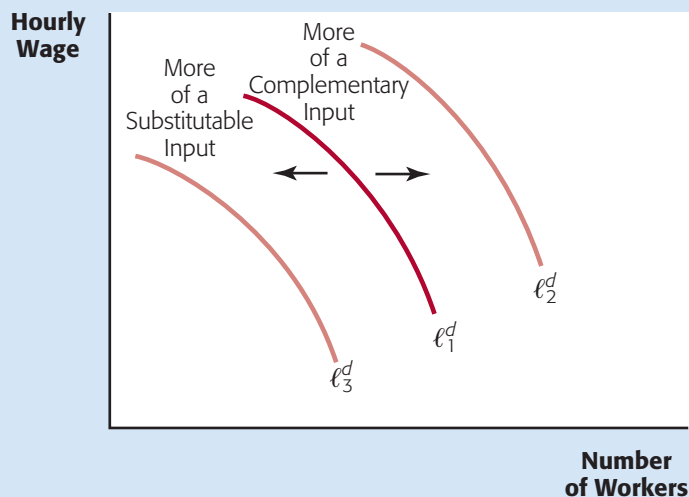
But a new input can also be **substitutable** for labor—*decreasing* marginal product at each employment level. For example, industrial robots—which tend to replace assembly workers—can decrease the marginal product of these workers. Introducing a substitutable input will shift the firm's *MRP* curve (its labor demand curve) leftward, as in the shift from ℓ_1^d to ℓ_3^d in Figure 7.

Once we know whether a new technology is complementary with or substitutable for labor, we can infer how it will affect the market demand for labor:

FIGURE 7

If a new input is introduced to the production process, the market demand for labor will shift. If the new input is complementary to labor—if it increases the marginal product of labor at each wage rate—the demand curve will shift outward. However, if the new input is a substitute for labor, the demand curve for labor will shift inward.

INTRODUCING A NEW INPUT



When many firms in a labor market acquire a new technology, the market labor demand curve will shift rightward if the technology is complementary with labor and leftward if the technology is substitutable for labor.

Determining whether a new technology is complementary with or substitutable for labor can be tricky, since firms often hire more than one type of labor. Think about what happens when retailers such as Macy's or Barnes & Noble acquire the inputs needed to sell over the Internet. Their demand for highly skilled labor—the kind that can operate and maintain hardware, and design and modify web pages—increases. But their demand for somewhat less skilled labor—salespeople, inventory clerks, and so forth—decreases, because online sales do not require these services to be performed by workers. Thus, the impact of technological progress on labor demand depends crucially on *which* labor market we are looking at—the market for high-tech workers, or the market for less-skilled salespeople.

A Change in the Price of Another Input. When the price of some input *other* than labor changes, the firm will generally adjust the quantities of *all* inputs, including labor. The impact on the labor demand curve will depend on whether the input is complementary with or substitutable for labor.

A drop in the price of computer hardware would cause retailers to hire more high-tech workers. Why? Since computer hardware is complementary with high-tech workers, the *MRP* curve (labor demand curve) for high-tech workers would shift rightward at each retail firm. If many of these retailers participate in the same high-tech labor market, then a drop in the price of computer hardware would cause a rightward shift in the demand curve for high-tech workers in that market.

But a drop in the price of computer hardware—which is substitutable for sales people—would have the opposite impact in the market for less skilled labor. As firms acquired the hardware to go on line, the marginal product of sales people would decrease, and the *MRP* curve (labor demand curve) would shift leftward at these firms, causing a leftward shift in the market labor demand curve for salespeople.

In general,

when the price of some other input decreases, the market labor demand curve may shift rightward or leftward. It will shift rightward if that other input is complementary with labor and leftward if the other input is substitutable for labor.

Interestingly, one such “other input” can be labor *from a different labor market*, such as foreign workers. Many people fear free trade agreements with low-wage foreign countries because they fear that it will make it easier and cheaper for U.S. firms to set up factories in those countries. This fear led to fierce political opposition to the North American Free Trade Agreement (NAFTA), which the United States signed with Mexico and Canada in 1993, and contributed to the protests (that led to street riots) when the World Trade Organization met in Seattle in late 1999.

Opponents of free trade claim that as U.S. firms are lured to set up production facilities in Mexico and other poor countries, jobs for American workers disappear. Their argument is that foreign labor is highly substitutable for U.S. labor, so that enabling U.S. firms to hire cheap foreign labor decreases the demand for U.S. labor. (We will dispute this argument—at least in part—in Chapter 16. But here's a hint: Is foreign labor substitutable for American labor *in general* or only in certain labor markets? Are there other labor markets in which foreign labor would be considered *complementary with* American labor?)

TABLE 2

SHIFTS IN THE LABOR DEMAND CURVE**An increase in**

demand for the firm's output
 the price of a complementary input
 the price of a substitutable input
 the number of firms in the market
 technology*

Will cause the market labor demand curve to

shift rightward
 shift leftward
 shift rightward
 shift rightward
 shift rightward if a new input is complementary with labor, leftward if the input is substitutable for labor

*An "increase" in technology here means the availability of a new input.

A Change in the Number of Firms. Within the United States, firms are continually entering and leaving local labor markets. The entry of new firms will shift the market labor demand curve rightward; exit will shift the curve to the left.

Sometimes, entry is due to the birth of an entirely new industry, as when the software industry expanded dramatically in the 1980s and the demand for labor shifted rightward in the area around Seattle. Other times, entry and exit occur when firms migrate from one local labor market to another. In the mid-1990s, firms in the computer chip industry began relocating to Oregon, shifting the demand for labor rightward in that state and leftward in the areas they abandoned.

Table 2 summarizes what you have learned about shifts in the market labor demand curve. Be careful as you look at the table; it shows only increases in each variable. A decrease in each variable would shift the labor demand curve in the opposite direction.

LABOR SUPPLY

So far, we've considered the demand side of the labor market and the behavior of firms that demand labor. Now we turn our attention to the *supply* side of the labor market and to the *households* that supply labor to firms. We begin with the individual's labor supply decision and then move on to discuss labor supply in the market as a whole.

INDIVIDUAL LABOR SUPPLY

In Chapter 5, the individual's problem was to choose the combination of goods and services that maximized his or her utility, subject to the constraints of a limited income and given prices for goods and services. Now we concern ourselves with an individual in the *labor* market who—once again—strives to maximize utility subject to constraints. Let's first look at the constraints that individuals face in a competitive labor market. Then we'll consider how the individual facing those constraints might make choices.

Identify Goals and Constraints



Individuals as Wage Takers. Think of the last time you looked for a job. Whether it was a professional job requiring considerable skill and experience or an

entry-level job such as bank teller, waiter, grocery bagger, or receptionist, there may have been hundreds—perhaps even thousands—of others looking for similar jobs in your labor market. Whether you decided to sell your labor in that market or not, your decision would be a very small drop in a very large bucket: The market wage rate would not be affected.

This characteristic—so many sellers that no single one can affect the market wage—is one of our conditions for perfect competition, and it is satisfied in most labor markets.

In a competitive labor market, each seller is a wage taker; he or she takes the market wage rate as given.

This is an important constraint on your job decision. You cannot choose your wage rate; it is determined by conditions in the market.

The Income–Leisure Trade-off. The wage rate you can earn plays an important role in a trade-off that we all face: The more time we spend enjoying leisure activities—talking with friends, going to the movies, reading, exercising, and so on—the less time we spend working and earning income. The wage rate determines the exact nature of this trade-off. For example, if you can earn \$10 per hour by working, then each additional hour of leisure time will cost you \$10 in foregone income. In a sense, \$10 is the *price* of an additional hour of leisure, since that is what you must give up, in money terms, to enjoy it.

Since different people are paid different wage rates, they will face different income–leisure trade-offs. An hour of leisure is “more expensive” to someone who earns \$100 per hour than to someone with a wage of \$10 per hour.

But in addition to differences in wage rates, there is another way that the income–leisure trade-off can differ among people: Some workers have considerable freedom to vary their weekly hours of work, and some do not.

For example, many self-employed professionals—doctors, lawyers, writers, and others—can adjust their work hours as they please, by increasing or decreasing the number of clients they serve. In addition, hourly workers can sometimes vary their hours of work by choosing to switch between part-time and full-time work or by accepting or refusing overtime. In these cases where hours can be varied, economists think about labor supply using a model of individual choice very similar to the one you learned for consumer theory in Chapter 5. However, instead of choosing the optimal combination of different *goods*, the individual chooses the optimal combination of *income* and *leisure*.

But in most labor markets, you will have relatively little freedom to vary your work hours because your employer will expect you to work a fixed number of hours—typically, eight hours a day, five days a week. In this case, your choice is not *how much* to work but rather *whether to offer your labor in a particular market*. Your choice of work hours in any labor market is constrained to 40 hours per week or zero hours per week. In this chapter, we’ll focus on *fixed-hours* labor markets like this, since they are so common in the real world.

The Labor Supply Decision. In a labor market with fixed hours, can we still view an individual as maximizing utility? It might seem that we cannot, at least not in the familiar way. If your hours are fixed, there are no marginal adjustments for you to make. Instead, you make a yes-no decision: to offer your labor services in a market, or *not* to offer them there. But even in this decision, utility maximization plays an important role: In deciding whether or not to work, or in which labor market to



Identify Goals and Constraints



Identify Goals and Constraints

supply your labor, you will always select the option that gives you the most utility. Let's explore this choice further.

Reservation Wages. One of the authors of this text, in his youth, spent six months working as an egg cleaner—cleaning the chicken droppings off fertilized eggs for eight hours a day, five days a week. It is not the most pleasant job, and chances are you are not currently planning to enter this line of work. But might you think again and decide that egg cleaning isn't all that bad if the job paid \$50 per hour? \$100 per hour? What about \$200 per hour? Surely there is *some* wage rate that would induce you to take a job as an egg cleaner. Economists call the *lowest wage rate* that would convince you to offer your labor services in a market your **reservation wage** for that labor market. Until you reach this wage rate, you are reserving your time for other uses that give you more utility—either not working at all or working in some other labor market. Whenever the wage rate in a market exceeds your reservation wage for that market, you will decide to work there. When the market wage rate is less than your reservation wage for that market, you will prefer not to work there.⁶

Reservation wage The lowest wage rate at which an individual would supply labor to a particular labor market.

MARKET LABOR SUPPLY

When we speak of the quantity of labor supplied in a market, we mean the number of qualified people who want jobs there. As we've seen, an individual will want to work in a market whenever the wage rate there is greater than his or her reservation wage. But because workers have different preferences over working conditions in different jobs, and different preferences for working at all, they will have different reservation wages for any particular market. For example, if you hate snakes, your reservation wage for a job as assistant snake trainer at a circus would be very high, perhaps \$200 per hour or more. If you like snakes, you might jump at the chance to work with them even at a wage of only \$10 per hour.

As the wage rate in a market rises, it will exceed more individuals' reservation wages, so more people will offer their labor in that market. Therefore,

the higher the wage rate, the greater the quantity of labor supplied.

Panel (a) of Figure 8 illustrates a **labor supply curve** in a hypothetical labor market, telling us the number of people who will want jobs there at each wage rate. In this market, the quantity of labor supplied at an hourly wage of \$10 is 1,000 workers, so we know that 1,000 people have reservation wages of \$10 per hour or less. At a wage of \$12, the quantity of labor supplied is 1,200, so we know that another 200 people have reservation wages between \$10 and \$12 per hour.

Labor supply curve Curve indicating the number of people who want jobs in a labor market at each wage rate.

SHIFTS IN THE MARKET LABOR SUPPLY CURVE

A change in the wage rate causes a movement *along* a labor supply curve, as in the move from point *C* to point *D* in Figure 8(a). But labor supply curves can (and of-

⁶ What happens if the market wage exceeds your reservation wage in more than one market at the same time? As long as preferences are *rational* (Chapter 5), this will never happen, for it would mean that you cannot decide which job-income combination is more attractive. For example, suppose the market wage for egg-cleaning jobs was \$25 per hour, and at that wage rate, egg cleaning is your most preferred job choice. Then your reservation wage for any other job *must*, by definition, be greater than the current market wage for those other jobs: You would not take any other job at its current wage as long as egg cleaning is available.

THE MARKET LABOR SUPPLY CURVE

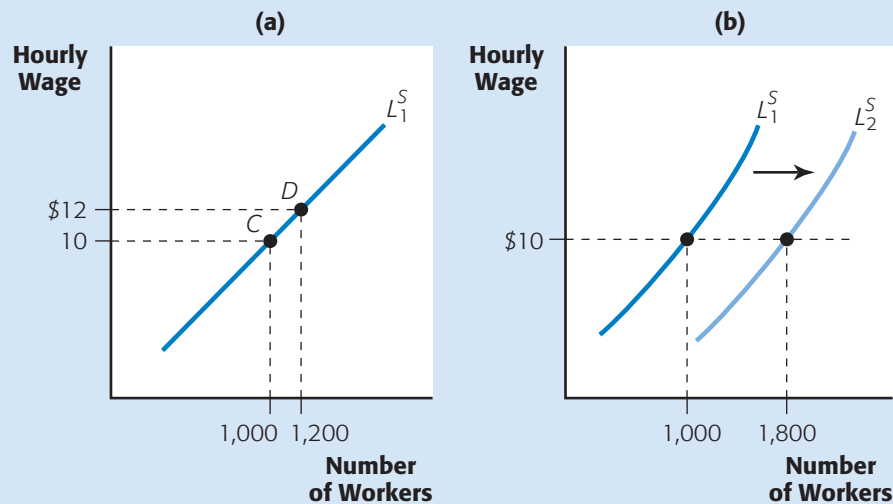


FIGURE 8

Panel (a) shows an upward-sloping market labor supply curve. A change in the wage rate causes a movement along that curve, as from point C to point D. In panel (b), the market supply curve shifts in response to a change in any nonwage determinant of labor supply. At any wage rate, more labor is supplied after the shift.

ten do) *shift*. Panel (b) illustrates an increase in labor supply in this market. Notice that, *at any given wage rate*, more people want to work in this market after the shift. For example, when the labor supply curve is L_1^S , 1,000 individuals want jobs at an hourly wage rate of \$10; after the shift to L_2^S , the number who want jobs at a wage of \$10 increases to 1,800.

What makes a labor supply curve shift? At the most general level,

a market labor supply curve will shift when something other than a change in the wage rate causes a change in the number of people who want to work in a particular market.

But let's be more specific. What, exactly, will cause a labor supply curve to shift?

A Change in the Market Wage Rate in Other Labor Markets. Imagine that you've just graduated from law school. You've always dreamed of working for a top-notch law firm. And because you did well in law school, you have good prospects of getting a job in that market, where average first-year salaries are \$105,000 (including typical bonuses). But one day, as you are sending out resumes, a friend calls you up on the phone. "Guess what," he says. "I just heard that Internet firms are looking for in-house lawyers, and are willing to pay first-year salaries of \$150,000." Upon hearing this news, you decide to go for a job at an Internet startup, instead of a law firm.

Would your behavior in this story be plausible? Absolutely. Many people will pull out of one labor market and enter another because of a widening wage differential between them. In our example, your behavior—and the behavior of hundreds of others like you—would cause the labor supply curve in the market for lawyers at top law firms to shift leftward.

And our example is *not* hypothetical. In early 2000, as Internet companies began to offer skyrocketing salaries to just-out-of-school lawyers, the number of applicants for jobs at top law firms decreased dramatically—a leftward shift in the

labor supply curve. In order to attract qualified applicants, the top law firms were forced to raise first-year salaries (including bonus) from \$105,000 to more than \$150,000 in a single year!⁷ (Note: The ultimate rise in salaries depends on the new intersection of the labor supply *and* labor demand curves, as you'll see a bit later in the chapter.)

The moral of the story is that labor supply behavior in *one* labor market may depend importantly on conditions in *other* labor markets. In general,

as long as some individuals can choose to supply their labor in two different markets, a rise in the wage rate in one market will cause a leftward shift in the labor supply curve in the other market.

Changes in the Cost of Acquiring Human Skills. To qualify for work in most labor markets, workers need special skills or training, which economists call *human capital*. This is obviously the case for highly paid professionals, like doctors, lawyers, engineers, architects, or business managers. But in most jobs you can think of—computer repair, plumbing, carpentry, language tutoring, and so on—a worker is expected to have specific skills before entering the labor market. Acquiring these skills can be costly—in time, money, or both. A change in the cost of acquiring human capital can affect the number of people who will decide to invest in training at any given wage rate and therefore shift the labor supply curve.

For example, suppose business schools across the country raised their tuition for MBA degrees by 20 percent, and there were no other changes in the economy. What would happen in the market for business managers? Initially, nothing. The suppliers of labor in this market are those who already have MBA degrees, and they would be unaffected by the tuition hike.

But now think about people deciding on careers. At any given wage rate, a career in business will look less attractive than before, now that tuition is higher. And at any given wage rate, fewer people would enroll in MBA programs. Within a few years—the time it takes to get through the program—the labor supply curve in the market for managers with MBA degrees would shift leftward, as retiring managers would not be fully replaced with new entrants.

More generally,

an increase in the cost of acquiring human capital needed to enter a labor market—say, due to an increase in school fees, fewer scholarships, or longer schooling requirements—will shift the labor supply curve leftward; a decrease in the cost of acquiring human capital will shift the labor supply curve rightward.

Population Changes. All else equal, the greater the population in any geographic area, the greater is the number of people who will want to work there. Population can grow naturally (in the United States, births exceed deaths by 1.6 million each year) or through immigration (about 900,000 more people immigrate to the United States each year than emigrate from it). In either case, population growth causes labor supply curves in both national and local labor markets to shift rightward over time.

⁷ David Leonhardt, “Law firms’ pay soars to stem dot-com defections,” *New York Times* (February 2, 2000).

TABLE 3

	Men	Married Men (Spouse Present)	Women	Married Women (Spouse Present)
1960	83.3	89.2	37.7	31.9
1970	79.7	86.1	43.3	40.5
1980	77.4	80.9	51.5	49.8
1990	76.4	78.6	57.5	58.4
1998	74.9	77.6	59.8	61.3

LABOR FORCE PARTICIPATION RATES (PERCENT OF THOSE OVER 16 WORKING OR LOOKING FOR WORK)

Source: U.S. Census Bureau, *Statistical Abstract of the United States, 1999* (Tables 657 and 658) B-37.

Labor supply curves can also shift due to migration *within* a country. Often, these shifts are a delayed response to an earlier change in relative wage rates. In the 1990s, higher wage rates in Oregon lured many workers to move there. This led to rightward shifts in labor supply curves in Oregon, and leftward shifts in regions that these workers came from.

Changes in Tastes. In any population, there is a spectrum of tastes for different types of jobs. Some part of the population will like working with numbers and hate working with people; another part will prefer just the reverse. Some like danger and excitement, whereas others like safety and routine. A change in these tastes can change people's reservation wages in a labor market and therefore change the number of people who want to work in a labor market at any given wage rate. That is, a change in tastes can shift the market labor supply curve.

A dramatic example of this is illustrated in Table 3, which shows the change in women's labor force participation from 1960 to 1998. In 1960, only 38 percent of women over 16 were in the labor force (working or looking for work), compared to 83 percent of men. By 1998, women's labor force participation rate had increased to almost 60 percent. The change was even more dramatic for married women. In 1960, only 32 percent were in the labor force; by 1998, the proportion had almost doubled, to 61 percent.

An important reason for this increase in labor supply appears to be a change in tastes. Many women changed their views of themselves and their economic role in society during this period and decided that they would prefer to work. As a result, low-cost day care centers sprang up around the country, reducing the *costs* of taking a job. Together, the change in tastes for work and the decrease in the opportunity cost of working shifted labor supply curves rightward in labor markets across the country. At any given wage rate, more women wanted jobs in these labor markets.

Changes in tastes can occur in more narrowly defined markets as well. In the midst of the social turmoil of the late 1960s and early 1970s, many college graduates wanted jobs that made a direct, visible contribution to community well-being. Certain careers—teachers, social workers, community organizers—were especially popular. As a result, the labor supply curves in these markets shifted rightward. At the same time, traditionally higher paying careers in corporate finance, marketing, and sales became relatively less popular; in these markets, labor supply curves shifted leftward.⁸ Starting in the early 1980s, and continuing today, tastes have



Lisa Barrow looks at a mother's labor supply decision in her "Child care costs and the return-to-work decision of new mothers" (http://www.ftbchi.org/pubs-speech/publications/periodicals/ep/1999/ep4Q99_3.pdf).

⁸ Since the number of workers with college degrees rises every year, the labor supply curve in most professional markets shifts rightward each year. The change in tastes discussed here actually caused labor supply curves in high-income jobs to shift rightward *more slowly* than they otherwise would have.

TABLE 4

**SHIFTS IN THE LABOR
SUPPLY CURVE**

An increase in	Will cause the market labor supply curve to
Tastes for work in a market	shift rightward
Population	shift rightward
Human capital costs	shift leftward
The wage rate in an alternative market	shift leftward

changed back: High-income jobs in business, law, and high-tech fields have become increasingly popular, reversing the labor-supply shifts of the 1960s.

Table 4 summarizes the causes of shifts in the market labor supply curve. See if you can *explain* each of the entries, rather than merely memorize them, and then reproduce this table—as well as Table 2 for labor demand—on your own.

SHORT-RUN VERSUS LONG-RUN LABOR SUPPLY

The quantity of labor supplied to a market depends crucially on the period we are considering. In general, when we adopt a longer time horizon, the quantity of labor supplied will be more sensitive to changes in the wage rate—labor supply will be more *elastic*. Why is this? We know that higher wage rates will increase the quantity of labor supplied to a market. But it often *takes time* for people to acquire the skills needed to qualify in a labor market or to move from one labor market to another.

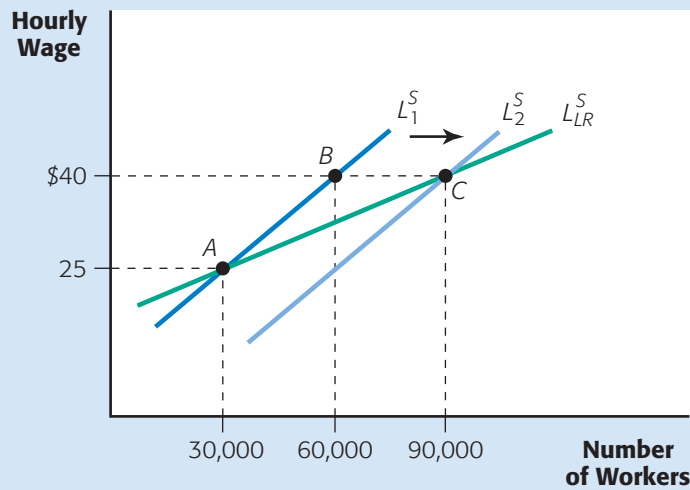
In some markets, the time needed to acquire skills can be considerable. To qualify as a lawyer requires three full years of post-college training, a college professor generally needs four years or more, and a physician requires at least seven years, and more in many specialties. Other jobs, such as secretary or construction worker, may have shorter training requirements, but it may still take considerable time before the full response to a wage change occurs.

For example, suppose the wage rate of secretaries increases. Before the full labor supply response occurs, people deciding on careers must *learn* about the change, *decide* to become secretaries, acquire the needed word-processing and other skills, prepare their resumes, find out which jobs are available, and, finally, begin looking. It is only at the last stage—where an individual begins *looking* for a job—that he or she becomes part of the total labor supply in a market. The full labor supply response to a wage-rate change can take many months or even years, depending on the adjustments required.

When analyzing a local labor market, there is another reason to expect a delayed labor supply response: It often takes considerable time to move from one local labor market to another. For example, suppose chefs' wage rates rise in Philadelphia relative to other areas of the country. Would chefs from Austin, Texas or Seattle, Washington be on the next flight to Philadelphia? Highly unlikely. Once again, there will be a variety of delays. First, chefs in other cities need to find out about the wage hike in Philadelphia. Second, they need to determine whether the higher wage rate there is permanent or temporary—few people would want to move to another town only to earn higher wage rates for a few weeks or months. Third, they must make the *decision* to move. (If you have ever been faced with this difficult choice, you can appreciate how hard it can be to decide to uproot oneself from friends and family and move to a new town.) Fourth, they need to wrap up affairs in their hometown: to give notice at their current jobs, to let the lease run out on

THE LONG-RUN LABOR SUPPLY CURVE

FIGURE 9



At a wage of \$25, 30,000 workers supply labor to this market. If the wage rises to \$40, additional people who *already* have the needed skills and who *already* live in the area will decide to supply labor to the market. Quantity supplied increases to 60,000 as a result of a movement to point *B* on the short-run labor supply curve. Over longer periods of time, however, a wage of \$40 will attract new entrants. With more people seeking work in this market, the short-run supply curve will shift to the right. Therefore if the wage remains at \$40, the quantity of labor supplied would increase beyond 60,000—to 90,000. The long-run labor supply curve is found by connecting points *A* and *C*.

their apartments or sell their homes, perhaps even to wait for their children to finish out the school year. All things considered, it could easily take years before the full labor supply response to the wage increase is completed.

To take account of these delays, it is convenient to define two periods for labor supply behavior. We define the *short run* as a period too short for people to move to a new locality or to acquire new skills. Thus, in the short run, the labor supply response to a wage-rate change comes from those who *already have the skills and geographic location* needed to work in a market.

The *long run*, by contrast, is enough time to acquire new skills or to change location. In the long run, the labor supply response to a wage-rate change includes those who will move into or out of the area and those who will acquire the skills needed to qualify in the labor market.

Figure 9 illustrates this distinction on a graph. When the wage rate is \$25, 30,000 workers supply labor in the market shown. Now suppose the wage rises to \$40. In the short run, the quantity of labor supplied will increase from 30,000 to 60,000 because more of those who *already* have the skills and who *already* live in the area will decide to work at the higher wage rate. These are people whose reservation wage in this market is greater than \$25, but no more than \$40. Thus, in the short run, we will move along the labor supply curve L_1^S , from point *A* to point *B*.

But as we proceed into the long run, the higher wage rate will attract entrants into the labor market. It will increase the number of individuals who, in skills and location, can realistically work there. As a result, the labor supply curve will shift rightward, until entry into the labor market stops. In Figure 9, this occurs when the labor supply curve reaches L_2^S , at which point all those who want to enter this

labor market at the wage of \$40 have done so. In the end, if the wage rate were to remain at \$40, the quantity of labor supplied would rise all the way to 90,000.

If we ask, “What is the *long-run* labor supply response to an increase in the wage rate from \$25 to \$40?” Our answer is “The amount of labor supplied increases from 30,000 to 90,000.” In other words, in the long run, we move from point A to point C in the figure. If we connect these two points with a line, we have the *long-run labor supply curve* labeled L_{LR}^S :

Long-run labor supply curve Curve indicating how many (qualified) people will want to work in a labor market after full adjustment to a change in the wage rate.

The long-run labor supply curve tells us how many (qualified) people will want to work in a labor market at each wage rate, after all adjustments have taken place. Specifically, all those who want to acquire new skills or who want to move to another location have done so.

Notice that, for a wage increase from \$25 to \$40, the long-run labor supply curve (L_{LR}^S) is more wage elastic than the short-run labor supply curve (L_1^S). That is, when the wage rate increases by a given percentage, labor supply rises by a greater percentage in the long run than in the short run. This will always be the case, because when the wage rate increases, the long-run labor supply response includes all those who will enter the labor market in the short run, *plus* the *additional* people who will enter the market in the long run. Thus,

the long-run labor supply response is more wage elastic than the short-run labor supply response.

Getting It Wrong: Ophthalmologists in Canada. The failure to recognize that labor supply is more elastic in the long run than the short run led to a serious mistake by Canadian policy makers in the mid-1980s. Canada has a national health insurance program that sets fees for medical services. As part of a cost-cutting measure, the system’s administrators decided to reduce the fees paid to ophthalmologists for routine eye care, to bring them more in line with optometrists’ fees, which were lower.

The reasoning was as follows: If optometrists were willing to provide the service at a low fee, then ophthalmologists should be willing to do the same. After all, ophthalmologists have already paid for their training, so these are *sunk* costs (see Chapter 6), irrelevant to any current decision. The *current* costs for conducting eye exams are the same for both ophthalmologists and optometrists. Therefore, ophthalmologists’ eye exam fees could be cut, and there should be very little change in the number of eye exams ophthalmologists offer to perform.

For a while, the policy seemed to work: Ophthalmologists grumbled, but continued to provide routine eye care to their patients. But after several years, a funny thing happened: The number of ophthalmologists declined, rather dramatically, and suddenly Canada’s health care administrators became concerned about having too few ophthalmologists.

The administrators’ mistake was to view their policy through a short-run lens only, when they should have been worried about the long run as well. From a short-run view, the labor supply response to any change in the wage rate is limited to those who already have the training. Since these doctors’ training costs are sunk costs, they would indeed continue to practice with low fees until they retired. In the short run, then, the labor supply curve might resemble L_2^S in Figure 9—not very wage elastic. When the wage rates of ophthalmologists were cut—say, from \$40 to \$25 as in the figure—the number practicing fell very little, along the curve L_2^S .

But taking a long-run view, we also consider the labor supply response among those who *could* acquire the human capital needed to qualify in the labor market, in this case, those *still deciding on a career*. To become an ophthalmologist, one must attend medical school for four years, pursue further specialized training in disorders of the eye, and serve a few years of residency. An optometrist, by contrast, needs only two years of post-college training. Thus, a higher wage rate (higher than optometrists') is needed to attract potential entrants into the ophthalmology labor market. When eye exam fees were equalized, the average income of ophthalmologists declined, and—over time—the short-run labor supply curve began shifting leftward. (In Figure 9, imagine a shift from L_2^S to L_1^S .) With the wage still at \$25, the number of ophthalmologists dropped further—to 30,000 (point A) in the figure.

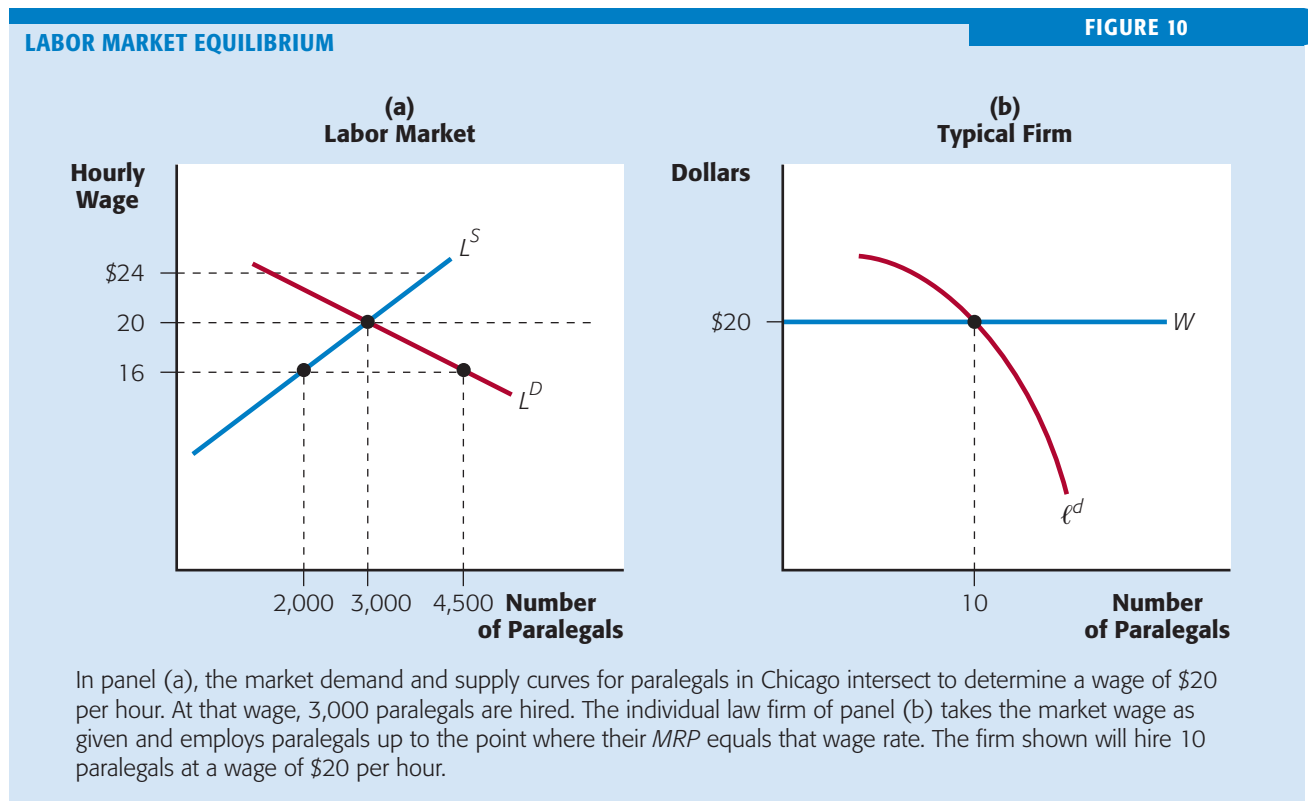
Once again, notice that the long-run labor supply curve L_{LR}^S is more wage elastic than the initial short-run supply curve (L_2^S in our story). Accordingly, the labor supply response was much greater in the long run than the administrators, thinking about the short run, had anticipated. Ultimately, the Canadian government was forced to reverse course and restore higher fees for ophthalmologists.

LABOR MARKET EQUILIBRIUM



Find the Equilibrium

Figure 10(a) illustrates the market for paralegals in the Chicago metropolitan area. (Paralegals are professionals with legal training, but no law degree, who assist lawyers.) The equilibrium in this market occurs where the supply and demand curves intersect. The equilibrium wage is \$20 per hour, and equilibrium employment



is 3,000 paralegals. Panel (b) illustrates a typical firm in this market. The firm takes the market wage rate of \$20 as given and hires the profit-maximizing number of paralegals—10—where its labor demand curve cuts the wage line, W .

How do we know that the equilibrium in this market is as we've described it? And how can we have confidence that the market will, indeed, reach this equilibrium? Suppose the wage rate is below \$20—say, \$16. Then law firms in Chicago would want to hire 4,500 paralegals, but only 2,000 people would want to work in this labor market. Competing with each other to hire paralegals, firms would drive the wage rate up. Therefore, \$16 cannot be the equilibrium wage rate in this market, since at that value, it would automatically begin rising. Once the hourly wage hit \$20, however, there would be no incentive for any firm to offer a higher wage, since every firm could hire all the paralegals it wanted at \$20.

Similarly, suppose the hourly wage were *greater* than \$20 (say, \$24). As you can see in panel (a) of the figure, at any wage rate greater than \$20, more people would want to work as paralegals than firms would want to hire. Firms would discover that they can pay less and still hire all the paralegals they want, so the hourly wage would begin to drop. Once again, when the wage dropped all the way to \$20, there would be no reason for any further change.

The forces of supply and demand will drive a competitive labor market to its equilibrium point—the point where the labor supply and labor demand curves intersect.

WHAT HAPPENS WHEN THINGS CHANGE?

Labor markets—like product markets—are in continual flux. A variety of events can cause the labor demand curve to shift (see Table 2 on p. 322) or the labor supply curve to shift (see Table 4 on p. 328). In this section, we explore how these shifts affect the equilibrium in a labor market.

What Happens When
Things Change?



A CHANGE IN LABOR DEMAND

What happens when a labor demand curve shifts? In Figure 11(a), the labor market is initially in equilibrium at point A , where the demand curve L_1^D intersects the short-run labor supply curve L_1^S . The equilibrium hourly wage is \$20, and equilibrium employment is 5,000. Panel (b) shows the typical firm facing the market wage of \$20 and maximizing profit by hiring 50 workers.

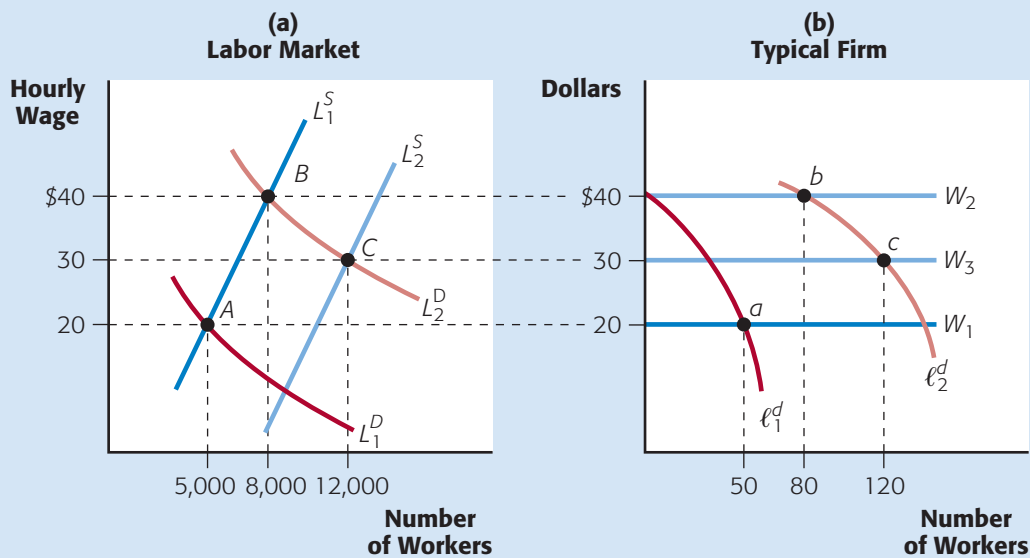
Now suppose that each firm in this labor market experiences a rightward shift in its labor demand curve. (What might cause this to happen? Look back at Table 2.) Each firm will want to hire more labor at any wage, so the market labor demand curve in panel (a) shifts rightward, driving the market wage up to \$40. This is a movement *along* the short-run labor supply curve L_1^S , from point A to point B , with employment rising to 8,000.

Meanwhile, the typical firm takes the new, higher wage of \$40 as a given. It decides to employ 80 workers. Who are the additional workers supplying labor in this market? Since this is the short run, they are individuals who are already qualified (in skills or geographic location) to work there and whose reservation wages are greater than \$20, but not greater than \$40.

But this is not the end of the story. In the long run, the higher wage rate will attract new entrants into the labor market—people who will acquire the needed train-

A CHANGE IN LABOR DEMAND

FIGURE 11



In panel (a), the market is initially in equilibrium at point A. The wage is \$20, and employment is 5,000. This equilibrium is disturbed by a rightward shift of each firm's labor demand curve. Each firm will want to hire more workers; as all the firms do, the market demand curve shifts right, driving the wage upward to \$40 at point B. Market employment rises to 8,000, and the typical firm hires 80 workers. In the long run, the higher wage will attract additional workers. Their entry shifts the market supply curve to the right; the wage falls to \$30 at point C. With the labor market once again in long-run equilibrium, market employment is 12,000, and the typical firm employs 120 workers.

ing or move to a new location. The population of qualified potential workers will increase, shifting the labor supply curve rightward.

This may seem a bit confusing. Aren't we looking at a labor *demand* shift in this section? Indeed, we are. But since the demand shift causes the wage rate to rise, and the higher wage rate attracts new workers into the market after some time, the short-run labor supply curve will eventually shift as well:

In the short run, a shift in labor demand moves us along a short-run labor supply curve. In the long run, the resulting increase in the wage rate will cause the short-run labor supply curve to shift as well.

The rightward shift in the labor supply curve will move us down along the curve L_2^D . Employment will expand further, and the market wage rate will gradually come down. When will this movement cease? Only when entry into this labor market is no longer attractive. In our diagram, this occurs when the labor supply curve reaches L_2^S , the wage settles at \$30, market employment is 12,000, and each firm is hiring 120 workers. Notice that entry stops *before* the wage falls all the way back to its original value, \$20. Why is this?

Largely because people have different tastes. In any labor market, those who want to be there the most—who have the lowest reservation wages—will be there *already*, before the wage rate changes. When the wage rate in a market increases,

the new entrants are those who would *not* have worked there at the old rate, but who would be willing to work there at a higher rate. In the long run, the wage cannot return to its original value of \$20, for at that value, many of the new entrants would leave, causing the wage rate to rise again. When the short-run labor supply curve has stopped shifting, the market wage rate, like \$30 in the figure, will be higher than the original wage, \$20.

As you can see, the consequences for wage rates are quite different in the short run than in the long run. In the short run, there is a relatively large rise in the wage—from \$20 to \$40. Indeed, the wage rate actually *overshoots* its long-run equilibrium value. Over time, as more people are attracted into the market and the labor supply curve shifts rightward, the wage rate falls to its long-run equilibrium value.

Wage rates, like the prices of goods and services, act as market signals—leading workers to move to areas where their work is most valued. When the labor demand curve shifts, the wage rate will overshoot its long-run equilibrium value. But as the signal begins to work, the temporary overshooting of the wage rate subsides.

Now that you understand how a change in labor demand can affect wages in a labor market, go back and read the paragraph that opened this chapter. Why did the salaries of U.S. medical specialists fall in the 1990s? Largely because of a leftward shift in labor demand. Beginning in the late 1980s, consumers began switching from expensive private medical practices to lower-cost health maintenance organizations (HMOs). As a result, the demand for medical care at private practices and hospitals—which employ most specialists—decreased (or, more accurately, began growing more slowly). This began happening in the late 1980s—just as today’s new specialists were deciding to become pre-med majors in college. Their decision, of course, was based partly on the higher earnings of physicians that prevailed during most of the 1980s.

The sequence of events was exactly the reverse of those depicted in Figure 11, including the drop in wage rates suffered by specialists. (To test yourself, draw a diagram that shows what has happened to these specialists’ salaries, starting with a decrease in demand for medical specialists. If the demand curve stays put after the initial shift, do you expect salaries to rise or fall over the next 10 years? *Hint:* The wage rate will overshoot, only this time in a *downward* direction.)

What Happens When
Things Change?



A CHANGE IN LABOR SUPPLY

Shifts in labor supply typically happen slowly. A look back at Table 4 shows why. While tastes for different jobs can and do change, the changes are usually very gradual. The cost of acquiring human capital can change more rapidly, but this will not shift a labor supply curve until some time later. For example, a drop in the price of going to law school will shift the labor supply curve rightward *three years later*—when those who enter law school now finally get their degrees and begin to enter the job market. Similarly, when the wage rate in some alternative labor market changes, such as the rate in another city, it takes time for people to move from one location to another and enter a new labor market.

Nevertheless, these shifts—as gradual as they may be—are important in understanding labor market changes, especially over the long run. Let’s explore an example: a leftward shift in the supply curve for business professors that occurred during the late 1990s.

Why did the supply curve for business professors shift leftward? One reason was a decrease in the number of people who were *qualified* to teach business at the college level—that is, a decrease in the number of people with Ph.D.s in business-related subjects. To understand why, we need to go back to the early 1990s, when salaries for individuals with MBA degrees from top universities rose sharply, draining away people who might otherwise have remained in school and obtained their Ph.D. In just three years, the number of new Ph.D.s awarded in business-related subjects declined by more than 10 percent—from 1,114 in 1996 to 1,006 in 1998. Suddenly, the number of new Ph.D.s coming onto the job market was smaller than the number of Ph.D.s retiring, so there was a decrease in the pool of labor qualified to teach business. In other words, the labor supply curve shifted leftward.

But there was more. Those who *did* have their doctorates and *were* qualified to teach in business schools also faced an alternative labor market: the market for private-sector business and financial analysts with Ph.D.s. In this alternative market, wages were rising even more rapidly than in the market for MBAs. In fact, by the end of the 1990s, many Wall Street firms were luring Ph.D.s from universities by offering them salaries two to four times greater than they could earn by teaching. Not surprisingly, at any given wage rate for teaching, fewer people wanted to remain in the university. So for this reason, too, the labor supply curve shifted leftward.

Let's look at a more specific market: the market for new finance professors. In Figure 12, the decrease in supply of these professors is illustrated by the leftward shift of the labor supply curve, from L_1^S to L_2^S . At the same time, there was an increase in demand for finance faculty as MBA enrollments continued to grow. This meant that the *demand* curve for these professors shifted to the right, from L_1^D to L_2^D . The decrease in labor supply and the increase in labor demand combined to move the equilibrium from point *A* to point *B*, causing the equilibrium wage of new finance professors to rise from \$66,900 per year in 1995 to \$85,300 in 1998.

Finally, one last question about Figure 12: What will happen in the long run? That's a trick question, because we *have* been discussing the long run. In this example, as in most cases of labor supply changes, the shift itself takes place in the long

THE MARKET FOR FINANCE PROFESSORS IN THE 1990s

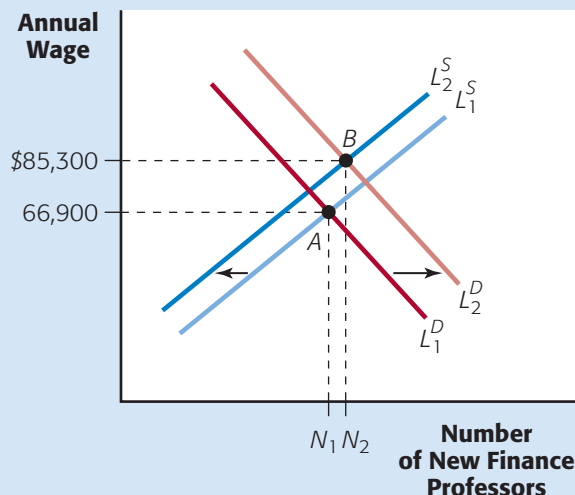


FIGURE 12

During the 1990s, the market supply of new finance professors fell, as shown by the leftward shift of the market supply curve. At the same time, increased demand for finance courses caused the demand curve to shift rightward. As a result, the equilibrium wage rose from \$66,900 to \$85,300.

run. Indeed, the labor supply curve stops shifting—at point *B*—only when all long-run adjustments have taken place. When a long-run change in labor supply is the cause of changes in the labor market, there is no separate short-run change in equilibrium to investigate.

LABOR SHORTAGES AND SURPLUSES

We sometimes hear about a shortage or a surplus of labor in some profession or trade: In the early 1990s there was a surplus of scientists, in the mid-1990s a shortage of software developers. Economists define a **labor shortage** as an *excess demand* for labor—a situation in which the quantity of workers demanded in a market is greater than the quantity supplied at the prevailing wage rate. Similarly, a **labor surplus** is an *excess supply* of workers, when the quantity of labor supplied is greater than the quantity demanded.

Labor shortage The quantity of labor demanded exceeds the quantity supplied at the prevailing wage rate.

Labor surplus The quantity of labor supplied exceeds the quantity demanded at the prevailing wage rate.

Look back at Figure 10 (p. 331). When the hourly wage is at its equilibrium value (\$20), quantities demanded and supplied are equal—there is neither an excess demand nor an excess supply of paralegals. We've argued that, in a competitive labor market, any excess demand or supply would be self-correcting. Competition for scarce jobs or competition for scarce workers would drive the wage rate to its equilibrium value. Now look back at Figures 11 and 12. There, we saw that changes in labor supply or demand cause changes in the equilibrium wage rate and employment level, but—as long as the wage can adjust—there is no shortage or surplus.

These observations suggest that a shortage or surplus can occur only when the wage rate fails to move to its equilibrium value for some reason. This is important because the media often attribute shortages and surpluses to the forces of supply and demand alone.

Shortages and surpluses in a labor market are not the natural consequence of shifts in supply and demand curves. A labor shortage will occur only when the wage rate fails to rise to its equilibrium value. Similarly, a labor surplus will occur only when the wage rate fails to fall to its equilibrium value.

Microeconomists are very interested in shortages and surpluses because they are costly for individuals, for firms, and for society as a whole. A shortage in a labor market makes it harder for firms to find workers and forces them to pay higher recruiting costs to fill job vacancies. In the end, some vacancies must remain unfilled—there are simply not enough workers to go around—which means that valuable output will not be produced.

Similarly, a surplus in a labor market makes it harder for workers to find jobs in that market. Time that could be spent earning income and producing output is instead devoted to sending out resumes, pounding the pavement, or waiting around for good fortune to strike.

Why would a wage rate sometimes fail to adjust to its equilibrium value? Toward the beginning of this chapter, it was pointed out that while the labor market is just like other markets in many respects, it also has some special features. First, the price of labor—the wage rate—is the chief source of most households' incomes. Most of us would not want to work for an employer who changed our wage rate every time there was a shift in labor demand or labor supply, because our income would change rather haphazardly. A firm that developed a reputation for frequent wage cutting would have difficulty attracting workers in the first place. It might have to pay higher wage rates, on average, than a firm with a better reputation. By

developing a reputation for wage stability, a firm has an easier time attracting labor and can earn higher profit in the long run.

Now consider what happens when the labor demand curve shifts rightward (as in Figure 11) or the labor supply curve shifts leftward (as in Figure 12). At the original wage rate, there would be a shortage of labor, which would ordinarily drive the wage rate up. But for a few weeks or even several months, a firm might resist a wage hike—even when it has unfilled vacancies—because it is reluctant to lock itself into a higher wage rate indefinitely, especially if the situation is believed to be temporary. However, if the shift in labor demand or labor supply is long lasting, firms will eventually (after a few months? years?) realize this and will then bite the bullet and fill their vacancies by paying the higher, equilibrium wage rate.

UNDERSTANDING THE MARKET FOR COLLEGE-EDUCATED LABOR

Students have many motives for attending college, but one of the most important motives is to invest in their own human capital. Put very simply, going to college will enable you to earn a higher income than you would otherwise be able to earn. How much higher? In 1998, the average high school graduate aged 25 or older earned \$19,735 per year, while the average college graduate earned \$36,708.⁹

The college *wage premium* is the percentage by which the average college graduate's income exceeds the average high school graduate's income. Figure 13 (next page) shows how this premium has behaved in recent years, separately for men and women. Notice the relative stability in the premiums during the late 1960s and through the 1970s, and the sharp increase during the 1980s and 1990s. By 1998, the premium for males had reached 72 percent, while that for women had reached 99 percent!¹⁰

The tools you've learned in this chapter can help you understand why the wage premium has behaved this way. The first step is to realize that, each year, the labor markets for those with college degrees and those with high school degrees experience changes like those shown in Figure 14 (p. 339). That is, each year, in each of these labor markets, both the labor supply curve and the labor demand curve shift rightward. The wage rate, however, may rise or fall in each market, depending on which curve shifts rightward more—the labor supply curve, or the labor demand curve.

Let's focus on the market for those with college degrees during the 1980s and 1990s. Why did the labor supply shift rightward each year? First, because there was an increase in the *proportion* of young people attending college. For example, the proportion of 18- and 19-year-olds in college has risen from 46 percent in 1980 to 62 percent in 1997. This, in turn, was partly caused by a change in tastes for college education and partly by a delayed, long-run response to the higher wage rates (due to overshooting) earned by college graduates in earlier years.

Using the THEORY



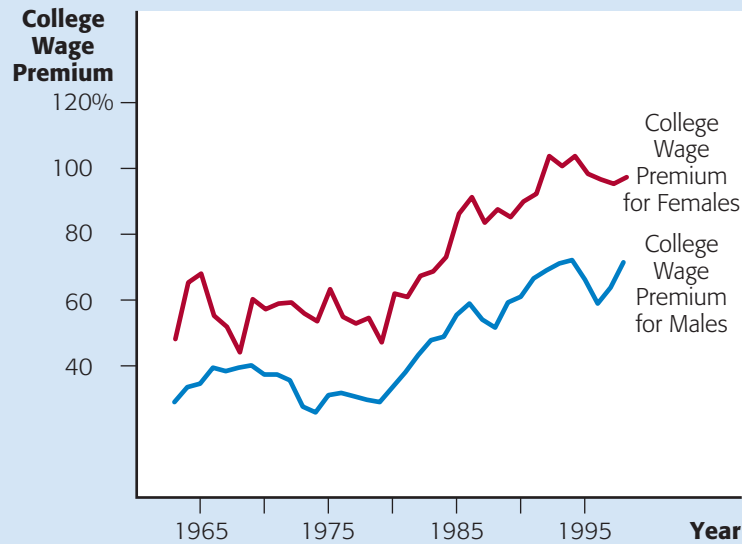
⁹ U.S. Bureau of the Census, "Measuring 50 Years of Economic Change" (Washington, DC: U.S. Government Printing Office, 1998), Table C-8.

¹⁰ This does not mean that female college graduates earn more than male college graduates. In fact, women earn less. Rather, it says that the percentage income *gap* between college and high school graduates is greater for women.

FIGURE 13

The college wage premium dropped during the late 1960s and the '70s, but turned around sharply during the 1980s and 1990s. In 1998, the premium for males reached 72 percent, while that for women reached 99%

THE COLLEGE WAGE PREMIUM



Second, the population itself increased. This would have increased the number of college graduates and shifted the supply curve rightward even if the *proportion* of young people attending college had remained the same.

Why did the labor demand curve shift rightward each year? In part, because of normal growth in the economy. As firms grow larger, and new firms are born, more labor will be demanded at any wage rate.

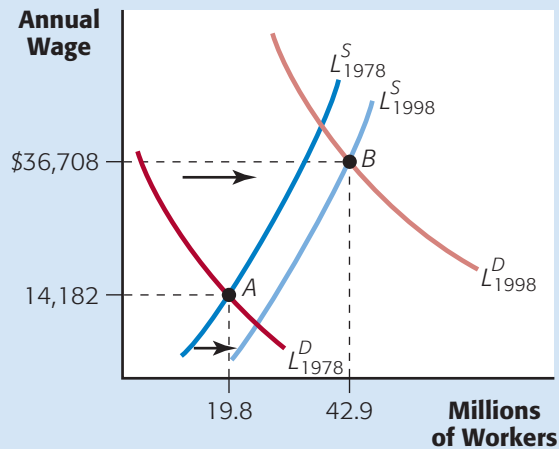
But another reason for increases in labor demand has been technological change. Over the last few decades, technological change has increased the skill requirements for many types of work. Routine jobs such as adding up numbers, handling simple requests for information over the phone, or connecting parts on an assembly line are increasingly being performed by computers and other machines. The jobs offered to *people*, meanwhile, have required greater skills than before. Instead of performing routine tasks, firms want to hire people who can write software, who can design and service computers and Web pages, and who know how to use high-tech equipment. As a result, many firms have shifted their hiring efforts toward college graduates, who are believed to have more skills and to be more capable of acquiring new skills.

Notice that, in Figure 14, the result of the shifts in labor demand and labor supply has been an increase in the yearly wage rate from \$14,182 in 1978 to \$36,708 in 1998. This is because, over the last two decades, the demand curve for college graduates shifted rightward faster than the supply curve. In the market for those with just high school diplomas (not shown), the opposite was occurring: The labor demand curve shifted rightward at about the same rate, and sometimes more slowly than, the labor supply curve. As a result, the wage rates of high school graduates have fallen.

What will happen in the future? There are two competing trends. The first trend is an acceleration in the rightward shift of the labor supply curve for college graduates. This will work to *decrease* the college wage premium. Why the acceleration of labor supply shifts for college graduates? In part, because young people are still responding to the overshooting of the wage rate in previous years. In addition, gov-

THE MARKET FOR COLLEGE-EDUCATED LABOR

FIGURE 14



Changes in the wage premium—the percentage by which the average college graduate’s earnings exceed the average high school graduate’s—can be explained in terms of shifts in labor supply and demand curves. During the 1980s and 1990s, demand increased as technological change increased skill requirements in many jobs. At the same time, supply increased because of a growth in the population of college-aged individuals and in the proportion of those individuals who chose to attend college. Because demand shifted faster than supply, the average wage rate for college graduates increased—from \$14,182 to \$36,708.

ernment subsidies to education—which have grown rapidly in the last decade, from \$20 billion in 1990 to about \$42 billion in 1999—are expected to grow further in the next decade. These subsidies—in the form of grants, work-study funding, and low-interest student loans—make college more affordable to students, and therefore increase the number who choose to enroll.

But there may also be a countervailing trend: an acceleration in the rightward shift of the labor *demand* curve for college graduates. This is due to further changes in technology. Many of the new technologies currently in the pipeline are complimentary with highly skilled labor, but substitutable for low-skilled labor. Students certainly acquire valued skills by going to college. Moreover, studies have shown that business firms invest more formal training in students with college degrees than in students with just high school degrees, further increasing the skill advantage of the college educated.

Most labor market analysts predict that, in the market for college-educated labor, the labor demand curve will shift rightward more rapidly than the labor supply curve over the next several years. Thus, the wage rate for college graduates is expected to rise. In the market for high school graduates, however, shifts in the labor supply curve are expected to outpace shifts in the demand curve. As a result, the wage premium for college students is expected to increase.

Interestingly, this wage premium for college graduates is one of the reasons behind a trend toward greater income inequality in the 1980s and 1990s. But it is not the only reason. Studies have shown that inequality has increased even within groups: greater inequality *among* college graduates and *among* high school graduates. What explains this growing income inequality?

To answer that question, we must extend and deepen our analysis of income inequality. We begin to do that in the next chapter.



Jeremy Greenwood explores the link between technology and earnings in “The Third Industrial Revolution” available at <http://www.clev.frb.org/research/review99/third.pdf>

S U M M A R Y

Firms need *resources*—land, labor, and capital—in order to produce output. These resources are traded in *factor markets* in which firms are demanders and households are suppliers. The *labor market* is a key factor market. Most households get most of their income from selling their labor. A *perfectly competitive labor market* is one in which there are many buyers and sellers, all workers appear the same to firms, and there are no barriers to entry or exit.

The demand for labor by a firm is a *derived demand*—derived from the demand for the product the firm produces. In a competitive labor market, each firm faces a market-determined wage rate. If labor is the only variable input, the firm hires up to the point at which the *marginal revenue product (MRP)* of labor—the change in total revenue from hiring one more worker—equals the wage rate. The firm’s *labor demand curve* is the negatively sloped portion of its *MRP curve*. If there is more than one variable input, the labor demand curve will be flatter, because changes in the usage of one input will affect the productivity of other inputs. Still, the firm will hire labor to the point where *MRP* equals the wage rate.

The *market demand for labor* is the horizontal sum of all firms’ individual labor demand curves. On the supply side, the upward-sloping *labor supply curve* reflects households’ *reservation wages*. A higher wage rate will attract more labor to a particular market. The market labor supply and demand curves intersect to determine the market wage rate and employment for a given category of labor.

Labor market equilibrium can change for a variety of reasons. Shifts in either curve will lead to a new equilibrium wage rate and employment combination. An increase in labor *demand* would result from an increase in the price of firms’ output, a technological change that increases the marginal product of labor, introduction of a new input that is complementary with labor, or an increase in the number of firms hiring in that market. In each case, the market labor demand curve would shift rightward, increasing both the wage rate and the level of employment. Market labor *supply* can increase as a result of a decrease in the wage rate in other labor markets, a reduction in the cost of acquiring skills needed for the labor market, an increase in the population, or a change in tastes in favor of work in that market. Such increase in labor supply would decrease the wage rate while increasing the level of employment.

It is important to distinguish between a short-run and a long-run change in labor market equilibrium. In the short run, we assume a fixed number of qualified people who are located in the geographic area of the labor market. The long run, by contrast, is a period of time long enough for workers to acquire new job skills or to move to new geographic locations. That is, in the long run outsiders can enter the market and supply labor there. After a shift in labor demand, the wage rate will generally *overshoot* its ultimate value in the short run, and then gradually move back toward its long-run equilibrium value.

K E Y T E R M S

product markets	derived demand	complementary input	long-run labor supply curve
factor markets	wage taker	substitute input	labor shortage
perfectly competitive labor market	marginal revenue product	reservation wage	labor surplus
	market labor demand curve	labor supply curve	

R E V I E W Q U E S T I O N S

1. What does it mean when we say that a firm’s demand for labor is a *derived demand*?
2. Explain how the introduction of a new input that is complementary with labor will affect labor demand. Do the same for an input that is substitutable for labor.
3. How does a firm in a perfectly competitive labor market decide how many workers to hire? What wage rate does it pay them?
4. How is the marginal revenue product of labor, *MRP*, calculated? Why is it usually stipulated that this formula is valid only “when output is sold in a competitive product

- market”? (For example, think about whether $MRP = P \times MPL$ would hold in the context of a monopoly.)
- Is there a difference between a firm's MRP curve and its labor demand curve? If so, what is the difference?
 - Why does a firm's labor demand curve become flatter when other inputs besides labor can be varied (such as in the long run)?
 - Look back at Figure 11 (p. 333). When the market wage increased from \$20 to \$40 in the short run, the typical firm increased the number of workers hired from 50 to 80. How could this be? Common sense dictates that when the wage rate increases, firms will want to hire *fewer*, not more, workers. Resolve the paradox.
 - What is a “reservation” wage? Why would your own reservation wage be different for different jobs? Why will two individuals' reservation wages generally be different when they are thinking about the *same* job?
 - In what sense is the wage rate a *signal*? How does this signal relate to short-run and long-run equilibrium in a labor market?
 - True or False. “When the labor demand curve shifts leftward, the inevitable result is unemployment.” Explain.

P R O B L E M S A N D E X E R C I S E S

- In the nation of Barronia, the market for construction workers is perfectly competitive. Explain what would happen to the equilibrium wage rate and equilibrium employment of construction workers under each of the following circumstances:
 - Young adults in Barronia begin to develop a taste for living in their own homes and apartments, instead of living with their parents until marriage.
 - Construction firms begin to use newly developed robots that perform many tasks formerly done by construction workers.
 - Because of a war in neighboring Erronia, Erronian construction workers flee across the border to Barronia.
 - There is an increased demand for automobiles in Barronia, and Barronian construction workers have the skills necessary to produce automobiles.
- The following gives employment and daily output information for Your Mama, a perfectly competitive manufacturer of computer motherboards.

Number of Workers	Total Output
10	80
11	88
12	94
13	97
14	99

A motherboard worker at Your Mama earns \$80 a day, and motherboards sell for \$27.50.

 - How many workers will be employed? How do you know?
 - Suppose the market wage for motherboard workers increases by \$5 per day per worker, but the market price of motherboards remains unchanged. What will happen to employment at the firm? Why?
- For the market for U.S. medical specialists discussed in the Section “A Change in Labor Demand” show what has happened, both in the short run and the long run, as demand for these specialists has declined. Use market labor supply and demand curves, as well as the labor demand curve of an individual hospital.
- Defense-related industries were a major employer of physicists throughout the Cold War. When tensions ended after the fall of the Soviet Union, however, defense cutbacks ensued.
 - Using graphs, illustrate the impact of defense cutbacks on the market for physicists. What happened to their equilibrium wage rate and the number employed? Assuming the market adjusted to the new equilibrium, would the cutbacks have caused unemployment among physicists? Why or why not?
 - In reality, many defense firms had long-term contracts with their professionals, locking them into specific salaries for years at a time. How does this fact alter your answer to (a)? Could it explain why unemployment occurred among physicists in the early 1990s? Explain.
- Draw a typical market labor supply curve for computer programmers over the next year. Now draw the curve looking over the next decade. What explains the principal difference between the two?

6. Suppose that dehydrated meat is an inferior good. Discuss the effects on the equilibrium wage rate and level of employment in the dehydrated meat industry of an increase in national income.
7. Re-read the section “A Change in the Market Wage in Other Labor Markets.” Then, set up two labor market diagrams—one showing the market for attorneys in law firms, and the other showing the market at Internet start-up firms. Show what would happen to the wage and employment of lawyers in both markets if the demand for Internet services *decreased*.

C H A L L E N G E Q U E S T I O N

1. Many people think that immigration into the United States—because it causes competition for jobs—will lower the wage rates of U.S. workers. Yet, even though the United States admits hundreds of thousands of immigrants each year, the average U.S. wage has continued to grow. Can you explain why? Are there any groups of workers within the economy for whom the fear of lower wages is justified? Explain.

E X P E R I E N T I A L E X E R C I S E S

1. In the late 1990s, firms and governments around the world worried about the Y2K problem—that computers would crash on January 1, 2000 because of glitches in the way some computer programs were coded. Many programmers were hired to check through software and eliminate bugs. But when January 1 rolled around and few problems were encountered, the demand for this kind of task disappeared.



Go to the Web site of the Bureau of Labor Statistics (<http://www.bls.gov>) and see if you can find data on employment and wages for computer programmers and systems analysts during the late 1990s and early 2000s. Then try to represent the situation, using a labor market diagram. Did “solving” the Y2K problem affect labor supply, labor demand, both, or neither? Review Tables 2 and 4 in this chapter. Which of the factors listed there came into play?

2. Use the *Wall Street Journal* or Infotrac to locate two articles describing the effects of technological changes on labor markets. In each case, identify a type of labor that is complementary with the new technology, and a type of labor that is substitutable for the new technology. In each of these labor markets, sketch a labor supply and demand diagram to show the effect of the change.

INCOME INEQUALITY

You sit down at a meeting with your video people and your international people and you crunch the numbers. With, say, Nicholas Cage and Ed Harris . . . you get one set of numbers. You put in [Sean] Connery's name, the numbers go way up.

A high-ranking Disney executive, explaining why Sean Connery was paid \$12 million to star opposite Nicholas Cage in The Rock.¹

Imagine, for a pleasant moment, that you are Sean Connery. Your typical workday begins in a limousine, escorting you to the site of the day's shooting, where you are fussed over by makeup artists and wardrobe staff. You memorize a few lines of dialogue, and then you stand around for several hours while the inevitable technical problems are resolved. During this time, you are doted on by assistants whose sole job is to keep you happy, who look at you respectfully, even worshipfully, and call you "Mr. Connery." Finally, you perform the day's work: maybe 10 minutes' worth of dialogue. If you make a mistake, you get another chance to get it right, as many chances as you need. And after doing this each day for four or five months, you pick up a check for \$12 million.

Now, switch gears and imagine that you have a less-rewarding job, say, a short-order cook at a coffeehouse. You spend the day sweating over a hot grill, spinning a little metal wheel with an endless supply of orders, each one telling you what you must do for the next three minutes. You cook several hundred meals that day, all the while suffering the short tempers of waiters and waitresses who want you to do it faster, who glare at you if you forget that a customer wanted french fries instead of home fries, and who call you everything but your proper name. At the end of the day, your face is covered with grease, your eyes are red from smoke, and your feet are sore from standing. And for toiling in this way day after day, for an entire year, you earn \$15,000.

And some people would consider you lucky: According to the U.S. Census Bureau, almost 36 million people live in poverty, with even smaller incomes—too small to achieve an acceptable standard of living.

We live in a country with extreme differences in wealth and income, where those at the very top ride in chauffeured limousines, while those at the bottom can barely afford to buy shoes. One reason for this is differences in wages—the subject of the first part of this chapter. Here, you will learn *why* Sean Connery earns more than a short-order cook. Indeed, you'll learn why most lawyers, doctors, and corporate managers earn more than most teachers, truck drivers, and assembly-line workers, and why these workers, in turn, earn more than farmworkers, store clerks, and

CHAPTER OUTLINE

Why Do Wages Differ?

- An Imaginary World
- Compensating Differentials
- Differences in Ability
- Barriers to Entry
- Union Wage Setting

Discrimination and Wages

- Employer Prejudice
- Employee and Customer Prejudice
- Statistical Discrimination
- Dealing with Discrimination
- Discrimination and Wage Differentials

Measuring Income Inequality

- The Poverty Rate
- The Lorenz Curve
- Problems with Inequality Measures

Income Inequality, Fairness, and Economics**Using the Theory: The Minimum Wage**

¹ *The New York Times*, September 18, 1995, p. D11.

waiters. As you'll see, we can explain much about wage differences, using the tools you learned in the previous chapter.

But labor is just one resource that people supply to the market, and labor earnings are just one source of income. Some people also own and supply other resources—such as capital or land—and they earn income from these, too. Still others—people at the very bottom of the economic ladder—can't supply *any* resources to the market. To understand income inequality more broadly, we must extend our reach beyond the labor market and look at earnings—or the lack of earnings—from *all* sources. We do this in the second half of the chapter.

WHY DO WAGES DIFFER?

At any time, some of the wage inequality we observe is *short-run* inequality. For example, suppose workers in two labor markets would earn the same wage *if* their markets were in long-run equilibrium. If adjustment to long-run equilibrium takes some time—as it often does (see Chapter 11), then wages can remain unequal for some time.

But there is also *long-run* wage inequality—differences in wages that persist after all adjustments have taken place. When thinking about income inequality, we are more concerned with long-run differences in wages, since the short-run differences, by definition, will disappear with time.

Table 1 shows average hourly earnings in several industries in 1980, 1990, and 1999. Notice the substantial differences in wages that have persisted over the last two decades. For example, the average worker in the mining industry has consistently earned almost twice as much as the average worker in retail trade. And inequality in labor incomes among *individuals* is much greater than the table shows. First, the figures are *average* figures; they ignore substantial differences in wages *within* each industry. In 1999, the highest-paid hourly workers in mining earned substantially more than \$17.09 per hour, and the lowest-paid in retail trade earned less than \$9.15. Second, the table ignores fringe benefits like health insurance and retirement benefits. This hidden income tends to be larger at the high end of the scale, increasing the degree of inequality still further. Third, the table includes payments to *hourly* workers only and excludes the (usually higher) labor incomes of supervisors and executives on a monthly salary. More accurate data—if it could be obtained—would reveal an even greater disparity in wages among U.S. workers. How can an hour of human labor have such different values in the market?

TABLE 1

AVERAGE HOURLY WAGES

Industry	1980	1990	1999
Mining	\$ 9.29	\$13.80	\$ 17.09
Construction	\$10.10	\$13.84	\$ 17.21
Manufacturing	\$ 7.42	\$10.91	\$14.04
Finance, insurance, and real estate	\$ 5.89	\$10.09	\$14.70
Retail trade	\$ 4.96	\$ 6.78	\$ 9.15
Average for private sector	\$ 6.76	\$10.08	\$13.35

Source: Bureau of Labor Statistics Data, <http://146.142.4.24/cgi-bin/surveymost>, accessed on January 2, 2000. Data are for September of each year.

AN IMAGINARY WORLD

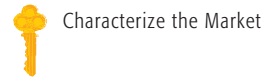
To understand why wages differ in the real world, let's start by imagining an *unreal* world, with three features:

1. Except for differences in wages, all jobs are equally attractive to all workers.
2. All workers are equally able to do any job.
3. All labor markets are perfectly competitive.

In such a world, we would expect every worker to earn an identical wage in the long run. Let's see why.

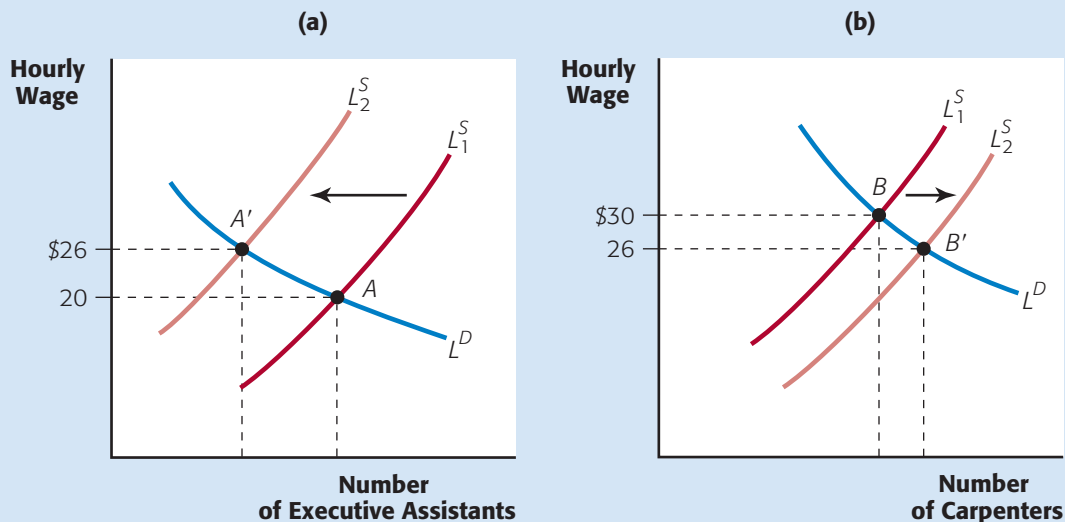
Figure 1 shows two different labor markets that, initially, have different wages. Panel (a) shows a local market for executive assistants, who earn \$20 per hour. Panel (b) shows the market for carpenters, who earn \$30 per hour. In our imaginary world, could this diagram describe the long-run equilibrium in these markets? Absolutely not.

Imagine that you are a word processor. By our first assumption, carpentry is just as attractive to you as secretarial work. But since carpentry pays more, you would prefer to be a carpenter. By our second assumption, you are *qualified* to be a carpenter, and by our third assumption, there are no barriers to prevent you from becoming one. Thus, you—and many other executive assistants—will begin looking for jobs as carpenters. In panel (a), the labor supply curve will shift leftward (exit from the market for assistants), and in panel (b), the labor supply curve will shift rightward (entry into the market for carpenters). As these shifts occur, the market wage of executive assistants will rise, and that of carpenters will fall.



DISAPPEARING WAGE DIFFERENTIALS

FIGURE 1



Initially, the supply and demand for executive assistants, in panel (a), determine an equilibrium wage of \$20 per hour—at point A. In panel (b), the initial wage for carpenters is \$30 per hour. If these markets are competitive, if the two jobs are equally attractive, and if all workers are equally able to do both jobs, this wage differential cannot persist. Some executive assistants give up that occupation—reducing supply in panel (a)—and become carpenters—increasing supply in panel (b). This migration will continue until the wage in both markets is \$26.

Find the Equilibrium



When will the entry and exit stop? When there is no longer any reason for an executive assistant to want to be a carpenter—that is, when both labor markets are paying the same wage—\$26 in our example. In the long run, the market for executive assistants reaches equilibrium at point A' and the market for carpenters at B' .

In our story, executive assistants actually *switch* jobs to become carpenters. But wages would equalize in the long run even if no one were to switch jobs. Why? If carpentry pays more, then new entrants into the labor force, choosing their trade for the first time, will pick carpentry over secretarial work. Then, as more executive assistants retire than enter the profession, their numbers will shrink. In carpentry, by contrast, there will be more new entrants than retirees, and the number of carpenters will grow. These changes will continue until wages are equal in both markets.

What is true of executive assistants and carpenters is true of *any* pair of labor markets we might choose: Doctors and construction workers, teachers and farmworkers—all would earn the same wage. In our imaginary world, different labor markets are like water in the same pool—if the level rises at one end, water will flow into the other end until the level is the same everywhere. In the same way, workers will flow into labor markets with higher wages, evening out the wages in different jobs . . . *if* our three critical assumptions are satisfied.

But take any one of these assumptions away, and the equal-wage result disappears. This tells us where to look for the sources of wage inequality in the real world: a *violation* of one or more of our three assumptions.

What Happens When Things Change?



COMPENSATING DIFFERENTIALS

In our imaginary world, all jobs were equally attractive to all workers. But in the real world, jobs differ in hundreds of ways that matter to workers. When one job is intrinsically more or less attractive than another, we can expect their wages to differ by a *compensating wage differential*:

A compensating wage differential is the difference in wage rates that makes two jobs equally attractive to workers.

To see how compensating wage differentials come about, let's consider some of the important ways in which jobs can differ.

Nonmonetary Job Characteristics. Suppose you work inside a skyscraper, and you find you could earn \$1 more per hour washing the building's windows . . . from the outside. Would you “flow” to the window washer's labor market, like water in a pool? Probably not. The higher risk of death just wouldn't be worth it.

Danger is an example of a **nonmonetary job characteristic**. It is an aspect of a job—good or bad—that is not easily measured in dollars. When you think about a career, whether you are aware of it or not, you are evaluating hundreds of nonmonetary job characteristics: the risk of death or injury, the cleanliness of the work environment, the prestige you can expect in your community, the amount of physical exertion required, the degree of intellectual stimulation, the potential for advancement . . . the list goes on and on. You will also think about the geographic location of the job and the characteristics of the community in which you would live and work: weather, crime rates, pollution levels, the transportation system, cultural amenities, and so on. What does all this suggest about differing wages in the long run? Remember that in long-run equilibrium, there is no automatic reason for the wage to change. This, in turn, requires that people have no incentive to leave one

Compensating wage differential A difference in wages that makes two jobs equally attractive to a worker.

Nonmonetary job characteristic Any aspect of a job—other than the wage—that matters to a potential or current employee.

labor market and enter another, for such changes would shift labor supply curves and change the wage in each market. But workers will be satisfied to stay in a job they consider less desirable only if it pays a compensating wage differential. The compensating differential will be just enough to keep workers from migrating from one labor market to another.

Let's see how compensating differentials figure into our example of word processors and carpenters. Look back at Figure 1. Earlier, we saw that if both jobs are equally attractive, both will pay the same wage in the long run. But now suppose that everyone prefers sedentary jobs such as word processing to physical labor such as carpentry, and it takes a \$10 wage differential in favor of carpentry to make the two jobs equally desirable. Then the two markets would settle at points *A* and *B*, where carpenters are paid a compensating wage differential of \$10 per hour to make up for the less desirable features of their jobs.

The nonmonetary characteristics of different jobs give rise to compensating wage differentials. Jobs considered intrinsically less attractive will tend to pay higher wages, other things equal.

What about unusually *attractive* jobs? These jobs will generally pay *negative* compensating differentials. For example, many new college graduates are attracted to careers in the arts or the media. Since entry-level jobs in these industries are so desirable for nonmonetary reasons, they tend, on average, to pay lower wages than similar jobs in other industries. For the same reason, people will accept lower wages when a job offers a high probability of advancement—and a higher salary—in the future. It comes as no surprise, then, that management trainees at large corporations are often paid relatively low salaries.

Of course, different people have different tastes for working and living conditions. While some prefer a quiet, laid-back work environment like a library or laboratory, others like the commotion of a loading dock or a trading floor. While most people are extremely averse to risking their lives, some actually prefer to live dangerously, as in police work or rescue operations. Therefore, we cannot use our own preferences to declare a job as less attractive or more attractive, or to decide which jobs should pay a positive or negative compensating differential. Rather, when labor markets are perfectly competitive, the entry and exit of workers automatically determines the compensating wage differential in each labor market.

This is one reason most economists are skeptical about the idea of *comparable worth*, which holds that a government agency should determine the skills required to perform different jobs and mandate the wage differences needed between them. Although this policy could correct some inequities when labor markets are imperfectly competitive, it could also introduce serious inequities of its own, since no one can know how different workers would value the hundreds of characteristics of each job. Economists generally prefer policies to increase competition and eliminate discrimination, so that the market itself can determine comparable worth.

A Digression: It Pays to Be Unusual. One implication of compensating wage differentials is that workers with unusual tastes often have a monetary advantage in the labor market. For example, only a small fraction of workers *like* dangerous jobs, such as police work. As long as the labor market is competitive, and there is relatively high demand for workers in dangerous jobs, police officers will earn more than those in other, similar jobs that have a lower risk of death or injury. But if you are one of those unusual people who *like* danger, you will earn the same

compensating wage differential as all other police officers, even though you would have chosen to be a police officer anyway.

Similarly, if you like the frigid winter weather in Alaska, if you like washing windows on the 90th floor, or if you think it would be fun to defend the cigarette industry in the media, you can earn a higher wage by putting your somewhat unusual tastes to work.

Cost-of-Living Differences. Many people would find living in Cleveland and living in Philadelphia about equally attractive. Yet wages in Philadelphia are about 10 percent higher than in Cleveland. Why? One major reason is that prices in Philadelphia are about 10 percent higher than in Cleveland. If wages were equal in the two cities, many people deciding where to live would prefer Cleveland, where their earnings would have greater purchasing power. The supply of labor in Philadelphia's labor markets would shrink, increasing the wage there, while the supply in Cleveland's labor markets would rise, driving down the wage in Cleveland. In the end, the wage difference would be sufficient to compensate Philadelphians for the higher cost of living in their city.

Differences in living costs can cause compensating wage differentials. Areas where living costs are higher than average will tend to have higher-than-average wages.

Differences in Human Capital. Suppose that you've decided to become an ontological prognosticator (no need to look it up—it's a hypothetical job). You've been informed that the job requires a Ph.D. degree and pays \$60,000 per year, and that's fine with you, since your reservation wage for this occupation is less than \$60,000. As you are applying for graduate school, you suddenly discover that ontological prognosticators must have *two* Ph.D. degrees, not one. Would your reservation wage increase? Absolutely—a second Ph.D. will require at least another four years of schooling, which means additional opportunity costs for tuition, books, and foregone income. And if your reservation wage rises above \$60,000, you would change your mind and seek another career.

This hypothetical story should convince you that higher training costs—like those facing doctors, attorneys, engineers, and research scientists—make a job less attractive. In order to attract workers, these professions must pay a wage greater than other professions that are similar in other ways, but require less training.

In terms of Figure 1 (p. 345), let's go back to our starting points, *A* and *B*, where carpenters earn a higher wage than word processors. In our imaginary world, this wage difference attracted workers to carpentry and repelled them from word processing, until wages were equal in the two markets. But now suppose that carpentry required more training than word processing. Then we would expect job shifting—and shifts in labor supply curves—to stop *before* wages were equalized. In the long run, after all adjustments had taken place, carpenters would earn more to compensate them for bearing higher training costs.

Differences in human capital requirements can give rise to compensating wage differentials. Jobs that require more costly training will tend to pay higher wages, other things equal.

Compensating differentials explain much of the wage differential between jobs requiring college degrees and those that require only a high school diploma. In 1998, the average college graduate earned an annual salary of \$43,782, while the

average high school graduate earned only \$23,594. The especially high earnings of the average dentist (\$92,350), lawyer (\$75,890), and physician (\$102,020)² reflect, at least in part, compensating differentials for the high human capital requirements—and human capital costs—of entering their professions.

The idea of compensating wage differentials dates back to Adam Smith, who first observed that unpleasant jobs seem to pay more than other jobs that require similar skills and qualifications. It is a powerful concept, and it can explain many of the differences we observe in wages . . . but not all of them.

DIFFERENCES IN ABILITY

In 1998, at the age of 35, Michael Jordan earned more than \$50 million—\$34 million playing basketball for the Chicago Bulls, and the rest for endorsing products such as Gatorade, WorldCom/MCI telephone service, and Nike shoes. Was this a compensating differential for the unpleasantness of playing professional basketball? For an unusually high risk of death on the job? Was the cost of living in Chicago hundreds of times greater than in other cities? Had Jordan, at the age of 35, spent more years honing his skills than the average attorney, doctor, architect, or engineer—or even more than the average basketball player?

The answer to all of these questions is no. We have overlooked the obvious explanation: Jordan is an *outstanding* basketball player, better than 99.999 percent of the population could ever hope to be with *any* amount of practice. This is partly because of his *endowments*—the valuable characteristics he possesses due to birth or childhood experiences but that did not require any opportunity cost on his part. In Jordan's case, these would include his natural speed, agility, and coordination. But Jordan also showed extraordinary perseverance in exploiting his talent. Together, his endowments of talent and his decision and work at exploiting them made Jordan an athlete of extraordinary ability.

While Michael Jordan may be an extreme case, the principle applies across the board. Not everyone has the intelligence needed to be a research scientist, the steady hand to be a neurosurgeon, the quick-thinking ability to be a commodities trader, the well-organized mind to be a business manager, or the talent to be an artist or a ballet dancer. This violates our imaginary-world principle that all workers have equal ability in all jobs and explains much of the wage inequality we observe in the real world.

We can understand this in terms of Figure 1 (p. 345). A wage differential between two otherwise equal jobs could persist if those working for lower wages (point *A* in panel (a)) cannot enter the high-wage market (point *B* in panel (b)) because—regardless of how much human capital they acquire—they can never perform well enough.

Many economists believe that income inequality has worsened in the 1990s. If this is true, differences in abilities may be playing an important role. Scientific discoveries and technological advances may have increased not only the skill requirements of many jobs, but also the abilities needed to acquire those skills. (For example, greater perseverance and intelligence are needed to master a word-processing program than to learn how to type.)

But Figure 1 only tells part of the story: Wages differ not only *between* different types of jobs, but also *within* job categories. And this is largely because, in any trade or profession, workers' talent, intelligence, and physical ability—and their value to firms—vary considerably.

² Salary figures for dentists, lawyers, and physicians are 1998 means, from the *Bureau of Labor Statistics* (http://www.bls.gov/oes/national/oes_prof.htm#b32000).

For example, suppose two architects have equal education and training, but architect A—being more talented—can design more innovative projects and attract twice as many high-paying clients than can architect B. Then a firm should be willing to pay architect A twice as much as architect B.

In general, those with greater talent, intelligence, or perseverance will be more productive on the job and generate more revenue for firms. Thus, firms will be willing to pay them a higher wage.

Take another look at the quote at the beginning of this chapter. Why was Disney willing to pay Sean Connery \$12 million to star in *The Rock*? The high-ranking executive explains it: “When you plug in Connery’s name, the numbers go way up.” The numbers he is referring to are *box office revenue*, about half of which flows to Disney. In large part because of his endowments of talent and looks, Sean Connery can earn more revenue for Disney than Ed Harris can—enough revenue to justify a salary of \$12 million, when Harris might have been hired for, say, \$1 million.

The Economics of Superstars. Sean Connery and Michael Jordan are examples of superstars—individuals who are almost universally regarded as the best, or among the top few, in their professions. In recent years, these individuals have included model Cindy Crawford, actress Gwyneth Paltrow, attorney Johnnie Cochran, talk show host Jay Leno, and writer John Grisham. (Whatever your own feelings about any of these people, the market—where people vote with their dollars—considers them at the very top of their professions.) Still, does outstanding ability fully explain the extremely high earnings of these superstars? Let’s see.

No doubt, NBC news anchor Tom Brokaw is better at delivering the news than most local news anchors. But is he better enough to justify a salary 20 or 30 times greater than the highest-paid local broadcaster? Similarly, Sean Connery is substantially better than the average actor, maybe even among the best. But is he better enough to justify a salary hundreds of times greater than the average? The same is observed among the top singers, doctors, attorneys, and so on: The additional earnings garnered by those at the very top seem out of proportion to their additional abilities. How can this be?

The explanation in all these cases *is* based on ability—and also by the exaggerated rewards the market bestows on those deemed the best or one of the best in a field.³ Say you like to read one mystery novel a month for entertainment. If you can choose between the best novel published that month or one that is almost—but not quite—as good, you will naturally choose the one you think is best. Only people who read *two* novels each month would choose the best



Although Sean Connery’s acting talent may not be a thousand times better than the average, he earns more than a thousand times the average actor’s salary because he is at the top of his profession.



It is tempting to think that jobs that require greater abilities or talents will *automatically* pay more than jobs that are easier and that more people can do. But this is not necessarily true. Fewer people can write good poetry than can write a good newspaper article, yet journalists earn substantially more than poets. Why? Very few people *read* poetry, and in that market, the derived demand for poets is very low relative to their supply. On the other hand, large numbers of people read newspapers. Compared to the market for poets, the derived demand for journalists is considerably higher relative to the supply. You will avoid much confusion if you remember that the equilibrium wage is determined by *both* sides of the market—supply *and* demand—rather than just one or the other.

³ See, for example, Robert H. Frank and Philip J. Cook, *The Winner Take All Society* (The Free Press: New York, 1995).

and the second best, and only those who read three will choose the top three. If most people rank recent mystery novels in the same order, then the best will sell millions of copies, the second best might sell hundreds of thousands, and the third best might sell only thousands. Even though all three novels might be very close in quality, the authors' earnings will be vastly different.

The same thing happens in the markets for rock concerts, action movies, and news broadcasts. In all these cases, where those at the top can sell their services to millions of people simultaneously, the reward for being best can be astronomical.

But this phenomenon is not limited to media stars. Suppose you needed a heart transplant, and the best surgeon is 10 percent better than the second-best surgeon. Wouldn't you be willing to pay more than a 10 percent premium to have the best, rather than the second best? The same applies to corporate executives. If Wal-Mart's chief executive officer can make decisions that are just slightly better than Kmart's, then Wal-Mart may gain significant market share over Kmart, and its earnings could be many times higher than Kmart's. This is one reason that, in the business world, small differences in perceived abilities of executives lead to huge differences in salaries.

BARRIERS TO ENTRY

In our imaginary world, there were no barriers to entering any trade or profession. The absence of barriers is an important element of our assumption that the labor market is competitive. But in some labor markets, barriers keep out would-be entrants, resulting in higher wages in those markets.

In Figure 1 (p. 345), we saw that if carpenters were paid higher wages than word processors, entry into the market for carpenters would equalize wages in the two jobs. But what if carpenters were *protected* from competition by a barrier to entry, one that kept newcomers from becoming carpenters? Then the labor supply curve in panel (b) would *not* shift rightward, and the higher wage for carpenters could persist. Going back to the analogy of water flowing to equalize the water level at both ends of a pool, a barrier to entry is like a wall in the middle of the pool. It blocks the flow, allowing one end to have a higher water level than the other.

Since barriers to entry help maintain high wages for those protected by the barriers—those who already have jobs in the protected market—we should not be surprised to find that in almost all cases, it is those already employed who are responsible for erecting the barriers. But it is not enough to simply put up a sign, “Newcomers, stay out!” The pull of higher wages is a powerful force, and preventing entry requires a force at least as powerful. What keeps newcomers out of a market, thus maintaining a higher-than-competitive wage for those already working there?

In many labor markets, occupational licensing laws keep out potential entrants. Highly paid professionals such as doctors, lawyers, and dentists, as well as those who practice a trade, like barbers, beauticians, and plumbers, cannot legally sell their services without first obtaining a license. In many states you cannot even sell the service of braiding hair without a license. In order to get the license, you must complete a long course in cosmetology and pass an exam.

The American Medical Association (AMA)—a professional organization to which almost half of American physicians belong—is perhaps the strongest example of occupational licensing as a barrier to entry. The AMA portrays itself as a vigilant defender of high standards in health care, through its regulation of medical schools, its certification of specialists, and its government lobbying. Economists

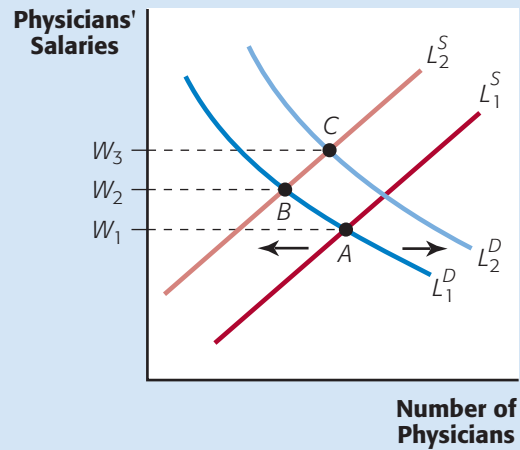


What Happens When Things Change?

FIGURE 2

Without the AMA, the labor supply and demand curves for physicians would intersect at point *A* to determine wage W_1 . AMA actions to restrict the supply of physicians have caused the supply curve to be L_2^S , which lies to the left of L_1^S . This implies a higher wage, W_2 . In addition, the AMA has sought to increase the demand for physicians' services by preventing non-physicians from competing. This shifts the demand curve to the right—to L_2^D —further increasing the wage to W_3 .

THE MARKET FOR PHYSICIANS



tend to have a much different view of the AMA. While not denying that some of its efforts do raise the quality of physicians, they see the association primarily as an instrument to maintain high incomes for doctors.

Figure 2 shows the market for physicians in the United States. In the absence of any income-raising activity, labor supply curve L_1^S would intersect labor demand curve L_1^D at point *A*, resulting in equilibrium wage W_1 . Whether this wage would be relatively high or low is not known; since 1847—when the AMA was founded—this competitive equilibrium has never been attained.

Much of the AMA's activity has been designed to decrease the *supply* of doctors. Immediately after its founding, it imposed strict licensing procedures that increased entry costs for *new* doctors; existing practitioners were exempted from the new requirements. In spite of these restrictions, there was a rapid increase in the number of physicians toward the end of the century. In response, between 1900 and 1920, the AMA closed down almost half of the nation's medical schools.⁴ These and other efforts to restrict the supply of doctors have resulted in a supply curve for physicians like L_2^S , lying to the left of L_1^S , moving the equilibrium to point *B*, and raising salaries to W_2 .

But this is not the end of the story. The AMA has also increased the *demand* for physicians' services by preventing nonphysicians from competing. Throughout its history, the association has moved aggressively to limit competition from midwives, chiropractors, homeopaths, and other health professionals. By limiting access to these alternative health professionals, the AMA increases the demand for the services of its own members. The impact of these policies has been a rightward shift in the demand curve for doctors, to L_2^D , moving the equilibrium to a point like *C* and raising salaries further, to W_3 .

(If you think maintaining high standards is the main motivation for these policies, consider this: AMA policy allows a physician to practice in *any* area of medicine, even one in which he has no specialized training. For example, a dermatologist with no training or experience in obstetrics can legally deliver a baby; a midwife

⁴ "Doctors Operate to Cut Out Competition," *Business and Society Review*, Summer, 1986, pp. 4–9.

with extensive experience might be arrested if she delivers a baby without the supervision of an M.D.)

In the late 1980s, rising health care costs led to increased public scrutiny of the AMA, and its anticompetitive practices came under heavy attack. Some restrictions were eased, and the number of doctors per 100,000 people increased from 169 in 1975 to 233 in 1990. At the same time, the Federal Trade Commission and the courts pressured the AMA to remove its ban on physician advertising. For the first time, new entrants could compete with established practices by advertising their prices and services. Not surprisingly, many physicians began to complain about falling incomes.

In the 1990s, physicians' advertising has intensified, and the number of physicians per 100,000 increased further, reaching 245 by 1997. Moreover, Health Maintenance Organizations have striven to decrease the demand for physicians' services by requiring prior approval for expensive tests and surgical procedures, especially those performed by specialists. And physicians' complaints have intensified.

UNION WAGE SETTING

A labor union represents the collective interests of its members. Unions have many functions, including pressing for better and safer working conditions, operating apprenticeship programs, and administering pension programs. But the foremost objective of a union is to raise its members' pay. Federal law prohibits a union from creating an overt barrier to entry—it is illegal for a firm to agree to hire only union members. Instead, the union negotiates a higher-than-competitive wage with the firm. But, as we know from the last chapter, at a higher wage, the firm will have a lower profit-maximizing employment level. Thus, many potential workers are kept out of union jobs because the firm will not hire them at the union wage.

The higher union wage is contrary to the interests of the employer—so why does the employer agree? Because the union has the power to strike. During a strike—when the firms' workers refuse to come to work—the firm suffers lost profits. Rather than take the risk of a strike, employers will often agree to the higher wage demanded by the union.

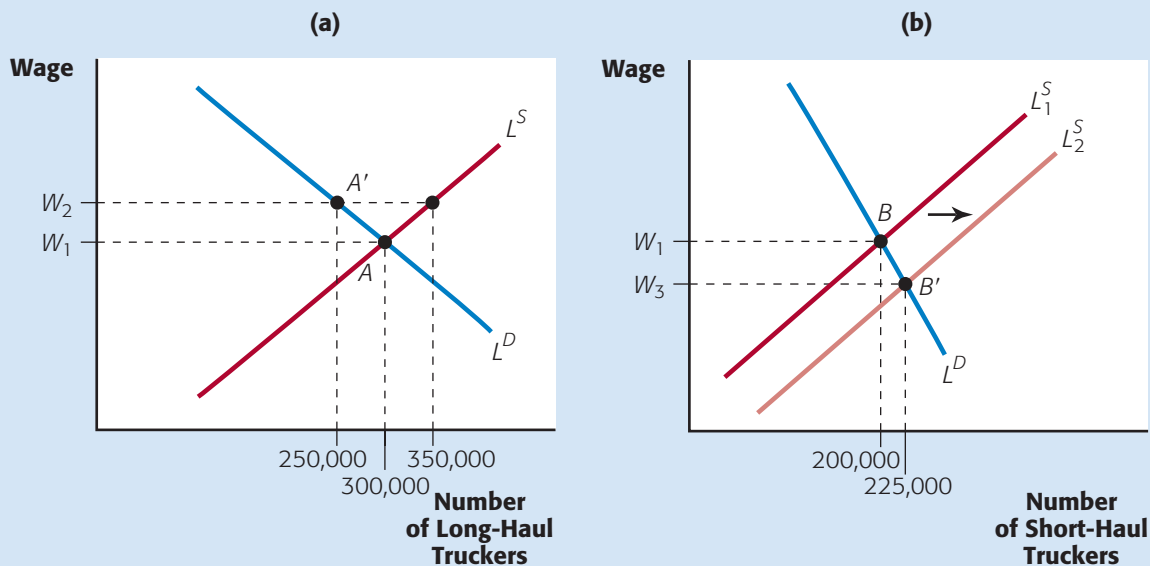
Figure 3 illustrates how unions can create wage differences. We assume that jobs in two industries—long-haul trucking and short-haul trucking—are equally attractive in all respects other than the wage rate. With no labor union, these two markets would reach equilibrium at points A and B , respectively, where both pay the same wage, W_1 .

Now suppose instead that long-haul truckers are organized into a union, which has negotiated a higher wage, W_2 , with employers. At this wage, there is an excess supply of long-haul truckers equal to $350,000 - 250,000 = 100,000$. Ordinarily, we would expect an excess supply of labor to force the wage down, but the union wage agreement prevents this. With fewer jobs available in the unionized sector, some former long-haul truckers will look for work as *nonunion*, short-haul truckers. Thus, in panel (b), the labor supply curve shifts rightward. In equilibrium, the number of short-haul truckers rises from 200,000 to 225,000, and the wage of short-haul truckers drops to W_3 . The end result is a union–nonunion wage differential of $W_2 - W_3$. Notice that only *part* of the differential ($W_2 - W_1$) represents an increase in union wages; the other part ($W_1 - W_3$) comes from a decrease in nonunion *wages*.

Through an increase in member wages and a decrease in nonmember wages, unions create a wage differential between union and nonunion labor markets.

FIGURE 3

UNION WAGE DIFFERENTIALS



In the absence of a union, the markets for short-haul and long-haul truck drivers would be in equilibrium at the same wage, W_1 . If long-haul truckers organize into a union, they can negotiate a higher wage— W_2 . At this wage, there is an excess supply of 100,000 long-haul truckers. With fewer jobs available in the unionized sector, displaced truckers seek work in the short-haul trucking industry, increasing supply there, and driving the wage down to W_3 .

In the end, how big is the union–nonunion wage differential? H. Gregg Lewis⁵ reviewed more than 200 studies that asked precisely this question and determined that between 1967 and 1979, union members earned, on average, about 15 percent more than otherwise similar nonunion workers.

Given the conflict surrounding many union–management wage negotiations, and the media attention devoted to them, this difference may seem rather small. But keep in mind that a 15 percent differential—about \$1.75 per hour at today's average wage—amounts to \$3,640 per year, and it continues year after year. After 40 years on the job, the average union member can expect to have earned about \$145,000 more than the average nonunion member, enough to put a down payment on a house *and* put a child through college with no student loans. And if each year's differential were put in the bank at 5 percent interest, it would amount to about \$465,000 after 40 years. The union–nonunion wage differential is nothing to sneeze at.

The differential has most likely declined since 1979, as unions' bargaining power has weakened. This is partly reflected in a decline in union membership: In the mid-1950s, 25 percent of the U.S. labor force was unionized; today, only about 13 percent of the labor force are union members. Nevertheless, unions still maintain a significant presence in many industries, such as automobiles, steel, coal, construction, mining, and trucking, and they are certainly responsible for at least *some* of the higher wages earned in those industries.

⁵ H. G. Lewis, *Union Relative Wage Effects: A Survey* (Chicago: University of Chicago Press, 1986).

DISCRIMINATION AND WAGES

Discrimination occurs when *the members of a group of people have different opportunities because of characteristics that have nothing to do with their abilities.* Throughout American history, discrimination against women and minorities has been widespread in housing, business loans, consumer services, and jobs. The last arena—jobs—is our focus here. While tough laws and government incentive programs have lessened overt job discrimination—such as the help wanted ads that asked for white males as late as the 1950s—less obvious forms of discrimination remain.

Our first step in understanding the economics of discrimination is to distinguish two words that are often confused. *Prejudice* is an emotional dislike for members of a certain group; *discrimination* refers to the restricted opportunities offered to such a group. As you will see, although prejudice is *sometimes* the cause of discrimination, it need not be. And discrimination can occur even without prejudice.

EMPLOYER PREJUDICE

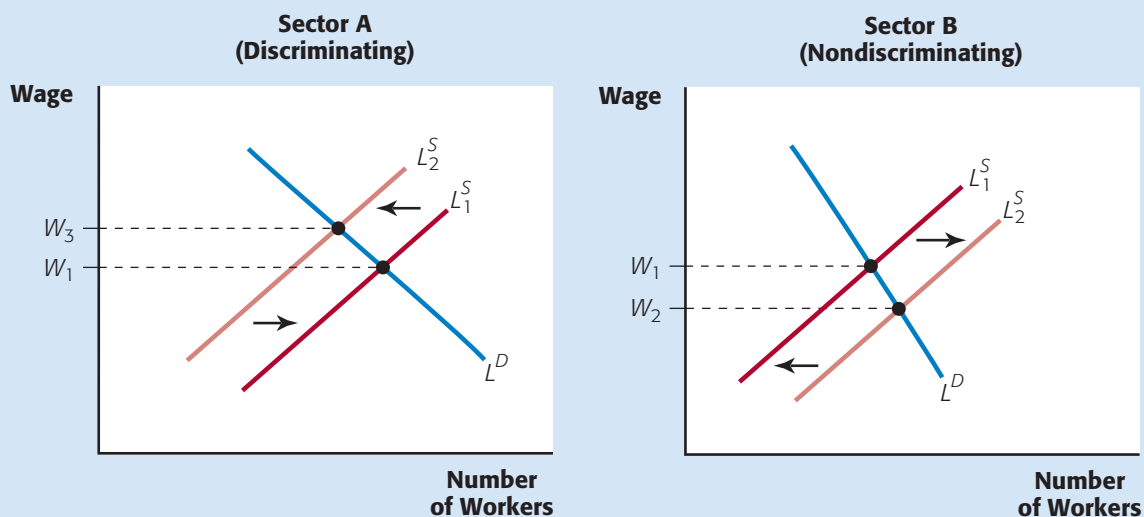
When you think of job discrimination, your first image might be a manager who refuses to hire members of some group, such as African-Americans or women, because of pure prejudice. As a result, the victims of prejudice, prevented from working at high-paying jobs, must accept lower wages elsewhere. No doubt, many employers hire according to their personal prejudices. But it may surprise you to learn that economists generally consider employer prejudice one of the *least* important sources of labor market discrimination.

To see why, look at Figure 4, which shows the labor market divided into two broad sectors, A and B. To keep things simple, we'll assume that all workers have

Discrimination When a group of people have different opportunities because of personal characteristics that have nothing to do with their abilities.

EMPLOYER DISCRIMINATION AND WAGE RATES

FIGURE 4



In the absence of discrimination, the wage rate would be W_1 in both Sector A and Sector B. If firms in Sector A discriminate against some group—such as women—the group would seek work in nondiscriminating Sector B. The increased labor supply in Sector B causes the wage there to fall to W_2 , while the decreased supply in Sector A causes the wage there to rise to W_3 . But only temporarily. As men migrate from Sector B to the now-higher wage Sector A, the labor supply changes in both sectors are reversed. The wage returns to W_1 in both sectors.

the same qualifications and that they find jobs in either sector equally attractive. Under these conditions, if there were *no* discrimination, both sectors would pay the same wage, W_1 . (Can you explain why?)

Now suppose the firms in sector A decide they no longer wish to employ members of some group—say, women. What would happen? Women would begin looking for jobs in the *nondiscriminating* sector B, and the labor supply curve there would shift rightward, decreasing the wage to W_2 . At the same time, with women no longer welcome in sector A, the labor supply curve there would shift leftward, driving the wage up to W_3 . It appears that employer discrimination would create a gender wage differential equal to $W_3 - W_2$.

But the differential would be only temporary. Why? With the wage rate in sector B now lower, *men* would exit that market and seek jobs in the higher-paying sector A. These movements would reverse the changes in labor supply, and, in the end, both sectors would pay the same wage again. Employer prejudice against women might lead to a permanent change in the *composition* of labor in each sector—with only men working in sector A and both sexes working in sector B—but *no change in wage rates*.

But employer prejudice might not even change the composition of labor in either sector, because there is another force working to eliminate this form of discrimination altogether: the output market. Since biased employers must pay higher wages to employ men, they will have higher average costs than unbiased employers. If biased firms sell their output in a competitive market, they will suffer losses and ultimately be forced to exit their industries. Over the long run, prejudiced employers should be replaced with unprejudiced ones. If the output market is imperfectly competitive, the firm will still have its stockholders or owners to contend with. Unless *their* prejudice is so strong that they are willing to forego profit, management will be under pressure to hire qualified women at a lower wage. In either case,

When prejudice originates with employers, market forces work to discourage discrimination and reduce or eliminate any wage gap between the favored and the unfavored group.

EMPLOYEE AND CUSTOMER PREJUDICE

What if *workers*—rather than employers—are prejudiced? Then our conclusions are very different. If, for example, a significant number of male assembly-line workers dislike supervision by women, then hiring female supervisors might reduce productivity, raise costs for any level of output, and therefore decrease profit. In a competitive output market, the *nondiscriminating* firm will be forced out of business. And even in imperfectly competitive output markets, stockholders will *want* the firm to discriminate against female supervisors, even if they themselves are not prejudiced. In this case, we cannot count on the market to solve the problem at all.

The same argument applies if the prejudice originates with the firm's *customers*. For example, if many automobile owners distrust female mechanics, then an auto repair shop that hires them would lose some customers and sacrifice profit. True, excluding qualified female mechanics is costly—it means paying higher wages to men and charging higher prices. But customers will be willing to *pay* a higher price, since they prefer male mechanics. Even in the long run, then, women might be excluded from the auto mechanics trade.

More generally, if worker or customer prejudice is common in high-wage industries, then women would be forced into low-wage jobs.

When prejudice originates with the firm's employees or its customers, market forces encourage, rather than discourage, discrimination and can lead to a permanent wage gap between the favored and unfavored group.

STATISTICAL DISCRIMINATION

Suppose you are in charge of hiring 10 new employees at your firm. Suppose, too, that young, married women in your industry are twice as likely to quit their jobs within two years than men (say, because they decide to have children) and that quits are very costly to your firm: New workers must be recruited and trained, and production is disrupted when there is a temporary gap in staffing. Let us say that 20 people apply for the 10 positions—half men and half women. All are equally qualified, and you have no way of knowing which individuals among them are more likely to quit within two years. Whom will you hire?

If your sole goal is to maximize the firm's profit, there is no question: You will hire the men. (If you have other goals, you may not last very long as a manager at that firm.) Notice that in this example, there was no mention of prejudice. Indeed, even if there isn't a trace of prejudice in you, in the firm's employees, or in its customers, profit maximization may still dictate hiring the men.

Statistical discrimination—so called because individuals are excluded based on the statistical probability of behavior in their group, rather than their own personal traits—is a case of discrimination without prejudice. It can lead an unbiased profit-maximizing employer to discriminate against an individual member of a group, even though that particular individual might never engage in the feared behavior.

But some observers have suggested that statistical discrimination is often a cover for prejudice. For example, consider statistical discrimination against women. True, women are more likely to leave work to care for their children. But men are more likely to develop alcohol and drug problems, which can lead to poor judgment and costly accidents on the job. If there were no prejudice, then the risks associated with hiring men would be thrown into the equation. According to critics of the statistical discrimination theory, the negative behavior of a favored group (such as men) is rarely considered by employers.

Statistical discrimination When individuals are excluded from an activity based on the statistical probability of behavior in their group, rather than their personal characteristics.

DEALING WITH DISCRIMINATION

As you've seen, discrimination due to pure employer prejudice is unlikely to have much of an impact on labor markets. As long as some employers are *not* prejudiced, those who *are* prejudiced will be at a competitive disadvantage. In the long run, the market helps to *eliminate* this type of discrimination.

But for other types of discrimination—such as statistical discrimination or discrimination due to worker or consumer prejudice—market incentives work in the opposite way, leading to a permanent and stubborn problem. In these cases, many economists and other policy makers believe that government action is needed. This is especially so when the groups discriminated against are already poor or disadvantaged in some way.

Some favor affirmative action programs, which actively encourage firms to expand opportunities for women and minorities; others favor stricter enforcement of existing antidiscrimination laws and stiffer penalties when discriminatory hiring

TABLE 2

**MEDIAN WEEKLY EARNINGS,
1998 (OF FULL-TIME WAGE
AND SALARY WORKERS)**

	Median Income	Percent of White Male Income
White Males	\$615	100%
Black Males	\$468	76%
Hispanic Males	\$390	63%
White Females	\$468	76%
Black Females	\$400	65%
Hispanic Females	\$337	55%

Source: U.S. Statistical Abstract, 1999, Table 702. (Note: Persons of Hispanic origin may be of any race.)

occurs. Both approaches to policy force *all* firms to bear the costs of nondiscriminatory hiring, so that no single firm is at a disadvantage. For example, by forcing *all* firms to hire women—and to bear the costs of greater quit rates or of alienating workers or customers who might be prejudiced—no single firm is put at a disadvantage by hiring women.

DISCRIMINATION AND WAGE DIFFERENTIALS

How much have the wages of victimized groups been reduced because of discrimination? As you are about to see, this is a very difficult question to answer.

A starting point—but *only* a starting point—is Table 2, which shows median earnings for different groups of full-time workers in the population. Notice the substantial earnings gap between men and women of either race and between whites and blacks of either sex. Doesn't this prove that the impact of discrimination on wages is substantial? Not necessarily.

Consider the black–white differential for men. In 1998, black men earned 24 percent less than white men, on average. But *some* of this difference is due to differences in education, job experience, job choice, and geographic location between whites and blacks. For example, the proportion of black adults with college degrees is about half that of white adults. Even if all firms were completely color blind in their hiring and wage payments, disproportionately fewer blacks would have higher-paying jobs requiring college degrees, and this would produce an earnings differential in favor of whites. The same would apply if blacks were more likely to live in low-wage areas or, on average, had fewer years of prior experience when applying for jobs.

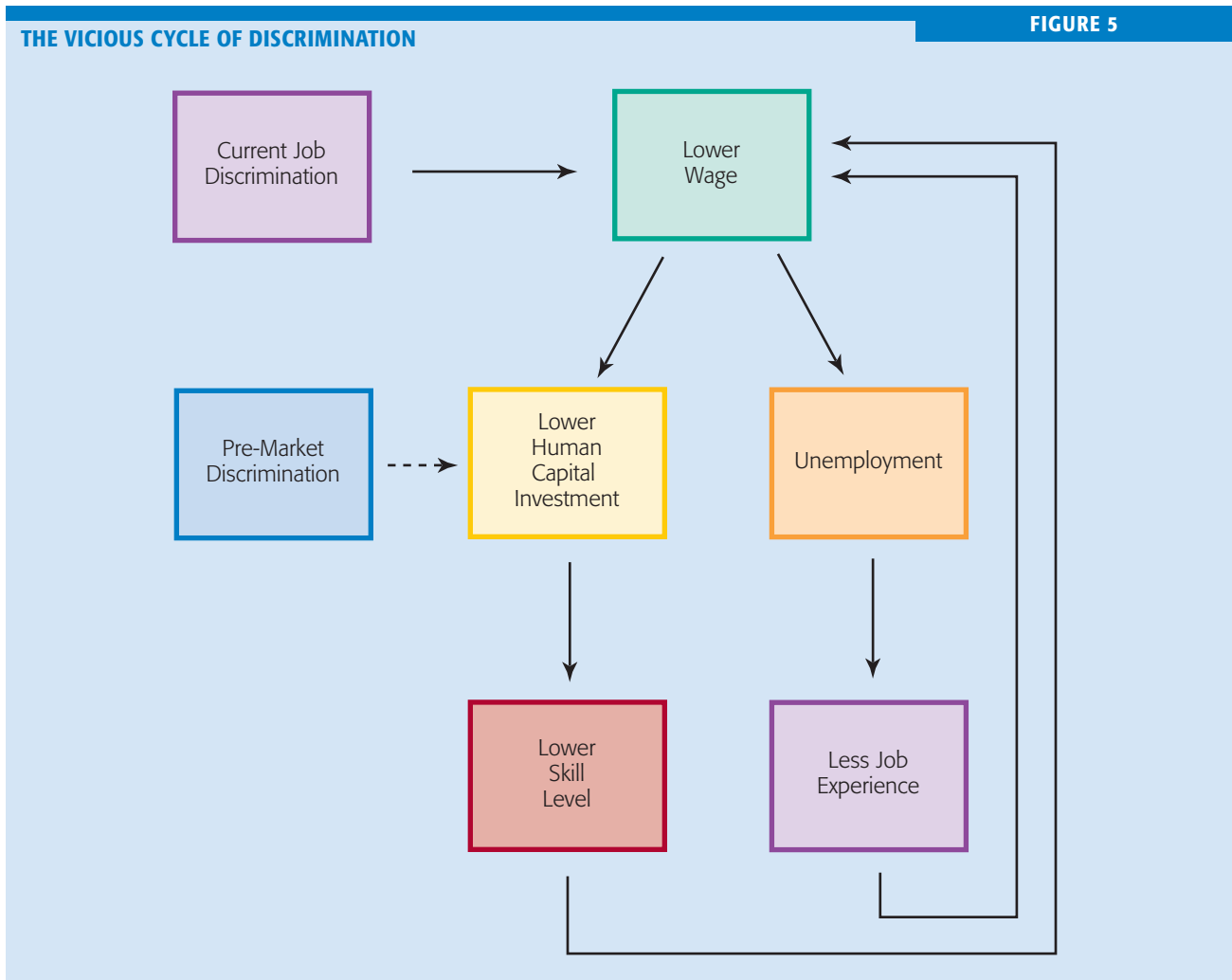
Several studies suggest that if we limit comparisons to whites and blacks with the same educational background, geographic location, and, in some cases, the same ability (measured by a variety of different tests), 50 percent or more of the earnings difference disappears.⁶

Does this mean that discrimination accounts for half or less of the earnings differential? Not at all: Many of the observed differences in education, geographic location, and ability are the *result* of job-market discrimination. Figure 5 illustrates a vicious cycle of discrimination in the labor market. First, job discrimination causes a wage differential between equally qualified whites and blacks. With a lower wage, blacks have less incentive to remain in the labor force or to invest in human capital, since they reap smaller rewards for these activities. The result is that blacks, on average, have less education and less job experience than whites, and even color-blind

⁶ See, for example, June O'Neill, "The Role of Human Capital in Earnings Differences between Black and White Men," *Journal of Economic Perspectives* (Fall 1990), pp. 25–45.

THE VICIOUS CYCLE OF DISCRIMINATION

FIGURE 5



employers will hire disproportionately fewer blacks in high-paying jobs, perpetuating their lower wages.

In addition to job-market discrimination, there is *pre-market* discrimination—unequal treatment in education and housing—that occurs *before* an individual enters the labor market. For example, regardless of black families' incomes, housing discrimination may exclude them from neighborhoods with better public schools, resulting in fewer blacks being admitted to college. Discriminatory treatment by teachers within a school may contribute to lowered aspirations and diminished job-market expectations. All of these contribute to the low-wage syndrome.

Similar reasoning applies to the earnings gap between women and men. On the one hand, we have a large earnings gap. In 1998, for example, the earnings of white, female workers were only 76 percent of those of white men. On the other hand, studies suggest that a third or more of the male–female wage gap is due to differences in skills and job experience. But for women, as well as blacks and other minorities, differences in skills and experience can be the *result* of lower wages, not just the cause of them: Since women know they will earn less than men and will have more trouble advancing on the job, they have less incentive to invest in human

capital and to stay in the labor force. Pre-market discrimination plays a role, too. Several studies have suggested that different treatment of girls in secondary school may lower their job-market aspirations. And even before school, girls may be socialized to prefer different (and lower-paying) career paths than boys, such as nursing rather than medicine.

In the end, we do not know nearly as much about the impact of discrimination on wages as we would like to know, but research is proceeding at a rapid pace. As we've seen, the data must always be interpreted with care:

In measuring the impact of job-market discrimination on earnings, the wage gap between two groups gives an overestimate, since it fails to account for differences in skills and experience. However, comparing only workers with similar skills and experience leads to an underestimate, since some of the differences are themselves caused by discrimination—both in the job market and outside of it.

MEASURING INCOME INEQUALITY

Wage differentials among households are an important cause of income inequality, but not the only cause. Two people with identical hourly wage *rates* may have vastly different wage or salary *incomes* because one is unemployed more often than the other or because one works more hours each week than the other.

Moreover, wages and salaries are not the only source of income. Some households supply capital (earning interest income or profit) or land and natural resources (earning rental income). These forms of income are often called *nonlabor income* or **property income**, to distinguish them from the wage and salary income derived from labor alone. Some of the largest incomes—such as Bill Gates's—are a mixture of labor and property income. Many households also receive **transfer payments** from the government, such as Social Security, unemployment insurance, or welfare payments.

When we are concerned with the fairness of our economic system, or the social problems that can result from inequality, we are ultimately concerned about inequality in *total* income, regardless of source. What can we say about income inequality in the United States? Although we have many measures of income inequality, they all leave much to be desired. Here, we consider the two most commonly cited measures.

THE POVERTY RATE

The **poverty rate** tells us the percentage of families whose incomes fall below a certain minimum, called the **poverty line**. The official poverty rate reported by the U.S. government is calculated as follows: The government determines the cost of feeding families of different types (number and ages of children, rural versus urban families, etc.). Then it is assumed that a family needs at least three times its food budget to pay for housing, clothing, transportation, and other basic requirements. Accordingly, the food budget is tripled to obtain the poverty line for each type of family. For example, in 1997, the poverty line for a family of two was an annual income of \$10,473; for a family of four, the poverty line was an annual income of \$16,400.

Finally, the poverty rate is then defined as the percent of U.S. families that fall below their respective poverty lines. In 1997, the official U.S. poverty rate was 10.3 percent, telling us that 103 out of every 1,000 families fell below the poverty line

Property income Income derived from supplying capital, land, or natural resources.

Transfer payment Any payment that is not compensation for supplying goods or services.

Poverty rate The percent of families whose incomes fall below a certain minimum—the **poverty line**.

Poverty line The income level below which a family is considered to be in poverty.

TABLE 3

POVERTY RATES FOR U.S. FAMILIES

Year	All Families	White	Black	Hispanic
1970	10.1%	8.0%	29.5%	Not Available
1980	10.3%	8.0%	28.9%	23.2%
1990	10.7%	8.1%	29.3%	25.0%
1997	10.3%	8.4%	23.6%	24.7%

Source: U.S. Statistical Abstract, 1999, Table 768.

defined for their characteristics. During the past two decades, the poverty rate has varied between 9 and 12 percent.

Poverty rates are important because they keep policy makers and the public aware of conditions at the bottom of the economic ladder. Of particular concern is the unequal *distribution* of poverty among different groups in the population. As you can see in Table 3, the poverty rate for black and Hispanic families has remained stubbornly above that for white families.

One reason for the persistently higher incidence of poverty among blacks and Hispanics is the lower wage rates earned by workers in these two groups, and this, in turn, is due to discrimination and differences in education levels (which may result from pre-market discrimination). In 1997, for example, 25 percent of the white population had college degrees, but only 13 percent of blacks and 10 percent of Hispanics had graduated from college. In addition to having lower wages, blacks and Hispanics earn much less nonlabor income than non-Hispanic whites.

The poverty rate gives us important information about the poorest families and how poverty is distributed among different groups within society. As a measure of income inequality, however, the poverty rate suffers from some serious drawbacks.

First, when calculating family income, the government leaves out the value of food stamps, Medicaid, and some other programs that help poor families. These programs are becoming more important over time. As a result, the poverty percentage tends to remain about the same, even though many poor people are becoming better off.

A second problem with gauging inequality with the poverty rate is that it ignores inequality among those *above* the poverty line. For a more comprehensive picture of inequality, we must turn to other measures, such as the one introduced in the next section.

THE LORENZ CURVE

Table 4 provides data that we can use to measure inequality across the entire spectrum of the income distribution. The table shows the percent of total income earned by each fifth of the population, when households are arranged by their incomes from lowest to highest. For example, the table shows that in 1998, the 20 percent of households with the lowest incomes earned only 3.6 percent of the total income, and the top 20 percent earned 49.2 percent of the total. If all households had earned identical incomes in every year, each entry in the table would be 20 percent. Notice, however, the high degree of *inequality* in the table. The inequality appears even greater when the top 20 percent is broken down further: In 1998, the top 5 percent of households (not shown in the table) earned 21.4 percent of total income, more than four times their proportional share.

TABLE 4

**PERCENT OF TOTAL
HOUSEHOLD INCOME
EARNED BY EACH FIFTH
OF U.S. HOUSEHOLDS**

	Lowest Fifth	Second Fifth	Third Fifth	Fourth Fifth	Highest Fifth	Gini Coefficient
1970	4.1%	10.8%	17.1%	24.5%	43.3%	0.394
1980	4.2%	10.2%	16.8%	24.8%	44.1%	0.403
1990	3.9%	9.6%	15.9%	24.0%	46.6%	0.428
1991	3.8%	9.6%	15.9%	24.2%	46.5%	0.428
1992	3.8%	9.4%	15.8%	24.2%	46.9%	0.433
1993	3.6%	9.0%	15.1%	23.5%	48.9%	0.447
1994	3.6%	8.9%	15.0%	23.4%	49.1%	0.456
1995	3.7%	9.1%	15.2%	23.3%	48.7%	0.450
1996	3.7%	9.0%	15.1%	23.3%	49.0%	0.455
1997	3.6%	8.9%	15.0%	23.2%	49.4%	0.459
1998	3.6%	9.0%	15.0%	23.2%	49.2%	0.456

Source: U.S. Census Bureau, *Historical Income Tables, 1967–98* (<http://blue.census.gov/hhes/income/histinc/h02.html> and <http://blue.census.gov/hhes/income/histinc/h04.html>), accessed January 3, 2000.

Lorenz curve When households are arrayed according to their incomes, a line showing the cumulative percent of income received by each cumulative percent of households.

Gini coefficient A measure of income inequality; the ratio of the area above a Lorenz curve and under the complete equality line to the area under the diagonal.

To get a clearer picture of what these numbers mean, look at Figure 6. The horizontal axis measures the cumulative percent of total households, and the vertical axis measures the cumulative percent of total income. For example, in 1998, the bottom 20 percent of households earned 3.6 percent of the total income, and the next 20 percent earned 9.0 percent, so the bottom 40 percent earned 3.6 percent + 9.0 percent = 12.6 percent of total income. Thus, one of the points in the figure is 40 percent on the horizontal axis and 12.6 percent on the vertical. The curve drawn through all the points obtained in this way is called the **Lorenz curve**.

If all households earned the same income, the Lorenz curve would be the thick straight line with a slope of 1 (marked “Line of Complete Equality”), since the bottom 20 percent would earn 20 percent of the total, the bottom 40 percent would earn 40 percent, and so on, until we reached 100 percent of all households, which—by definition—always earn 100 percent of the income. By contrast, the Lorenz curve in an economy with inequality will always be bowed out in the middle, although it will start and end at the same point as the line of complete equality. This gives us a visual representation of income inequality: The more bowed out the Lorenz curve—or the greater the area marked *A* in the figure—the greater will be the degree of inequality.

One of the most popular numerical measures of income inequality—the **Gini coefficient**—is obtained from the Lorenz curve in a very simple way: We divide area *A* in Figure 6 by the total area underneath the diagonal (area *A* plus area *B*). The more unequal the income distribution, the larger will be area *A* relative to area *A* + *B*, and the larger the Gini coefficient. If there were complete income equality—where everyone earned the same income—then area *A* would equal zero, so the Gini coefficient, $A/(A + B)$, would equal zero as well. The highest degree of inequality—where one person earned all the income, and the rest earned none—would give a Gini coefficient of 1.0. (Prove this to yourself by drawing the Lorenz curve for this case.) In general:

The larger the Gini coefficient—up to a maximum of 1.0—the greater is the degree of income inequality.

In 1998, the Gini coefficient for U.S. household income was 0.456 (see Table 4).

THE U.S. LORENZ CURVE

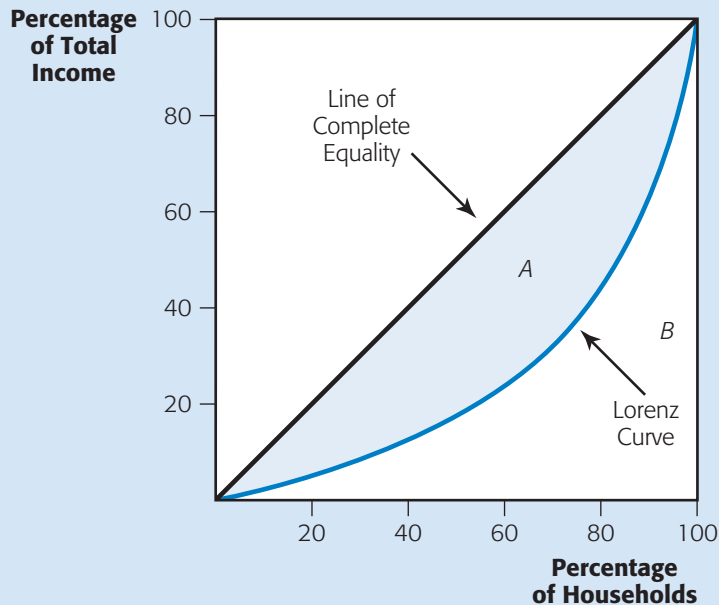


FIGURE 6

The Lorenz curve measures the cumulative percentage of income received by a particular cumulative percentage of the population, arranged from lowest income to highest.

Lorenz curves (and their associated Gini coefficients) are often used to compare income inequality among nations, as in Figure 7. Notice the similarity among the Lorenz curves of the three developed countries: the United States, France, and Japan. Japan has a Gini coefficient only slightly lower than that of the United States, whereas France's is only slightly higher. Much greater inequality is seen in Brazil, where society is sharply divided among the extremely poor—tribal members in the country and ghetto dwellers in urban areas—the rather wealthy middle class, and the superrich.

Within countries, Gini coefficients typically change little from year to year. They are most useful in indicating trends over long periods of time. For example, the slow, but steady, increase in inequality in the United States over recent decades has been reflected in a coefficient rising from 0.394 in 1970 to 0.456 in 1998. During this time, the top fifth gained more than 5 percentage points, most of which came from the share of the bottom three-fifths. The change from 1992 to 1993 was particularly striking, violating the general rule of slow changes: Between those two years, the Gini coefficient increased from 0.433 to 0.447. This set off a national debate about the causes of increasing income inequality, and what should be done about it, that continued through the decade.

Does anything you've learned in this chapter help to explain the rise in income inequality in the 1990s? Indeed it does. Income inequality arises largely from wage inequality. And among the reasons for wage inequality are three—differences in human capital, differences in ability, and the economics of superstars—that have very likely become more significant in the 1990s. As we discussed in Chapter 11, the increasingly rapid pace of technological change has probably increased the relative rewards to those with higher-than-average ability and education. And the revolution in telecommunications—especially the Internet—has created new media for superstars to reach a larger audience, and achieve even higher incomes.

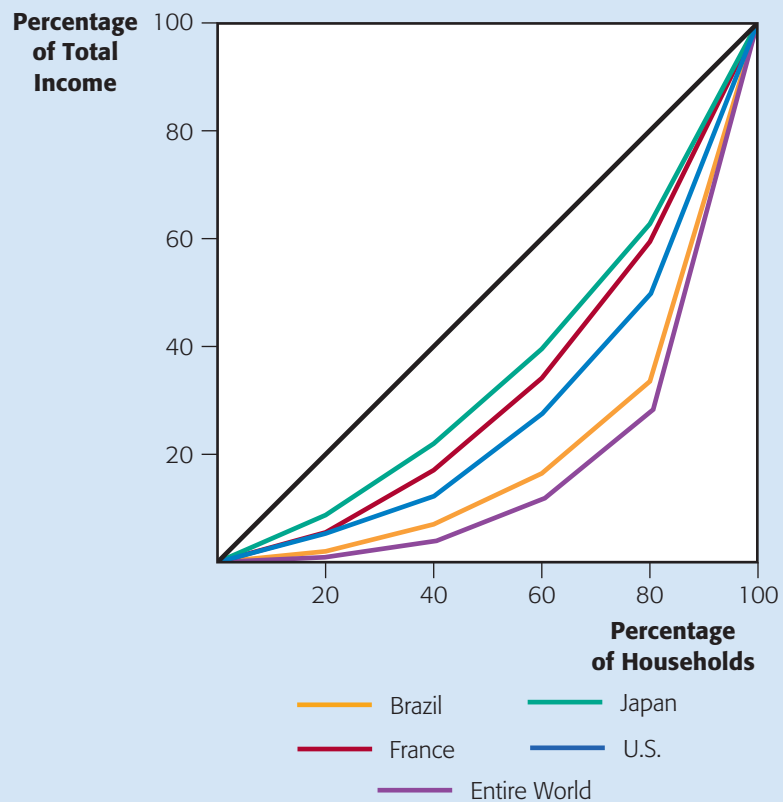


[http://](http://www.swcollege.com/bef/policy_debates/income_inequality.html)

For a comprehensive review of recent changes in income distribution, check the Policy Debates page at http://www.swcollege.com/bef/policy_debates/income_inequality.html.

FIGURE 7

LORENZ CURVES FOR SELECTED NATIONS



But wage inequality is not the only source of income inequality. People earn income not just from their labor, but also from their wealth. So we can gain further insight into the origins of income inequality by considering the distribution of *wealth*—the value of families’ assets (their houses, stocks, bonds, bank accounts, etc.) at some point in the year. Figure 8 plots a Lorenz curve for household wealth next to the Lorenz curve for household income, using data for 1998.⁷ As you can see, wealth is much less equally distributed than income, especially at the very top, where the top 10 percent of families owned more than half of total wealth.

The distribution of wealth is important in its own right. Wealth provides financial and psychological security, allows one to provide for one’s heirs, and may confer political influence on those who hold it.

In addition, as discussed above, *wealth provides income for its holders each year*. Indeed, this is where nonlabor income comes from. If you own shares of stock in a corporation, you will receive a share of the firm’s profit; if you own bonds, you will receive interest payments; if you own an apartment building or a mini-mall, you will receive rent. Since those who earn low wages and salaries also tend to have little wealth, the inequality in wealth *adds to* the inequality in total income.

⁷ “Recent Changes in U.S. Family Finances: Results from the 1998 Survey of Consumer Finances,” *Federal Reserve Bulletin* (January 2000).

U.S. LORENZ CURVES FOR INCOME AND WEALTH

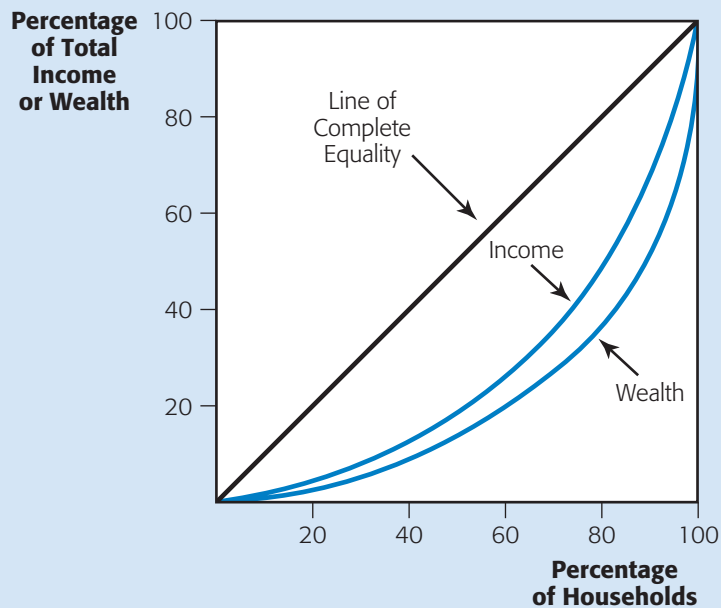


FIGURE 8

In the United States, income is more equally distributed than wealth.

PROBLEMS WITH INEQUALITY MEASURES

Ideally, we would like our measures of income inequality to tell us something about inequality in *economic well-being* and to use them as a guide for social and economic policy. For these purposes, however, the poverty rate, the Gini coefficient, and other gauges of income inequality—at least the way they are measured in practice—suffer from serious deficiencies.

Earned Income Versus Available Income. Our inequality measures are based on the income *earned* by different groups, not the income *available* for spending. For a variety of reasons, these can be very different.

First, the United States, like virtually all developed countries, has a progressive income tax (such that higher-income families pay a greater percentage of their income). However, since our inequality measures are based on income before tax, they tend to overstate inequality in available income.

Our measures also ignore some government transfers, like free medical care, food stamps, and subsidized housing, which free up income for other uses. While no one would argue that transfer payments can propel someone from the lowest fifth to the highest, they certainly increase the share of income going to those at the bottom. Ignoring transfers, like ignoring taxes, leads to an *overstatement* of inequality. Sweden's reputation as an egalitarian society, for example, is based on a highly progressive income tax and generous government transfers and programs for those at the bottom. None of this is reflected in Sweden's Lorenz curve.

On the other hand, our inequality measures ignore fringe benefits, which go mostly to those in the middle, and income from capital gains, which goes mostly to those at the top. (An individual has a capital gain when she sells assets—stocks, bonds, or real estate—at a price higher than the original purchase price.) If we

included fringes and capital gains in our measures, they would show a greater proportion of total income going to the middle and the top. For these reasons, our measures may *understate* income inequality.

Income Mobility. It is one thing to say that the bottom 20 percent of households earn only 3.6 percent of the income and quite another to say that, year after year, *the same households remain at the bottom*. The United States has a relatively mobile society—people switch careers, change jobs, and start new businesses more often than in most other countries. These changes—as well as pure chance—will give people good years and bad years. If many of those at the bottom or top are there only temporarily, then over a longer time horizon, there is less inequality than our measures suggest.

Moreover, one's own income tends to change in a predictable pattern over one's lifetime. Most workers start out earning low incomes, which then rise as they acquire more skills and experience, and, finally, fall sharply in retirement. This, too, can distort our measures of income inequality. To take an extreme example, imagine an economy that always has just five workers, each of whom passes through the same five phases of income over their lives: \$40,000 per year in the first decade, \$60,000 in the second decade, \$80,000 in the third, \$100,000 in the fourth, and then \$20,000 in the decade of retirement. Suppose, too, that at any point in time, one worker is in each phase. Then total yearly income in the economy will be $\$20,000 + \$40,000 + \$60,000 + \$80,000 + \$100,000 = \$300,000$. Each year, the bottom fifth (the retired worker) would earn just $\$20,000/\$300,000 = 0.066$ of total income. The top fifth (the worker at the height of her earning power) would have $\$100,000/\$300,000 = 0.333$ of total income. Even though everyone would have an identical income profile over his or her lifetime—total equality of *lifetime earnings*—the Lorenz curve would show substantial inequality.

The problem is that Lorenz curves, Gini coefficients, poverty rates, and most other measures of inequality give us a snapshot picture of the distribution of income when what we would ideally like is a moving picture—a picture of the distribution of lifetime earnings. But such information would be very difficult to gather; it would require tracing the incomes of a large sample of people over their entire lifetimes.

However, a few studies have tracked earnings over several years, and the findings are interesting. Look at Table 5. Each column of data represents the individuals in a particular income fifth over the years 1968–70. (Their income was averaged over the three years 1968, 1969, and 1970.) The data entry in the column tells us where those families ended up two decades later, in 1989–91. For example, the 53.8

TABLE 5

**INCOME MOBILITY
BETWEEN 1968–70 AND
1989–91**

		1989–91 Position				
		Bottom 20%	2nd 20%	3rd 20%	4th 20%	Top 20%
1968–70	Bottom 20%	53.8	21.8	18.8	4.8	0.9
Position	2nd 20%	22.7	25.4	18.5	25.8	7.7
	3rd 20%	11.1	21.4	24.4	27.8	15.4
	4th 20%	5.3	22.6	23.0	19.3	29.8
	Top 20%	7.0	8.6	16.2	22.2	46.1

Source: Peter Gottschalk and Sheldon Danziger, "Family Income Mobility—How Much Is There and Has it Changed?" *Boston College Working Papers in Economics*, No. 398 (December 1997). Number of individuals: 1,840.

in the upper left-hand corner tells us that 53.8 percent of the households that were in the bottom fifth in 1968–70 were still there in 1989–91. The entry next to it tells us that 21.8 percent of those who were in the lowest fifth in 1968–70 had moved up to the second fifth by 1989–91.

What do these numbers tell us overall? First, that the U.S. income distribution has been at least somewhat mobile. About half of those in the bottom fifth moved up within a generation, and half of those in the top fifth moved down. This does not mean that those in the top have switched places with those on the bottom; on the contrary, there was very little movement from one extreme to another. But there was substantial movement from the extremes to the middle three-fifths. For example, over the period studied, 45 percent of those at the bottom and 47 percent of those at the top moved to the middle.

The data in the table may actually understate income mobility in the United States, since it does not tell us what happened between the beginning and end periods of the study. A certain number of households that were in the same fifths in 1968–71 and 1989–91 actually moved out of their fifths in the interim. In other words, the table gives us two snapshots separated in time, but it is still not a moving picture.

An older study,⁸ but one less plagued by this problem, focused on movements into and out of poverty. The study found that while 24.4 percent of families had fallen below the poverty line in at least one year between 1969 and 1978, only 5.4 percent were below the line in five or more of those years, and only 0.7 percent were below the line during all 10 years. This suggests that, at least since the 1970s,

poverty has been far from a lifetime sentence for the majority of the American poor. But for a small minority, poverty is, indeed, a stubborn problem. All of this information is hidden by the simple poverty rate itself.

Careless Interpretations. Another problem with measures of income distribution is a criticism not of the measures themselves, but of how they are interpreted. So far, we have tried to be entirely descriptive, avoiding value judgments about the words “equality” and “inequality.” But ask yourself: As you have been reading this chapter, have you made the implicit assumption that more inequality is bad and more equality good? Many people (but few economists) automatically react in this way. They confuse *equality*—which means that everyone gets the same result—with *equity*—which implies *fair and equal treatment*. Even if we had a perfect measure of income inequality—say, one based on the lifetime income actually available to each citizen—extreme caution would be needed in drawing conclusions about equity or fairness. The reasons for this are the subject of the next section.

INCOME INEQUALITY, FAIRNESS, AND ECONOMICS

Fairness is difficult to define, in large part because we all have such different ideas about what it is. Witness the conflicts—which often come to blows—among kids at play, where the accusation “That’s not fair” is invariably answered with “Yes, it is.” Or think about the conflicts over marital property in divorce proceedings, over business property in the dissolution of a partnership, or over the grades given by

⁸ Greg Duncan, et al., *Years of Poverty, Years of Plenty: The Changing Fortunes of American Workers and Families* (Ann Arbor, MI: University of Michigan Institute for Social Research, 1984).



For an update on women’s earnings, see Mary Bowler, “Women’s earnings: An overview,” in the *Monthly Labor Review* (12/99) available at <http://stats.bls.gov/opub/mlr/art2full.pdf>.

teachers. In all of these cases, highly emotional disputes center around entirely different definitions of fair.

For the most part, economics steers clear of the fairness controversy, since so much of the field emphasizes positive (descriptive and predictive) issues, rather than normative (prescriptive) ones. But there is no avoiding the problem of fairness when one discusses income inequality. After all, what is the purpose of measuring inequality in the first place, except to compare *what is* to some standard of *what should be*?

Since the controversy over fairness is based on conflicting values, what can economics possibly contribute to this debate? Actually, quite a bit.

First, despite the controversy, there are *some* issues of fairness on which almost everyone agrees. By identifying the many different causes of income inequality—as we’ve done in this chapter—we can at least pinpoint those types of inequality that almost all of us would regard as fair and those we would regard as unfair. This is no small accomplishment, and it can help us avoid policies that would, when properly understood, actually make the distribution of income more *unfair*.

For example, almost everyone would agree that income inequality due solely to compensating wage differentials is entirely fair. If one worker must put up with longer hours, a greater risk of death, more unpleasant weather, a greater risk of unemployment, or more years of schooling than another, it is only fair that he or she be paid more. Thus, eliminating compensating wage differentials, which would make incomes more *equal*, would also make them less *equitable* to most of us.

The same holds for some of the inequality in property income. Remember the fable of the grasshopper, who fiddled all day, and the ant, who prepared for winter? Although many well-to-do Americans have inherited their wealth, many others have acquired theirs through years of working long hours, saving, or bearing risk. If some of us could have chosen to make these sacrifices, but did not, is it really fair for all of us to have the same wealth? Is it fair for the grasshopper to end up as wealthy as the ant? Most of us would say no.

These examples suggest the key to our common ground:

Inequality that results from choices that any of us can make is generally regarded as fair.

What about the other side of the coin: Do we all agree that inequality arising not from different choices, but from different *opportunities*, is inherently unjust? Actually . . . we do not seem to agree about this.

In some cases where opportunities are restricted—as in discrimination—there is widespread agreement, and social policy is often directed toward removing or weakening these barriers—for example, giving victims of employment discrimination the right to sue.

In other cases, there is no consensus about fairness. For example, some see large differences in inherited wealth as a social evil, creating an uneven playing field from the very beginning of life. Others believe that the freedom to use one’s property as one wishes—including passing it on to one’s heirs—is a fundamental human right. Similar disagreements occur over inequality arising from inherited talent, intelligence, beauty, or physical strength.

In a democracy, conflicts of values like these are resolved in the voting booth. Does economics have anything to contribute here? Yes, it does: Once we decide on our goals, there is the very difficult matter of designing policies to achieve them. A fuller understanding of the impacts of different policies, and the opportunity costs they require us to pay, can help us avoid serious and harmful mistakes.

Suppose, for example, that the majority of citizens, applying their own definitions of fairness, were to decide that it is unjust for superstars—the most talented or intelligent or physically gifted among us—to reap the huge rewards the market bestows upon them. The question of fairness, in the majority’s mind, has been resolved.

But then would come the tough questions. What are the options for limiting these high incomes? What would be the consequences of each action? In particular, what would be the opportunity cost for the rest of us? Would the reduced incentives mean fewer new discoveries of lifesaving drugs or new technology? Would fewer entrepreneurs be willing to devote the time, money, and energy to discover which goods and services we want? Would the quality of our culture gradually decline, as many talented singers, writers, artists, musicians, and actors decided that—with a smaller prize at the very top—a career in the arts is no longer worth it? In other words, would our effort to distribute the economic pie more equally result in a smaller pie overall? If so, how much smaller? And how would the burden be shared among the rest of us?

These are questions that economics, more than any other social science, is equipped to answer. An example of this type of analysis is discussed in the following “Using the Theory” section.

THE MINIMUM WAGE

One policy motivated by a desire for a more equitable distribution of income—at least for those at the bottom of the distribution—is the federal minimum wage law. When it was first established in 1938, the minimum wage was 25 cents per hour and applied to industries employing only 43 percent of the workforce. In 1999, the minimum wage was \$5.15 per hour and covered almost 90 percent of the workforce. Does the minimum wage create greater equality among our citizens? Let’s see.

To understand the effect of the minimum wage, we’ll divide the U.S. labor market into three parts: (1) the market for skilled labor; (2) the market for unskilled labor in industries covered by the minimum wage law; and (3) the market for unskilled labor in industries *not* covered by the law, either because it does not apply (waiters, house cleaners, and nannies) or because firms routinely violate it (typically, very small firms that are difficult to monitor).

Figure 9 shows what the initial equilibrium in all three markets would be if there were no minimum wage. The wage rate in the skilled labor market, where demand is high relative to supply, is \$20 per hour. In the unskilled labor market, where demand is lower relative to supply, we assume that the wage would be \$4.00. Notice that wages in both unskilled labor markets are equal, since—in the absence of a minimum wage law—workers would migrate to the market with the higher wage.

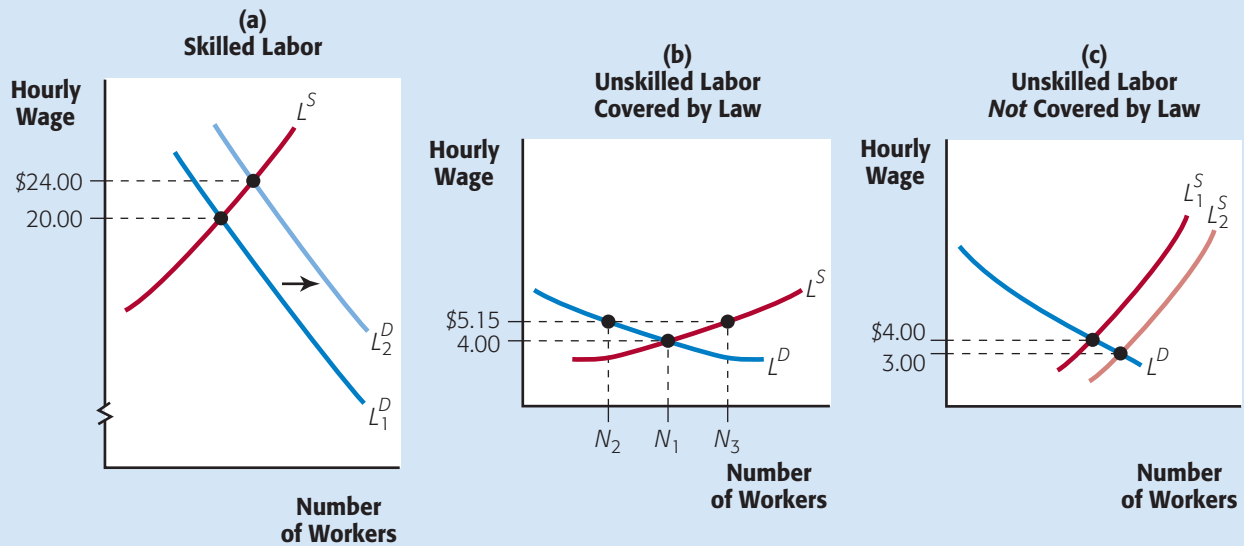
Now let us impose a minimum wage of \$5.15 in the covered unskilled sector and trace through the effects in the figure. First, employment in the covered unskilled sector falls, from N_1 to N_2 in panel (b). Since quantity demanded is less than quantity supplied at a wage of \$5.15, and since no firm can be forced to hire more workers than it desires, there will be an excess supply of labor equal to $N_3 - N_2$. Part of this excess is due to an increase in quantity supplied from N_1 to N_3 ; with a higher wage, more people want to work. But part of it is also due to a decrease in quantity demanded from N_1 to N_2 . You can already see that while some unskilled workers benefit—they earn a higher wage—others are hurt—they lose their jobs. But this is just the beginning.

Using the THEORY



FIGURE 9

THE MINIMUM WAGE



In the absence of a minimum wage law, the wage in the market for skilled labor is high—say, \$20 per hour. In the unskilled labor market, the wage is low—\$4 per hour. If a minimum wage of \$5.15 per hour is imposed, employment in the covered unskilled sector—panel (b)—falls from N_1 to N_2 . With a higher wage for unskilled labor, some firms will substitute skilled workers and capital equipment. This increases the demand for skilled labor and increases the hourly wage in panel (a) to \$24. At the same time, some individuals who lose their jobs in the covered, unskilled market of panel (b) will move to the uncovered labor market—panel (c)—further depressing the wage there.

Some of those who lose their jobs in the covered sector will move to the only sector where jobs are still available—the uncovered sector. There, the labor supply curve will shift rightward, from L_1^S to L_2^S in panel (c), and the market wage will fall below its initial value—to \$3.00 in our example. Thus, the impact of the minimum wage spills over into the sector not covered by it. Increased competition for jobs drives down the wages of *all* workers there, even those who were already employed before the minimum wage was imposed. More specifically, we would expect a decline in the wages of waiters, housecleaners, and unskilled workers who work in law-breaking firms.

What about skilled workers? Are they affected by minimum wage legislation? You might think not, since they are already earning more than the minimum. But when the wage of unskilled labor rises in the covered sector, employers there will, to some degree, substitute skilled workers and capital equipment for unskilled labor. For example, a dishwasher might be replaced by a sophisticated dishwashing machine that requires maintenance and repair by skilled workers. An unskilled floorwasher with a mop might be replaced by a cleaning service that uses skilled workers operating high-tech equipment to clean and wax floors. Substitution toward skilled labor will shift the labor demand curve in panel (a) rightward, from L_1^D to L_2^D . Further, skilled labor is needed to design, produce, and market the capital equipment itself, contributing to a further increase in demand. As a result, the wage rate in the skilled sector will increase, from \$20 to \$24 in our example.

You can see that the minimum wage sets off a chain of events. In the end, some unskilled workers benefit in the form of higher pay. Other unskilled workers are

harmled by lower pay or unemployment. There is only one group of workers in which everyone benefits: skilled workers. It should come as no surprise, then, that for many decades, the most vocal advocates of raising the minimum wage have been labor unions, whose membership consists almost entirely of skilled workers.

What do economists think about the minimum wage? There is both agreement and disagreement. Surveys consistently show that a large majority of economists agree with the analysis presented here, as well as its conclusion: that a minimum wage causes unemployment among unskilled workers. For example, in a 1995 survey of economists who specialize in the study of labor markets, 87 percent agreed that “a minimum wage increases unemployment among young and unskilled workers.”⁹

But *how much* unemployment is caused by a hike in the minimum wage? Here, the results of economic research vary and, accordingly, economists disagree. In the 1995 survey, when asked about the effect of a 10 percent increase in the minimum wage (say, from \$5.15 to \$5.65), the median estimate of the rise in unemployment among teenagers was 2 percent. But there was considerable variation around the median; only a quarter of the predictions fell between 1 and 3 percent.

What about policy advice? Here, too, there is disagreement. A slight majority of the labor economists (57 percent) believed that the minimum wage should be increased, in spite of any rise in unemployment. Not surprisingly, those who favored an increase in the minimum believed in a rather small unemployment effect; their median forecast for teenage unemployment was a 1 percent increase. Those who opposed any increase in the minimum thought that the impact on teenage employment would be larger (3 percent).

⁹ Robert Whaples, “Is There Consensus among American Labor Economists? Survey Results on Forty Propositions,” *Journal of Labor Research*, Fall 1996 (Vol. XVII, No. 4), pp. 730–731.

S U M M A R Y

In all nations, incomes vary markedly. Partly, that’s because of differences in wages that can be traced to differences in the attractiveness of jobs, differences in productivity, and imperfections in labor markets. When the attractiveness of two jobs differs, *compensating wage differentials* will emerge to offset those differences. When the productivity of workers differs, the more productive workers will earn higher wages. And in some cases, barriers to entry contribute to higher wages for protected workers.

Another reason for wage differentials is prejudice. When employer prejudice exists, market forces work to discourage discrimination and reduce wage gaps between groups. However, employee and customer prejudice encourage discrimination and can lead to permanent wage gaps.

Incomes also differ because of differences in nonlabor income. The *poverty rate* is the fraction of families whose incomes—however measured—fall below a certain minimum *poverty line*. The *Lorenz curve* and the associated *Gini coefficient* are comprehensive measures of income inequality. In the United States, as in most countries, wealth is less equally distributed than income.

All income measures tell us something about income inequality, but all suffer from deficiencies. A progressive income tax, government transfer programs, and fringe benefits all mean that *earned* income differs from income available for spending. Also, most measures of inequality pertain to a given point in time. Income itself tends to change in a predictable pattern over each individual’s lifetime.

K E Y T E R M S

compensating wage differential
nonmonetary job characteristic

discrimination
statistical discrimination
property income

transfer payment
poverty rate
poverty line

Lorenz curve
Gini coefficient

R E V I E W Q U E S T I O N S

1. For each of the following jobs, would you expect the compensating wage differential to be positive or negative? (In each case, compare to a job as a computer programmer.) Describe what nonmonetary job characteristics and human capital requirements might be at work in each case.
 - a. Worker in a slaughterhouse
 - b. College professor
 - c. Attorney
 - d. Bartender at a tropical resort
 - e. New York City police officer

2. In this chapter, you learned about several explanations for wage inequality. Which explanation (or explanations) best explains each of the following?
 - a. A paralegal in New York earns more than a paralegal doing the same work in Keokuk, Iowa.
 - b. Although they work on the same cases and do many of the same things, an attorney's salary is many times that of a paralegal.
 - c. Larry King earns more as a talk show host than Goofy Gary, the morning man on a New York radio show.
 - d. A professor of philosophy with a Ph.D. earns less than an accountant with only a B.A.
 - e. Construction workers in Germany, which has strong unions and extensive apprenticeship programs, are paid higher wage rates than American workers in the same trades.

3. Why are earnings not always proportional to ability?

4. True or False? Discrimination does not always arise from prejudice. Explain.

5. Explain how market forces tend to:
 - a. Encourage discrimination when the prejudice comes from a firm's employees or customers.
 - b. Discourage discrimination when the prejudice comes from employers.

6. What is "statistical discrimination"? What are some possible remedies for it?

7. How did technological advances contribute to increased income inequality in the 1990s?

8. Explain how the union–nonunion wage differential can arise. Illustrate with relevant graphs.

9. List and describe all possible income sources other than wages and salaries.

10. Discuss some of the problems associated with the inequality measures you studied in this chapter.

11. What would the Gini coefficient be if income in a country were equally distributed?

P R O B L E M S A N D E X E R C I S E S

1. The labor markets for factory workers and construction workers are in equilibrium: The wage in both is W_0 , and the number employed is N_0 . Assume that both labor markets are perfectly competitive, there are no barriers to entry or exit of workers, and factory skills are very similar to construction skills.
 - a. Unexpectedly, demand for factory output soars. Using graphs, show the short-run effect on the equilibrium wage and number employed in factories.
 - b. Draw graphs that illustrate the long-run equilibrium position in the two industries.

2. The following table lists the annual income of the 10 citizens of the little town of Dismal Seepage.

Joe	\$10,000	Dick	\$18,000
Jim	\$15,000	Ellen	\$ 3,000
Sue	\$ 4,000	Ann	\$30,000
Jack	\$25,000	Ralph	\$ 8,000
Roy	\$ 7,000	Bill	\$50,000

 - a. Draw the Lorenz curve for this community.
 - b. On the same graph, draw another (hypothetical) Lorenz curve that would reflect the effects of transfer payments but not fringe benefits.
 - c. Make a rough estimate of the Gini coefficients for both (a) and (b).
 - d. Assume that all the people in town live alone and that the yearly cost of food for a single person in Dismal Seepage is \$3,000. What is the official poverty rate in the town?

CHALLENGE QUESTIONS

1. You are designing a society from scratch. Issues of income distribution, equality of opportunity, and so on are completely up to you. The one catch is, you have to live in the society you design, and you don't know in advance your relative economic position in that society (high or low income, talented, average, or below average, etc.). What kind of society would you fashion? Discuss how your plan reflects your values concerning fairness, equity, efficiency, and so forth, as well as your concern over whether you end up on top or at the bottom.
2. Some advocates of the minimum wage argue that any decrease in the employment of the unskilled will be slight. They assert that an increase in the minimum wage will actually increase the total amount paid to unskilled workers (i.e., $\text{wage} \times \text{number of unskilled workers employed}$). Discuss what assumptions they are making about the wage elasticity of labor demand.

EXPERIENTIAL EXERCISES

1. Visit the Census Bureau's Web page on poverty statistics at www.census.gov/hhes/www/poverty.html. Look at the Small Area Income and Poverty Estimates and find the latest estimates for your county. How does the poverty rate there compare to the overall rate in your state and in the United States as a whole? Use economic reasoning to explain the differences you find.



2. The front page of the Marketplace section of the *Wall Street Journal* sometimes carries articles on income distribution and the personal impact of poverty. Pay particular attention to the Work & Family and Business and Race columns in the Wednesday paper. Pick one of the articles, and see if you can identify some of the economic forces affecting income distribution and poverty.



CHAPTER

13

CAPITAL AND FINANCIAL MARKETS

CHAPTER OUTLINE

Physical Capital and the Firm's Investment Decision

The Value of Future Dollars
The Firm's Demand for Capital
What Happens When Things Change: The Investment Curve

Investment in Human Capital

General Versus Specific Human Capital
The Decision to Invest in General Human Capital

Financial Markets

The Bond Market
The Stock Market
The Economic Role of Financial Markets

Using the Theory: Can Anyone Predict Stock Prices?

Predicting Stock Prices: Fundamental Analysis
Predicting Stock Prices: Technical Analysis
The Economist's View: Efficient Markets Theory

In January 2000, the following events made headlines in newspapers across the country.

- America Online, the nation's largest Internet Service Provider, acquired Time Warner, a media conglomerate that owned *People* magazine, Elektra Records, CNN, Cinemax, and dozens of other companies. The purchase price for Time Warner was \$183 billion, making it the largest corporate takeover in U.S. history.
- Corning Inc. announced that it would spend \$750 million on plant and equipment over the next few years to expand its optical fiber manufacturing capacity by more than 50 percent.
- The Department of Education reported a boom in distance education, with enrollments more than doubling between 1995 and 1998.
- Extensivity, a new software company, sold shares of stock to the public for the first time. Before the month was over, the price of its shares had jumped by 260 percent.

These events might seem to have little to do with one another. But in fact, all of them arose from a similar source. In each case, the event occurred because some decision maker was able to put a value on money to be received in the *future*.

In this chapter, we will study decisions about *streams of future payments*. More specifically, we'll study two types of decisions: (1) the decision to invest in productive capital, such as factory buildings, equipment, or in skills and training; and (2) the decision to purchase financial assets, such as stocks and bonds. To understand all of these decisions, we need new concepts and techniques to help us place a value on income to be received in the future.

PHYSICAL CAPITAL AND THE FIRM'S INVESTMENT DECISION

The concept of *capital* was introduced in the first chapter of this book. There, you learned that capital is one of society's *resources*, along with land and labor. More specifically, capital is any long-lasting *tool* that people use to produce goods and

services. You also learned that we can classify capital into two categories: physical *capital*, such as the plant and equipment owned by business firms, and *human capital*—the skills and training of the labor force. In this section, we'll focus on firms' decisions about physical capital, and we'll take up human capital in the next section.

How does a business firm decide how much physical capital to buy? In the same way that it makes any other decision. The firm's goal is to maximize its profit—not just this year, but over many years into the future. But in trying to select the best quantity of capital to purchase, the firm faces constraints.

First, the firm faces some given technology, as represented by its production function. The technology tells us how much output the firm can produce with each quantity of capital it might purchase and put in place.

Second, the firm faces a constraint on the price it can charge for its *output*. This constraint is determined by the demand curve it faces. As we did when we studied labor markets, we'll keep our analysis simple by assuming that firms sell their output in perfectly competitive product markets. That is, each firm takes the price of its product as a given.

Finally, the firm must *pay* for its capital, just as it must pay for its other inputs. Here again, we'll assume that it faces a perfectly competitive market for physical capital. As a consequence, the firm takes the market price of the capital it buys as a given.

Given these constraints—on technology, the price at which it can sell its output, and the price it must pay for its capital—the firm tries to buy the best quantity of physical capital. How does it make this decision?

Let's make this question more specific. Suppose you are the fleet manager at Quicksilver Delivery Service. Your firm delivers packages for small retailers in the Chicago metropolitan area for which it charges the market rate of \$4 per package. You are in charge of buying new trucks as the firm expands. How many trucks should you buy?

Think back a few chapters to when you learned about the demand for labor. There we saw that—in determining the profit-maximizing number of workers to employ—the firm should keep hiring additional workers as long as their benefit to the firm—measured by the marginal revenue product of labor (MRP_L)—exceeds the cost to the firm—the wage rate.

Something very similar is true of your demand for new trucks. You want to determine—for each additional truck—whether the benefit exceeds the cost. So, there are some obvious parallels between your firm's demand for trucks and its demand for labor. But there are some important differences as well.

First, your firm does not *own* the labor it employs. Instead, it *rents* the labor by paying each worker a certain wage each week. With capital, things are different. Although it is possible to rent equipment, most firms choose to *purchase* their capital outright. And because capital is durable and, by definition, long lasting, the firm must think about the future when deciding whether to buy plant and equipment.

Let's rephrase the firm's decision about capital using language that, by now, should be familiar to you. The additional yearly revenue you get from a unit of capital—such as a truck—is called the **marginal revenue product of capital** (MRP_K). But the MRP_K tells us only part of the story. It measures the additional revenue in any one year, but to find the *total* benefit of capital, we need to measure the additional revenue over *many* years—as many years as the capital will last. For example, when you purchase a truck, that truck will contribute to your firm's income this year, next year, and on into the future. So when we measure the benefits of buying another truck, we must find a way to value the revenue that the truck earns for your firm over several years.

“That's easy,” you might think. “I'll just add up the revenue that an additional truck will earn for me in each of the years that I'll own it.” For example, suppose



Identify Goals and Constraints

Marginal revenue product of capital

The increase in revenue due to a one-unit increase in the capital input.



A truck—like most types of physical capital—will increase a firm's revenue for many years. As a result, the firm must calculate the present-dollar equivalent of future receipts.

Present value The value, in today's dollars, of a sum of money to be received or paid at a specific date in the future.

you are trying to decide whether to buy a truck that will last 15 years, and its MRP_K is \$10,000. According to this method of determining the truck's benefits, you would just multiply the yearly MRP_K of \$10,000 by the number of years the truck will last. The total benefit of the truck would then be $15 \times \$10,000 = \$150,000$. Is this right?

Not quite. The problem with this approach is that it adds each year's revenue to the total, regardless of *when* the revenue is earned. But in reality, the value of a future payment depends on when that payment is received. To see why, we'll have to take a detour from Quicksilver Delivery and explore the issue of future payments more generally. We'll come back to Quicksilver and its trucks when we're done.

THE VALUE OF FUTURE DOLLARS

To see why the value of a future payment depends on *when* that payment is received, just run through the following thought experiment. Imagine that you are given the choice between receiving \$1,000 now and \$1,000 one year from now. Do you have to think hard before making up your mind? Regardless of when you will actually spend the money, it is always better to have the dollars earlier rather than later. For example, say you don't plan to spend the money until next year. Then, if you get the \$1,000 now, you could put it in the bank and earn interest for a year, giving you *more* than \$1,000 when you finally spend it. On the other hand, say you plan to spend the money right away. Then receiving it *now* rather than later saves you the interest you would have to pay to borrow the money for immediate use.

Because present dollars can earn interest, and because interest must be paid to borrow present dollars, it is always preferable to receive the same sum of money earlier rather than later. Therefore, a dollar received now is worth more than a dollar received later.

Knowing that dollars received in the future are worth less than dollars received today is an important insight. But when analyzing capital markets, we need to know precisely *how much* a given future payment is worth—that is, how many of today's dollars you would trade for the future payment.

The present value (PV) of a future payment is the value of that future payment in today's dollars. Alternatively, it is the most anyone would pay today for the right to receive the future payment.

To understand this concept better, let's work out a simple example: What is the present value of \$1,000 to be received one year in the future? That is, what is the most you would pay *today* in order to receive \$1,000 one year from today? The answer is certainly *not* \$1,000. Why not? If you paid \$1,000 today for a guaranteed \$1,000 in one year, you would be giving up \$1,000 that you *could* lend to someone else for interest. If you lent the money, you'd end up with *more* than \$1,000 one year later. So it never makes sense to pay \$1,000 now for \$1,000 to be received one year from now.

But would you pay \$900 for the guaranteed future payment? Or \$800? In fact, the most you would pay is the amount of money that, if lent out for interest, would get you exactly \$1,000 one year from now. That amount of money is the *present value* of \$1,000 to be received in one year, since that is the most you would part with today in exchange for the future payment.

This observation suggests that the present value of a future payment depends on the interest rate at which you can lend funds. Suppose this interest rate is 10 percent per year. Then the present value of \$1,000 to be received one year from today is an

amount of money that—if lent out at 10 percent annual interest—would give you precisely \$1,000 in one year. At 10 percent interest, each dollar you lend out will give you 1.10 dollars in one year, so the PV we seek will satisfy the following equation:

$$PV \times 1.10 = \$1,000.$$

Solving for PV , we get

$$PV = \frac{\$1,000}{1.10} = \$909.$$

In words, if you lent out \$909 at 10 percent interest, you would have \$1,000 one year from today. Therefore, \$909 is the most you would be willing to give up today for \$1,000 in one year, or *\$909 is the present value of \$1,000 received one year from now.*

We can generalize this result by noting that, if the interest rate had been something other than 0.10—we'll call it i —or the amount of money had been something other than \$1,000—say, Y dollars—then the present value would satisfy the equation

$$PV \times (1 + i) = Y$$

or

$$PV = \frac{Y}{(1 + i)}.$$

But what if the payment of \$ Y were to be received *two* years from now instead of one? Then we can use the same logic to find the present value. In that case, each dollar lent out would become $(1 + i)$ dollars after one year, and then—when the dollar plus the earned interest was lent out again for a second year—it would become $(1 + i)(1 + i) = (1 + i)^2$ dollars at the end of the second year. Thus, the PV will satisfy

$$PV \times (1 + i)^2 = Y$$

and solving for PV , we obtain

$$PV = \frac{Y}{(1 + i)^2}.$$

Finally, for payments to be received one, two, or any number of years in the future, we can state that

the present value of \$ Y to be received n years in the future is equal to

$$PV = \frac{Y}{(1 + i)^n}.$$

For example, with an interest rate of 10 percent, the present value of \$1,000 to be received three years in the future would be

TABLE 1

PRESENT VALUES OF \$1
FUTURE PAYMENTSValue of \$1 to Be Received at Various
Numbers of Years in the Future,
at Different Discount Rates

No. of Years in Future	5 Percent	10 Percent	15 Percent
0	\$1.00	\$1.00	\$1.00
1	\$0.95	\$0.91	\$0.87
2	\$0.91	\$0.83	\$0.76
3	\$0.86	\$0.75	\$0.66
4	\$0.82	\$0.68	\$0.57
5	\$0.78	\$0.62	\$0.50
10	\$0.61	\$0.39	\$0.25
20	\$0.38	\$0.15	\$0.06

$$PV = \frac{\$1,000}{(1.10)^3} = \$751.$$

Discounting The act of converting a future value into its present-day equivalent.

Discount rate The interest rate used in computing present values.

The process of making dollars of different dates comparable is called **discounting**. Since the interest rate is used to compute the present value of future dollars, the interest rate itself is called the **discount rate**.¹ Table 1 shows the present value of a dollar to be received at different times in the future, at different discount rates.

For example, what is the present value of \$1 to be received 10 years from today? If the interest rate is 10 percent, the present-day equivalent is \$1 *divided by* $(1.10)^{10}$, or $\$1/2.59 = \0.39 . This tells us that, when the interest rate (discount rate) is 10 percent, anyone expecting to receive \$1 ten years from today might just as well accept \$0.39 now. After all, when loaned at 10 percent interest per year, 39 cents will become \$1 in 10 years.

From the logic of present-value calculations, and from the entries in Table 1, we can see that

the present value of a future payment is smaller if (1) the size of the payment is smaller, (2) the interest rate is larger, or (3) the payment is received later.

Why does postponing a future payment decrease its present value? Because the later you receive your money, the greater the sacrifice of interest you *could* have earned in the meantime. Why do higher interest rates decrease the present value of a future payment? Because the higher the interest rate, the greater the interest you *could* have earned by lending out your money today and, therefore, the more interest income you *sacrifice* by waiting.

Finally, there is one more way in which we use the formula for present value calculations: to determine the value of a stream of future payments, with each individ-



First Interstate Bank maintains on-line present and future value calculators. You can find them at <http://www.firstinterstatebank.com/planning/index.htm>.

¹ In macroeconomics, the term *discount rate* has a completely different meaning: It's the interest rate that the Federal Reserve charges banks when it lends them reserves. There is no connection between the two different meanings of the term.

ual payment to be received at a *different* time in the future. For example, how can we calculate the value, in today's dollars, of the following stream of future payments: \$1,000 to be received one year from now, \$900 to be received two years from now, and \$600



Be careful when working with interest rates: They can be expressed in either percentage form or decimal form. An interest rate of 5 percent (5%) can also be expressed in decimal form as 0.05. This is why the expression “1 + *i*” is equal to 1.05 when the interest rate is 5 percent. Similarly, an interest rate of 0.5% (*one-half* of one percent) would translate to 0.005 in decimal form, and “1 + *i*” would then equal 1.005.

to be received three years from now? The answer is: We first calculate the present value of each payment, and then we add those present values together:

$$PV = \frac{\$1,000}{(1 + i)} + \frac{\$900}{(1 + i)^2} + \frac{\$600}{(1 + i)^3}$$

With an interest rate of 10%, the *total* present value of the entire stream of payments is equal to:

$$\begin{aligned} PV &= \frac{\$1,000}{(1.10)} + \frac{\$900}{(1.10)^2} + \frac{\$600}{(1.10)^3} \\ &= \$909.09 + \$743.80 + \$450.79 \\ &= \$2,103.68 \end{aligned}$$

The logic of present value shows us why anyone who expects to receive a stream of future payments must discount each of those payments before adding them together. The next section provides an example of how firms use present value to make decisions about investing in new capital.

THE FIRM'S DEMAND FOR CAPITAL

Let's return to your problem at Quicksilver Delivery Service. How many new trucks should you buy? Suppose that the first new truck you purchase would be used to serve the Northern territory. Buying this truck would enable your firm to deliver 2,500 additional packages each year, thereby generating $\$4 \times 2,500 = \$10,000$ in additional yearly revenue.² So the marginal revenue product of that first new truck is \$10,000 per year.

A second new truck would be used in a new Northeast territory. It turns out that this truck, too, would generate \$10,000 in additional revenue, so its annual *MRP* also equals \$10,000. If you purchase a third truck, you would use it in the Eastern territory where, in the course of a typical year, it would generate \$8,000 of additional revenue. A fourth truck could generate \$5,000 of revenue on a new Southern route. And a fifth truck would be used in the Southeastern territory, but only partly for delivering packages. The rest of the time, it would be used to pick up packages that were shipped to the wrong addresses, to drop off mail, and for other miscellaneous purposes. It would generate \$2,000 of additional revenue each year.

² When we report how much additional revenue a truck will contribute, we are referring to *net* revenue. That is, we've already subtracted off any additional costs that go along with having another truck, such as the costs of gasoline, maintenance and repairs, and hiring another driver.

TABLE 2

THE PRESENT VALUE OF TRUCKS AT QUICKSILVER DELIVERY SERVICE (WITH A DISCOUNT RATE OF 10%)

Truck	Additional Annual Revenue (MRP_K)	Total Present Value of Additional Revenue over 15 years
1	\$10,000	$\frac{\$10,000}{(1.1)} + \frac{\$10,000}{(1.1)^2} + \dots + \frac{\$10,000}{(1.1)^{15}} = \$76,060.80$
2	\$10,000	$\frac{\$10,000}{(1.1)} + \frac{\$10,000}{(1.1)^2} + \dots + \frac{\$10,000}{(1.1)^{15}} = \$76,060.80$
3	\$ 8,000	$\frac{\$8,000}{(1.1)} + \frac{\$8,000}{(1.1)^2} + \dots + \frac{\$8,000}{(1.1)^{15}} = \$60,848.64$
4	\$ 5,000	$\frac{\$5,000}{(1.1)} + \frac{\$5,000}{(1.1)^2} + \dots + \frac{\$5,000}{(1.1)^{15}} = \$38,030.40$
5	\$ 2,000	$\frac{\$2,000}{(1.1)} + \frac{\$2,000}{(1.1)^2} + \dots + \frac{\$2,000}{(1.1)^{15}} = \$15,212.16$

Let's suppose that each truck has an expected useful life of 15 years,³ so that Quicksilver can look forward to 15 years' worth of additional revenue for each additional truck that it buys. If the appropriate discount rate for Quicksilver's *PDV* calculations is 10 percent, then we can calculate the total *PDV* of the future revenue as in Table 2. Notice that, after the firm buys 2 trucks, the totals in the last column get smaller as the number of trucks increases. This occurs because of a property of inputs that should be familiar to you—diminishing marginal productivity. As more and more capital is employed, the marginal product of capital (*MPK*) declines—each additional truck can deliver fewer additional packages than the truck before. Because the price you charge for delivery services (*P*) is constant at \$4 per package, this means that the marginal revenue product of capital— $MRP_K = P \times MPK = \$4 \times MPK$ —also decreases as additional trucks are purchased. Finally, with a decreasing MRP_K , the present value totals in the last column must also decrease as more trucks are added.

Now that we know the total present value that you gain from each truck, do we know how many trucks you should buy? Almost, but not quite. There is still the matter of how much each truck *costs*. Remember that you should buy a truck whenever the benefit from that truck exceeds its cost. So you would buy all trucks for which the numbers in the last column of Table 2 exceed the price of a truck. For example, suppose delivery trucks cost \$65,000 each. Then it certainly makes sense to buy the first two trucks, each of which generates a total present value of \$76,061 in additional revenue. But it does *not* make sense to buy the third, since it generates a total present value of only \$60,849—less than the price of the truck. If, on the other hand, trucks cost \$50,000 each, it would make sense to buy three of them,

³ Actually, it is not quite right to assume that a truck will generate the same amount of revenue each year of its useful life. Trucks, like all capital goods, *depreciate*—they wear out gradually. As a result, the additional revenue contributed by a given truck will grow smaller with each passing year, rather than stay the same as we've assumed. Ignoring depreciation helps keep the math simple. If you go on in economics, you'll learn how to incorporate depreciation into present value calculations.

since the first three trucks all generate a total present value for the firm that is greater than \$50,000.

Our examples have focused on a special type of capital—delivery trucks—but the same logic works for any other type of physical capital—automated assembly lines, desktop computers, filing cabinets, locomotives, and construction cranes. In each of these cases, the first step in making a decision about a capital purchase is to put a value on an additional unit of capital. This value is the total present value of the future revenue generated by the capital.

This first step—putting a value on physical capital—is so important and so widely applicable that we can refer to it as a general principle:

The principle of asset valuation says that the value of any asset is the sum of the present values of all the future benefits it generates.

Principle of asset valuation The idea that the value of an asset is equal to the total present value of all the future benefits it generates.

The principle of asset valuation tells us how to determine the marginal benefit from buying another unit of capital, such as another truck. The next step is to compare this marginal benefit with the cost of the capital itself. As you've seen, the firm should then buy any capital for which the marginal benefit (total present value of future revenue) is greater than the cost.

WHAT HAPPENS WHEN THINGS CHANGE: THE INVESTMENT CURVE

Investment is the term economists use to describe firms' purchases of new capital over some period of time. In the example above, if trucks cost \$50,000 each, Quicksilver should buy three of them. If it bought all three trucks this year, its investment expenditures for the year would be $\$50,000 \times 3 = \$150,000$.

Investment Firms' purchases of new capital over some period of time.

But this conclusion about investment is based on the assumption that the interest rate—and Quicksilver's discount rate—is 10 percent. With a lower interest rate—say, 5 percent—each year's revenue would have a higher present value, so the total present value of any truck would be higher. Our conclusion about Quicksilver's investment spending might then change. Similarly, a rise in the interest rate—say, to 15 percent—would *lower* the present value of each year's revenue, and *decrease* the total present value of a truck.

Table 3 shows how our total present value calculations for each truck change as the interest rate changes. The table assumes that the other ingredients in the firm's decision making do not change. Each package delivered still generates revenue of \$4, and the productivity of each truck is still what it was before. For instance, a

TABLE 3

Truck	Additional Annual Revenue	Total Present Value with a Discount Rate of:		
		5%	10%	15%
1	\$10,000	\$103,797	\$76,061	\$58,474
2	\$10,000	\$103,797	\$76,061	\$58,474
3	\$ 8,000	\$ 83,037	\$60,849	\$46,779
4	\$ 5,000	\$ 51,898	\$38,030	\$29,237
5	\$ 2,000	\$ 20,759	\$15,212	\$11,695

PRESENT VALUE CALCULATIONS FOR VARIOUS INTEREST RATES

truck used on the Northern route would still allow Quicksilver to deliver 2,500 additional packages each year.

The numbers in the last three columns are each calculated just as the numbers we calculated in Table 2. The only difference is that, instead of always assuming a discount rate of 10 percent, Table 3 shows the total present value for each truck under three different interest rates. Notice what happens as we move from left to right in the table for any particular truck: the interest rate rises, from 5 percent to 10 percent to 15 percent, and the value of the truck to the firm falls.

Now, if trucks cost \$50,000 each, how much will Quicksilver invest (spend on new trucks) at any given interest rate? Let's see. If the interest rate is 5 percent, Quicksilver should buy four trucks, because each of the first four trucks has a total present value greater than \$50,000 at that interest rate. The fifth truck, however, has a total present value of only \$20,759, so the firm should not buy that one. Thus, if the interest rate is 5 percent, Quicksilver's investment spending will be $\$50,000 \times 4 = \$200,000$.

If the interest rate rises to 10 percent, Quicksilver should buy only three trucks. (Can you see why? *Hint*: What is the total present value of the fourth truck when the interest rate is 10 percent?) At this higher interest rate, Quicksilver's investment spending would fall to $\$50,000 \times 3 = \$150,000$. Finally, if the interest rate rises to 15 percent, Quicksilver should buy only two trucks, so its total investment spending is $\$50,000 \times 2 = \$100,000$.

What is true for Quicksilver is true for every truck-buying firm in the economy: The higher the interest rate, the fewer trucks delivery services and other truck-buying firms will want to purchase, and the smaller will be investment expenditures in trucks during the year.

Take a moment to think about why this happens. The trucks themselves are the same, and they are just as productive as before. But each truck is less valuable to firms in *present-dollar* terms. That's because waiting to receive future revenues now has a greater opportunity cost, whereas the truck is still paid for in today's dollars, whose value is unaffected by the interest rate. So each firm will want fewer trucks at any given price.

Moreover, the same logic applies to other capital purchases. At high interest rates, U.S. firms end up buying less of all different kinds of capital—not just delivery trucks, but also other durable goods such as computers, machine tools, combines, and printing presses. It should be no surprise, then, that we come to the following conclusion:

As the interest rate rises, each business firm in the economy—using the principle of asset valuation—will place a lower value on additional capital, and decide to purchase less of it. Therefore, in the economy as whole, a rise in the interest rate causes a decrease in investment expenditures.

The relationship between the interest rate and investment expenditure is illustrated by the economy's investment curve, shown in Figure 1. The curve slopes downward, indicating that a rise in the interest rate causes investment spending to fall. When you study *macroeconomics*, you'll learn that the investment curve is important for the performance of the overall economy, for several reasons. But here's a hint as to one of them: When the interest rate falls, the increased investment in new capital means that the nation's *capital stock*—the total quantity of installed capital—will end up larger than it otherwise would be. With more capital, labor will be more productive, and our standard of living will be higher. This relationship between the

THE INVESTMENT CURVE

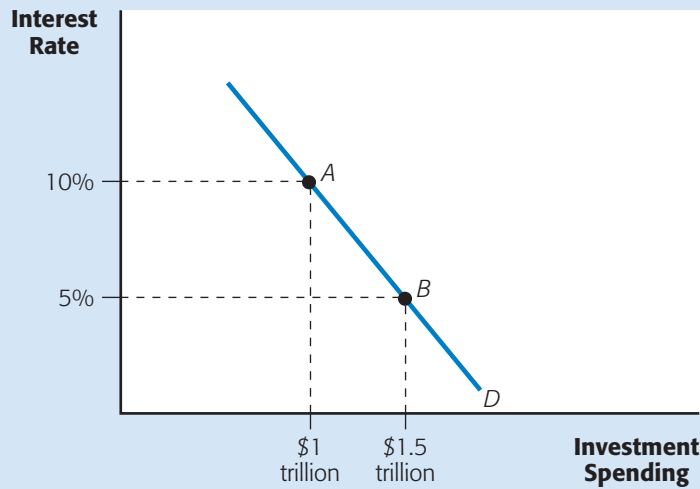


FIGURE 1

As the interest rate falls from 10 percent to 5 percent, each firm that buys a particular type of capital will buy more of it. As a result, the economy's total investment in physical capital rises from \$1 trillion to \$1.5 trillion. This is shown as the movement from point *A* to point *B* along the investment curve in the figure.

interest rate, investment spending, and the ultimate size of our capital stock is one reason that policy makers pay so much attention to the overall level of interest rates in the economy.

To recap:

Lower interest rates increase firms' investment in physical capital, causing the capital stock to be larger, and our overall standard of living to be higher.

INVESTMENT IN HUMAN CAPITAL

So far in this chapter, we've explored investment in *physical* capital. But now let's consider investment in *human* capital—the skills and abilities of the workforce. Like physical capital, these skills and abilities are long-lasting tools that make labor more productive in producing output. But unlike physical capital—which is owned by firms—human capital is ordinarily possessed by individual workers.

Economists are very interested in human capital investment. Here, we will concentrate on just two questions: First, who pays for workers to acquire human capital—the workers themselves, or the firms who employ them? We'll see that some types of human capital are usually paid for by firms, and other types are paid for by workers. Second, when an individual must pay to acquire human capital on his own, how does he make the decision? That is, how does an individual decide whether or not to acquire skills that would make him more valuable to an employer?

GENERAL VERSUS SPECIFIC HUMAN CAPITAL

Economists classify human capital into two categories, according to how broadly it can be applied in the workplace. Human capital that makes you more productive at *many* different firms is called **general human capital**. If you study engineering at college, for example, your knowledge will help you at any of thousands of

General human capital Knowledge, education, or training that is valuable at many different firms.

TABLE 4

TYPES OF HUMAN CAPITAL NECESSARY TO BE A SUCCESSFUL AEROSPACE ENGINEER AT GENERAL ELECTRIC

General	Ability to reason logically Mastery of mathematics and physical reasoning Knowledge of general engineering design principles Courses in thermodynamics, fluid mechanics, and heat transfer
Specific	Experience with General Electric jet engines Knowledge of the skills and abilities of other GE engineers Familiarity with the kinds of aircraft that use GE jet engines Understanding of GE's unique corporate structure and decision-making process

Specific human capital Knowledge, education, or training that is valuable only at a specific firm.

manufacturing firms across the country, including those that make aircraft, automobiles, and computer chips.

But there is also **specific human capital**, which is chiefly of value at a *specific firm*. For example, suppose you take a job as an engineer working on jet engines at General Electric and you learn specific details about the GE90 engine. That knowledge is specific human capital because it will be useful only if you continue working on GE jet engines. If you move to Pratt & Whitney, the specific details you've learned about the GE90 will be useless because they don't apply to Pratt & Whitney's engines.

Table 4 shows the types of human capital that you might need to be a successful aerospace engineer at General Electric.

The entries in the table that are general human capital would be useful not just at GE, but in many other firms as well, including other aircraft makers. But the entries that are specific human capital would have no value to any firm other than GE.

There is a very good reason for distinguishing between general and specific human capital. Firms have limited incentive to invest in general human capital because they cannot be sure of capturing all the benefits. To see why, suppose that a firm like General Electric were to pay for its employees to get engineering degrees. That would require a tremendous expenditure on tuition payments, to say nothing of the cost of the lost output for GE while its employees were in school rather than working, or the cost of replacing them with other, perhaps temporary workers. But once the employees graduate, there is no law requiring them to use their new skills as *General Electric* employees. They might decide to test the job market and find that a rival firm is willing to pay them a higher wage than GE pays. This rival firm, after all, did not bear the cost of educating the GE employees and therefore is better positioned to pay higher wages than is General Electric.

As you can see, a firm gains little by investing in its employees' general human capital. And few firms do so. In practice, individuals usually pay the cost of acquiring their own general human capital. You are probably doing that now as you study economics.

More generally,

employers have limited incentives to provide general human capital, since it increases the worker's value to many firms, and the worker will capture the benefits in the form of a higher wage. Therefore, workers must acquire general human capital on their own—or with the help of government subsidies.⁴

⁴ Of course, there are exceptions: Some firms *do* pay for their employees to finish college or to get professional degrees. But this is not very common, and the employees often must sign special contracts promising to work for a given number of years.

Notice the last phrase in the shaded statement. Governments often subsidize colleges and schools and provide grants and loans to students. Why? As individuals acquire general human capital, they become more productive and capable of earning higher wages. Therefore, they should be willing to pay the full cost of their education, even if it means borrowing money to do so.

But there is a problem: Most of those who attend college are young, and have not yet accumulated much wealth to be used as collateral for a student loan. Without government help, they would find it difficult to borrow for their educational spending. Since banks know through bitter experience that the default rate on student loans is quite high, they are reluctant to extend such loans unless the government guarantees them.

Because society as a whole benefits as students acquire more formal education, governments have stepped in to help people obtain schooling at all levels. They do this by running elementary and secondary schools, subsidizing colleges and universities, and providing low-interest loans to college students.

Now let's turn our attention to specific human capital, which is of value only to one specific employer. Individual workers are usually *not* willing to pay the cost of specific human capital. Why not? Because unlike general human capital, which ends up benefiting workers, specific human capital ends up benefiting the firm. For example, suppose an engineer at GE develops knowledge about the skills and abilities of other GE engineers—an example of specific human capital. Then she is no more valuable to Boeing or any other aircraft firm than she was before she acquired this knowledge. These other firms will *not* be willing to pay her any higher wage because of this specific human capital, and therefore, GE will not have to pay her a higher wage in order to keep her. Thus, the specific human capital does not benefit the worker in the form of a higher wage. But it *does* benefit GE, since the worker—although she is paid the same wage as before—is now more productive.

Of course, both workers and firms know that specific human capital benefits the firm, and so the firm is the one that ends up paying for it.

Individuals have little incentive to pay for specific human capital, since it increases their value to only one firm, and that firm will capture the benefits. Therefore, firms provide their workers with specific human capital at their own expense.

THE DECISION TO INVEST IN GENERAL HUMAN CAPITAL

Now that we've seen that individuals typically pay to acquire their own general human capital, how is the decision made? Let's take a specific example: Suppose an accountant must decide on purely economic grounds whether to take a specialized course in how to handle the books of entertainment companies. It's a costly course: \$30,000 in tuition and another \$25,000 in foregone income during the three months he is enrolled in the course. But the course will increase his income by \$10,000 per year for each of the next eight years, after which he plans to retire.

The principle of asset valuation plays a central role in the accountant's decision. That is,

to the worker that possesses it, human capital is an asset that generates higher income in the future. Therefore, the benefit of any given human capital investment is equal to the total present value of the additional future income.

At an annual interest rate of 10 percent, the total present value of the stream of extra revenue would be \$53,349. Since the course costs \$55,000, it's not worth it: The total present value of the additional income is *less* than the cost of the course. In purely economic terms, the accountant would be better off not taking the course.

But what if the annual interest rate were lower, say 8 percent? The cost of taking the course—\$55,000—would remain the same, because that cost is paid *now*. But the present value of future revenue would change. With a lower interest rate, the total present value of the additional income would be higher, at \$57,466. Thus, at an interest rate of 8 percent, the investment is worth it, since the benefit (measured in total present value) is now greater than its cost. In general:

Investment in human capital, like investment in physical capital, is inversely related to the interest rate. The lower the interest rate, the greater the benefits of any human capital investment, and the more human capital workers will want to acquire.

Moreover, the consequences of the change in investment are much the same for human capital as for physical capital. Recall what we learned earlier about physical capital: It makes us more productive as workers, and, as firms acquire more of it, the economy and our living standard grows. The same thing is true of human capital. The more we acquire, the more we can produce. Thus:

Lower interest rates encourage individuals to invest in general human capital. As a result, the total amount of human capital—and our overall standard of living—will be higher if interest rates are lower.



<http://>

At South-Western College Publishing's Finance Web site (<http://swcollege.com/finance/finance.html>) you can find a wide variety of material on financial markets.

Financial asset A promise to pay future income in some form, such as future dividends or future interest payments.

FINANCIAL MARKETS

You may be wondering what financial markets—like the markets for stocks and bonds—have to do with the other subject of this chapter: markets for capital. After all, capital—like machines and factories—is something *real*; it enables firms to produce real goods and services. The same is true of human capital: It enables real people to produce more real goods and services.

But in financial markets, the things being traded are just *pieces of paper*, which don't directly help anyone to produce anything. So what do these pieces of paper have to do with capital?

Actually, quite a bit. The pieces of paper being traded in financial markets are **financial assets**—promises to pay future income to their owner. Accordingly, the value of a financial asset is calculated in the same way as the value of any other asset, such as a truck or a computer: We find the *total present value* of the future payments that the asset will generate. Thus, our method of valuation is one connection between markets for capital and markets for financial assets.

But there is another connection between these two types of markets as well. Because capital lasts for many years, most firms fund their capital purchases by taking on financial obligations that themselves last many years. That is, to get the money to purchase trucks, factory buildings, office furniture, and other forms of capital, firms will usually issue long-term IOUs and exchange them for the needed funds. This leaves the firm with long-lasting capital, but also a long-lasting obligation to make future payments. Of course, the more capital a firm purchases, the greater the

value of the IOUs the firm will issue. So there is a close *economic* connection between a firm's decision to be a demander in a capital market and its decision to be a supplier in the financial markets.

In the rest of this chapter, we'll explore two types of financial assets: bonds and stocks. We'll also analyze the very well-publicized markets in which these assets are traded.

THE BOND MARKET

If a firm wants to buy a new fleet of trucks, build a new factory, or upgrade its computer system, it must decide how to finance that purchase. One way to do this is to sell **bonds**. A bond is simply a promise to pay a certain amount of money, called the **principal** or **face value**, at some future date. Although \$10,000 is the most common principal amount, you can also find bonds with face values of \$100,000, \$5,000, and other amounts.

A bond's **maturity date** is the date on which the principal will be paid to the bond's owner. If a bond has a maturity date 30 years after the date on which it was first sold, we'd call it a 30-year bond. Other bonds have shorter maturities—15 years, 10 years, 1 year, 6 months, or even 3 months.

Some bonds, including many of those sold by the U.S. federal government, are **pure discount bonds**. A discount bond is one that does not make any payments except for the principal it pays at maturity. For example, at some time in your life, you may have gotten a gift of a U.S. savings bond, issued by the federal government and sold at most banks. A \$100 savings bond is a promise by the federal government to pay \$100 to the bond's owner in, say, 30 years. If the savings bond sells for \$40 and pays \$100 at maturity, the total interest on the bond is \$60—the difference between what the bond originally sold for and what the owner will receive at maturity.

Most bonds, however, promise—in addition to repayment of principal—a series of interim payments called **coupon payments**. For example, a 30-year, \$10,000 bond might promise a coupon payment—say, \$600—each year for the 30 years, and then pay \$10,000 at maturity.

A bond's **yield** is the effective interest rate that the bond earns for its owner. The yield on a bond, as you will see later on, is closely related to the price that someone pays for the bond.

How Much Is a Bond Worth? To determine the value of a bond, let's start with a simple example: a pure discount bond that promises to pay \$10,000 when it matures in exactly one year. The \$10,000 is a future payment, and our method of calculating its value should not surprise you: It involves *present value*. Let's suppose the interest rate at which you can borrow and lend funds is 10 percent. Then we can determine the present value of the bond with our discounting formula as:

$$PV = \frac{\$Y}{(1 + i)} = \frac{\$10,000}{1.10} = \$9,091.$$

Since the present value of \$10,000 to be received in one year is \$9,091, that is the most you should pay for the bond. Assuming the bond's current owner can borrow and lend at the same 10 percent interest rate as you, then \$9,091 is the lowest price at which she will sell the bond to you. We conclude that this bond will sell for \$9,091—no more and no less.

Bond A promise to pay a specific sum of money at some future date.

Principal (face value) The amount of money a bond promises to pay when it matures.

Maturity date The date at which a bond's principal amount will be paid to the bond's owner.

Pure discount bond A bond that promises no payments except for the principal it pays at maturity.

Coupon payments A series of periodic payments that a bond promises before maturity.

Yield The rate of return a bond earns for its owner.

The same principle applies to more complicated types of bonds, such as discount bonds that don't pay off for many years, or coupon bonds. For example, suppose a bond maturing in five years has a principal of \$10,000, and also promises a coupon payment of \$600 each year until maturity, with the first payment made one year from today. The total present value of this bond would be:

$$PV = \frac{\$600}{(1.10)} + \frac{\$600}{(1.10)^2} + \frac{\$600}{(1.10)^3} + \frac{\$600}{(1.10)^4} + \frac{\$600}{(1.10)^5} + \frac{\$10,000}{(1.10)^5} = \$8,483.69.$$

Once again, this total present value—\$8,483.69—is what the bond is worth, and this is the price at which it will trade, as long as buyers and sellers use the same discount rate of 10 percent in their calculations.

Bond Prices and Bond Yields. There is an important relationship between the price of a bond and the yield or rate of return the bond earns for its owner. This is easiest to see with a pure discount bond, such as the bond that pays \$10,000 in one year in our example above. Suppose you were to buy this bond for \$8,000. Then, at the end of the year, you would earn interest of \$10,000 – \$8,000 = \$2,000 on an asset that cost you \$8,000. Your yield would be \$2,000/\$8,000 = 0.25 or 25 percent.

But now suppose you paid \$9,000 for that same bond. Then your interest earnings would be \$10,000 – \$9,000 = \$1,000, and your yield would be \$1,000/\$9,000 = 0.111 or 11.1 percent.

As you can see, the yield you earn on a bond depends on the price you pay for it. For each price, there is a different yield. And the greater the price of a bond, the lower the yield on that bond. This applies not only to simple discount bonds, but also to more complicated bonds with coupon payments. And the reasoning is the same in both cases: A bond promises to pay fixed amounts of dollars at fixed dates in the future. The more you end up paying for those promised future payments, the lower your rate of return.

More generally:

There is an inverse relationship between bond prices and bond yields. The higher the price of any given bond, the lower the yield on that bond.

What is true for a single bond is also true for bonds in general: When many bonds' prices are rising together, so that the average price of bonds rises, then the average *yield* on bonds must be falling.

Primary and Secondary Bond Markets. Every type of financial asset is traded in two different types of markets. The **primary market** is where newly issued financial assets are sold for the first time. But once a financial asset is sold in the primary market, the buyer is free to sell it to someone else. When a previously issued asset is sold again, the sale takes place in the **secondary market**. Most of the trading that takes place in financial markets on any given day is *secondary market trading*.

Applying this distinction to bonds, we would say that the *primary bond market* is where newly issued bonds are sold to their original buyers, while the *secondary bond market* is where previously issued bonds change hands.

It is only in the primary market that a firm actually obtains funds for its investment projects. Once a firm has issued and sold a bond, that bond can change hands many times in the secondary market, but the firm will not benefit directly from these sales. Secondary market trading is an exchange between private parties, and the original issuing firm or government agency is not involved.

Primary market The market in which newly issued financial assets are sold for the first time.

Secondary market The market in which previously issued financial assets are sold.

Still, firms and government agencies follow secondary bond markets closely. Why? It turns out that the secondary market has important feedback effects on the primary market, and thus affects those that want to borrow money by issuing bonds. The link between these two markets arises because most bonds offered for sale in the primary market have very close substitutes available in the secondary market. For example, suppose that IBM wants to borrow funds by issuing 10-year, \$10,000 bonds in the primary market. In order to attract buyers, it will have to sell these *new* bonds at the same price as any *old* \$10,000 IBM bonds trading in the secondary market that still have 10 years left before maturity. After all, there is no reason for a bond buyer to prefer a new, 10-year bond to an old bond that has 10 years left to run if both are issued by the same corporation, and both have the same face value. Thus,

while bond issuers are not directly participants in secondary market trading, they are affected by what happens in the secondary market. More specifically, if a bond's price rises in the secondary market, the price one can charge for similar, newly issued bonds in the primary market will rise as well.

Since there is such a close relationship between bond prices and bond yields, we can also express this idea in terms of yields.

If a bond's yield falls in the secondary market, the yield of similar, newly issued bonds in the primary market will fall as well.

A bond's yield is what a firm ends up paying when it issues bonds and sells them in the primary market. So a firm would like its bond yield to be as small as possible (its bond price to be as high as possible).

Why Do Bond Prices (and Bond Yields) Differ? Thousands of different kinds of bonds are traded in financial markets every day. There are corporate bonds of various maturities and bonds issued by local, state, and federal governments and government agencies. Bonds issued by foreign firms and governments are also traded in the United States. And each bond has its own unique yield. Why is this? Why don't all bonds give the same yield? That is, why doesn't each bond sell at a price that makes its yield identical to the yield on any other bond?

The answer is found in the principle of asset valuation, which tells us that a bond—like any asset—is worth the total present value of its future payments. Imagine that you are a bond trader and you are trying to determine the maximum price you should offer for a bond. You know the face value of the bond and its maturity date, as well as the values and dates of any coupon payments it might make. Your problem then boils down to determining what discount rate to use in calculating the total present value of those future payments. That is, you must determine which discount rate will accurately reflect the opportunity cost of your funds.

If you were *absolutely certain* that you would receive the promised future payment, then your discount rate should be the interest rate you *could* earn on *other*, absolutely certain investments. The promises made by the U.S. government are generally considered the most reliable, and the interest rate on U.S. government securities is often called the *riskless rate*. So, if you have the same faith in the bond you are considering buying as you would in U.S. government bonds, then you should use the interest rate on government bonds as your discount rate, and calculate the total *PV* accordingly.

However, few bonds are as safe as U.S. government bonds. Indeed, private firms do occasionally go bankrupt and default on their obligations—some recent examples include Barneys in 1996, Montgomery Ward in 1997, and Boston Chicken in



Don't think that the risk of default is the only risk in owning a bond. A bond holder also takes a risk because interest rates in the economy might change in the future. Why is this a risk for a bond holder? Remember that the price of any bond is equal to the total present value of its future payments. As you've learned, higher interest rates result in lower present values and lower prices for bonds. Thus, if you want to sell a bond after you've bought it, a rise in interest rates will force you to sell it at a lower price—possibly even a price lower than the price you paid. Even government bonds—which carry no default risk—still carry interest rate risk.

Of course, interest rates can move in your favor, too. But the fact remains that buying a bond is always a gamble because interest rates can change in your favor or your disfavor.

1998. The bond market is alert to the likelihood of default, and bonds are rated according to this likelihood. Moody's, one of the services that rates bonds, classifies bonds as Aaa (the least likely to default), followed by Aa, A, Baa, and so on. When a bond has a higher likelihood of default, the opportunity cost of your funds to buy it is greater than just the interest foregone because you

are also foregoing safety—you risk losing the entire value of the bond. Therefore, for riskier bonds, your discount rate should include the opportunity cost of foregone interest that you could have earned on U.S. government bonds, *plus* an extra premium reflecting the higher risk. And the riskier the bond, the higher the discount rate you should apply to it, and the lower will be its total present value.

To put a value on riskier bonds, market participants use a higher discount rate than on safe bonds. This leads to lower total present values and lower prices for the riskier bonds. With lower prices, riskier bonds have higher yields.

Table 5 shows that the market does value bonds in this way. It shows the yields on different types of bonds as of January 14, 2000. Notice how the yields diverge. The difference between the riskless yield of 6.68 percent and the risky Baa yield is more than 1.7 percentage points. That difference is the premium that compensates investors for the chance that a Baa bond will go into default in a given year.

The bonds of economically unstable foreign governments often have high risks of default, and these bonds can carry high yields as a result. For example, throughout the mid-1990s, the yield on Russian government bonds that promised repayment in U.S. dollars was more than triple the yield on U.S. government bonds. Buyers of Russian bonds did not have complete faith that the Russian government would be able to obtain the dollars to make good on its promise of repayment. Therefore, they needed to be compensated for the risk of default. Sure enough, in August 1998, the Russian government *did* default on its debt, causing bond holders, both individuals and large money-center banks, to lose billions of dollars.

Riskiness is only one reason that bond prices and bond yields differ. If you go on to study financial economics, you'll learn that two bonds with equal default risk can

TABLE 5

**INTEREST RATE ON BONDS,
JANUARY 14, 2000**

Rating	Interest Rate
Federal government bond	6.68 percent
Aaa corporate bond	7.84 percent
Aa corporate bond	8.00 percent
A corporate bond	8.18 percent
Baa corporate bond	8.42 percent

Source: Moody's Investors Service Web site
<http://www.moodys.com/economics.nsf>

have different yields for a variety of reasons, including differences in their maturity dates, differences in their frequency of coupon payments, or because one bond is more widely traded (and therefore easier to sell on short notice) than another.

THE STOCK MARKET

A **share of stock**, like a bond, is a financial asset that promises its owner future payments. But the nature of the promise is very different for these two types of assets. When a corporation issues a bond, it is *borrowing* funds and promising to pay them back. But when a corporation issues a share of stock, it brings in new ownership of the firm itself. In fact, a share of stock *is*, by definition, *a share of ownership* in the firm. Those who pay for their shares provide the firm with the funds, and in return, the firm owes them—at some future date or dates—a share of the firm's profits.

For example, in January 2000, Lycos, Inc.—the Web media company and developer of the Lycos Internet search engine—wanted to raise funds to finance a further expansion. It could have borrowed the funds by issuing bonds and selling them in the primary market, but instead, it issued new shares of stock, thereby bringing in new owners.

When a firm wishes to raise money in the stock market, it gets in touch with an investment bank. Investment banks are firms that specialize in assessing the market potential of new stock issues. Together, the firm and its investment banker develop a prospectus that describes the offering—the nature of the firm's business, the number of shares that will be sold, and so on. The purpose of the prospectus is to inform potential investors of the risks involved. It must be reviewed by the *Securities and Exchange Commission*, the principal regulatory agency that oversees financial markets.

Once the prospectus is approved, the firm can sell shares to the public. If it is the first-ever offering of shares by this firm, the sale will be called an *initial public offering (IPO)*. The firm's investment banker usually tries to line up buyers for the offering before the securities are actually released for sale. In practice, it's usually large institutional investors, such as mutual funds, who first purchase new shares.

Primary and Secondary Stock Markets. When a corporation issues new shares—as part of an IPO or a secondary offering—they are sold in the *primary stock market*. When Lycos, Inc. sold 6 million shares at \$77.375 each, the firm hoped to receive $\$77.375 \times 6$ million, or about \$464.3 million. Out of those proceeds, it paid its investment bankers and kept the rest to spend on expanding its operations. This is the only time the corporation received any income from sales of this stock. From then on, the stock traded in the *secondary market*—the market in which previously issued shares are sold and resold.

As in the bond market, the issuing corporation has no *direct* relationship with the secondary market. But the secondary market is very important to firms that raise funds in the primary market, for two reasons. First, because of the secondary market, people who buy shares know they can easily sell them when they want. This makes people more willing to hold stock, including the new shares that firms issue to raise funds.

Second, price changes in the secondary market affect the price a firm can get from selling shares in the primary market. In fact, when a firm's shares are already trading in the secondary market, a small offering of new shares will always sell at the secondary market price. That's because the firm's new shares are perfect substitutes for the shares already trading in the secondary market. If the price drops in the secondary market, the price of new shares must drop the same amount in order to be as attractive to buyers as secondary market shares. But this means that the firm will

Share of stock A share of ownership in a corporation.

raise less money with its public offering. A serious drop in share prices in the secondary market can even lead a firm to cancel a public offering, and postpone any investment projects that the offering was supposed to fund.

Direct and Indirect Ownership of Stock. Many people own shares of stock directly. You or a family member may have purchased stock for your own account, by calling a broker or going online and ordering, say, 200 shares of barnesandnoble.com stock. The stock is then held by your brokerage firm, and you are free to buy more or sell it any time you want, with a phone call or an online order.

Mutual fund A corporation that specializes in owning shares of stock in other corporations.

But you can also own stock *indirectly*, by purchasing shares of a **mutual fund**. A mutual fund is a corporation that, in turn, buys shares of stock in *other* corporations. There are mutual funds that specialize in Internet companies, in foreign companies located in specific regions like Europe or Asia, and in long-lived companies that have a reputation for stable, if not growing, profits. Most mutual funds suggest that, by doing careful research into companies and making professional predictions about the future, they can pick stocks within their specialty more wisely than a nonprofessional can. (We'll discuss the accuracy of this claim in the Using the Theory section of this chapter.)

A final way that households can—indirectly—own stock is through retirement funds that are managed by their employers. These accounts should not be confused with retirement accounts—called 401(k) and 403(b) accounts—that employees manage for themselves, in which the stock is owned directly or indirectly through mutual fund shares. By contrast, when a worker's retirement fund is managed by an employer, the total funds available for retirement will depend on the performance of the stock and bond markets, but the worker has no ability to buy and sell shares of individual bonds, stocks, or mutual funds on his own. It is not unusual for half or more of the funds in such retirement accounts to be held in stocks, with most of the rest in bonds.

Stock ownership in the United States is growing rapidly. In 1999, Americans held more wealth in the stock market than in the value of their own homes. Fully 48 percent of Americans owned shares of stock or mutual fund shares that they managed themselves, up from about 19 percent in 1983. If we included those with employer-managed retirement accounts that include stocks, the percentage of Americans with a stake in the stock market would be much higher.

Why Do People Hold Stock? Why do so many individuals and fund managers choose to put their money into stocks? You already know part of the answer: When you own a share of stock, you own part of the corporation. Indeed, the fraction of the corporation that you own is equal to the fraction of the company's total stock that you own. For example, in January 2000, Tommy Hilfiger, the clothing manufacturer, had 95 million shares outstanding. If you owned 10,000 shares of Tommy Hilfiger stock, then you owned $10,000/95,000,000 = 0.000105$, or about one one-hundredth of one percent of that firm. That means that you are, in a sense, entitled to a hundredth of a percent of the firm's after-tax profits.

Dividends Part of a firm's current profit that is distributed to shareholders.

In practice, however, most firms do not pay out *all* of their profit to shareholders. Instead, some of the profit is kept as *retained earnings*, for later use by the firm. The part of profit that is distributed to shareholders is called **dividends**. A firm's dividend payments benefit stockholders in much the same way that interest payments benefit bondholders, providing a source of steady income. Of course, as a part owner of a firm, you are part owner of any retained earnings as well, even if you will not benefit from them until later.

Aside from dividends, a second—and usually more important—reason that people hold stocks is that they hope to enjoy **capital gains**—the returns someone gets when they sell an asset at a higher price than they paid for it. For example, if you buy shares of Compaq computer at \$30 per share, and later sell them at \$35 per share, your capital gain is \$5 per share. This is in addition to any dividends the firm paid to you while you owned the stock.

Some stocks pay no dividends at all, because the management believes that stockholders are best served by reinvesting all profits within the firm so that *future* profits will be even higher. The idea is to increase the value of the stock, and create capital gains for the shareholders when the stock is finally sold. America Online, for example, pays no dividends but had a total stock value of \$136.8 billion in January 2000. It got this value because investors expected it to pay dividends at some point—and they expected the dividend stream to grow thereafter. Another example is Microsoft, which has never paid a dividend but had a value of around \$595 billion in early 2000. Microsoft's shareholders had great faith that they would eventually start to get cash from the company.

Over the past century, corporate stocks have generally been a good investment. They were especially rewarding during the 1990s, enjoying (on average) a 15 percent annual return. That means that the average \$1,000 invested in the stock market on January 1, 1990 would have increased in value to \$4,045 by the beginning of 2000.

Valuing a Share of Stock. The value of a share of stock, like any other asset, is the total present value of its future payments. For a share of stock, the future payments are all the profits that the share is expected to earn for its owner. But over what time horizon should stocks be valued? Unlike a bond, which has a maturity date, a share of stock is expected to remain an earning asset for some owner for as long as the company exists—forever, unless market participants anticipate the firm will go out of business at some future date. Fortunately, economists and mathematicians have developed formulas to measure the total present value of a firm's future profits under a variety of different assumptions. For example, the simplest formula tells us that, if a firm will earn a constant \$ Y in profit after taxes each year forever, then the total present value of these future profits is $\$Y/i$, where i is the discount rate. So, for example, if a firm is expected to earn \$10 million in after-tax profit for its owners per year forever, and the discount rate is 10 percent, then—according to the formula—the total *PV* of those future profits is $\$10 \text{ million}/0.10 = \100 million . If there are 1 million shares of stock outstanding for this firm, then each share should be worth $\$100 \text{ million}/1 \text{ million} = \100 .

The value of a share of stock in a firm is equal to the total present value of the firm's after-tax profits, divided by the number of shares outstanding.

Note that we are valuing a share of stock by future profits, not by dividends. Remember that firms often plow their profits back in the firm in order to increase the firm's growth rate further. What counts is after-tax profits, because these belong to the firm's shareholders, whether they receive them in cash or not.

However, when valuing the shares of real-world companies—companies whose earnings are expected to grow, and companies whose future earnings involve some risk—the simple formula we've just used is too limiting. Other, more complicated formulas have to be used, and you will learn some of them if you go further in your study of economics or business. But even without knowing the detailed formulas,

Capital gain The return someone gets by selling a financial asset at a price higher than they paid for it.

we can come to four important conclusions about the factors that can affect a stock's value.

First, earnings forecasts are usually based on the firm's current earnings. The total present value of the firm's future profits will be greater if those profits are rising from a higher base of current profit. Thus,

an increase in current profits increases the value of a share of stock.

Second, for any given base value of current profit, a higher anticipated growth rate will raise the profit expected in each future year, which will raise the total present value of the firm's profits. Hence,

an increase in the anticipated growth rate of profits increases the value of a share of stock.

Third, as you've learned, a higher discount rate decreases the present value of any payment to be received in the future. Thus,

a rise in interest rates—or even an anticipated rise in interest rates—decreases the value of a share of stock.

Finally, there is the matter of risk. In making financial decisions, most people prefer a sure thing to a gamble (although there are exceptions). We adjust for risk in our *PV* calculations by applying a higher discount rate to future payments that are more risky. This means that the *PV* of any future year's profits will be lower when the amount of those future profits is less certain. Accordingly,

an increase in the perceived riskiness of future profits decreases the value of a share of stock.

Reading the Stock Pages. In the United States, financial markets are so important that stock and bond prices are monitored on a continuous basis. If you wish to know the value of a stock, you can find out instantly by checking with a broker or logging on to a Web site that reports such information. One such site is Thomson Investors Network (<http://www.thomsoninvest.net/index.sht>) but there are dozens of others. In addition, stock prices and other information are reported daily in local newspapers and in specialized financial publications such as the *Wall Street Journal*.

To some people, the pages that cover the stock market look as impenetrable as Egyptian hieroglyphics. But in fact, the information on the stock pages is very easy to understand, once you decide to learn it.

Figure 2 shows an excerpt from the New York Stock Exchange Composite Transactions reported in the February 9, 2000 *Wall Street Journal*. The data refer to the previous trading day—Tuesday, February 8, 2000.

Let's focus on one typical stock—The Gap (listed as Gap Inc), a large retailer. The first two columns in the table show the highest and lowest prices paid for the stock during the previous 52 weeks. You can see that The Gap's stock ranged in price from a high of $\$53\frac{3}{4}$ per share to a low of $\$30\frac{1}{16}$. By tradition, stock prices were always quoted in fractions of a dollar. However, beginning in 2000, the Secu-

STOCK MARKET TABLE, FEBRUARY 9, 2000

FIGURE 2

52 Weeks	Hi	Lo	Stock	Sym	Div %	PE	30d Vol	Hi	Lo	Close	Net Chg
11%	8 1/4		Goldilocks	GCY	2.8	8.1	31	80	9 1/4	9 1/4	+ 1/4
25%	23 1/2		Goldilocks pl		2.0	8.3	6	24 1/4	24	24 1/4	+ 1/4
12%	10%		Goldilocks	GAB	1.4	11.9	1407	12 1/4	12 1/4	12 1/4	+ 1/4
25%	20%		Goldilocks pl		1.81	7.9	67	23	22 1/4	23	+ 1/4
19%	11		Goldilocks	GOT	3.0	10.8	481	18 1/4	18 1/4	18 1/4	+ 1/4
25%	22		Goldilocks pl		1.98	8.3	14	23 1/4	23 1/4	23 1/4	+ 1/4
10%	7%		Goldilocks	GUT	.88	7.5	185	8 1/4	8	8	—
25%	20%		Goldilocks	GBP	2.12	9.4	14	22 1/4	22 1/4	22 1/4	— 1/4
25%	19%		Goldilocks pl		2.38	11.1	88	18 1/4	18 1/4	18 1/4	— 1/4
8%	3 1/4		Goldilocks	GMA	.87	1.2	del	15	6 1/4	6 1/4	—
5%	3 1/4		Goldilocks	GML	del	280	2 1/4	2	2 1/4	+ 1/4	
55%	22 1/4		Goldilocks	GMC	36	1.8	11	20 1/4	20 1/4	20 1/4	— 1/4
65%	4 1/4		Goldilocks	AUG	1.85	2.8	18	52 1/4	52 1/4	52 1/4	— 1/4
25%	12%		Goldilocks	GLM	1.46	10.9	1378	14 1/4	13 1/4	13 1/4	— 1/4
83%	80%		Goldilocks	EDC	.84	1.2	21	53 1/4	53 1/4	53 1/4	+ 1/4
53%	20%		Gap Inc	GPS	.09	.2	35	54 1/4	50 1/4	50 1/4	+ 1/4
21%	11		Gap Inc	EDI	del	12	231	18 1/4	17 1/4	17 1/4	+ 1/4
25	9 1/4		Gap Inc	IT1	1.9	8.2	20	24 1/4	24 1/4	24 1/4	— 1/4
25	9 1/4		Gap Inc	ITB	del	424	12 1/4	12 1/4	12 1/4	—	
84	28 1/4		Gap Inc	GTW	del	45	28 1/4	28 1/4	28 1/4	— 3/4	
30%	23%		Gap Inc	GET	.80	2.8	10	408	29 1/4	28 1/4	— 1 1/4
13%	7%		Gap Inc	GT	.36	4.8	4	10 1/4	7 1/4	7 1/4	— 1/4
102%	58%		Gap Inc	DNR	del	1876	158	153 1/4	158 1/4	+ 7 1/4	
22%	13%		Gap Inc	CHR	del	1837	18 1/4	18 1/4	18 1/4	+ 1/4	
39%	30%		Gap Inc	GAM	5.98	13.8	171	37 1/4	36 1/4	36 1/4	— 1/4
25%	20%		Gap Inc	GAM pl	1.88	7.7	78	23 1/4	23 1/4	23 1/4	— 1/4
20%	9%		Gap Inc	BGC	26	2.2	8	205	9 1/4	9 1/4	+ 1/4
5%	1%		Gap Inc	BGC pl	28	8.4	3	88	2 1/4	2 1/4	+ 1/4
-14%	5%		Gap Inc	BPP	del	3	281	14 1/4	14 1/4	14 1/4	— 1/4
9%	2%		Gap Inc	BDC	del	1448	9 1/4	9 1/4	9 1/4	+ 1/4	
75%	42%		Gap Inc	GD	36	2.2	9302	44 1/4	42 1/4	43	— 1 1/4
189%	95%		Gap Inc	GE	1.64	1.2	42	138 1/4	136 1/4	137 1/4	+ 1 1/4
38%	25		Gap Inc	GEF	2.04	7.8	15	1734	29	28 1/4	— 1/4
25%	17%		Gap Inc	GEF pl	1.81	8.8	del	66	20 1/4	20 1/4	+ 1/4
43%	23%		Gap Inc	GIS	1.18	3.8	16	113 1/4	31	30 1/4	— 1/4
47%	55%		Gap Inc	GM	2.08	2.5	94	405	80 1/4	78 1/4	— 1
			Gap Inc				1436	139 1/4	134 1/4	137 1/4	— 2

Source: *The Wall Street Journal* (February 9, 2000), page C6.

urities and Exchange Commission began requiring that U.S. stocks be priced in decimals, as the rest of the world does. Under the system, a $\$53\frac{3}{4}$ price would be reported as $\$53.75$.

The next columns show the stock's name—abbreviated to Gap Inc—followed by its stock symbol—GPS. You may need to know the stock symbol if you want to find a stock's price online, or on a “ticker tape”—the continuous report of stock trades that runs from wall to wall in many financial institutions or at the bottom of the screen on CNBC television network.

The next two columns report the firm's most recent cash dividend—in this case .09, or 9 cents per share—and the corresponding *dividend yield*, obtained by dividing the most recent year's dividends by the current stock price. For The Gap, this was .2, meaning that if the dividend had been paid on January 31, each share would pay a dividend equal to two-tenths of one percent of the stock's current value. The price-earnings (PE) ratio, shown in the next column, is the stock's current price divided by its after-tax profit per share during the previous 12 months. Or, put another way, the PE ratio tells us the cost of each dollar of yearly after-tax profits. The figure of 35 means that The Gap's current stock price was 35 times the size of its most recent earnings per share, or that—if you bought this stock—you would be paying \$35 for each dollar of yearly profits that the firm was currently earning.

Many people watch PE ratios closely. They theorize that a stock with a low PE ratio is a better deal, since it costs less per dollar of profit. But this strategy can be deceiving. A company's PE ratio might be very low because its future prospects aren't very good. People may not be expecting much growth in the firm's future profits, or they may even be expecting its profits to fall, so they won't pay a very high price for each dollar of *current* earnings. On the other hand, a company whose profits are expected to grow rapidly might command a very *high* PE ratio. People are willing to pay a higher price for this stock because they expect profits to grow, but the PE ratio will be high because it measures the price of a dollar of *current* profits, rather than future profits. In general, an unusually high or unusually low PE ratio does not tell people whether the stock is relatively expensive or relatively cheap; one must also consider the stock's future prospects.

The remaining columns tell us about the most recent day's transactions in this stock—February 8, 2000 in this case. The column headed *Vol 100s* indicates how many shares, in hundreds, traded on that day. Multiplying the tabled figure of 25472 by 100, we find that about 2.55 million shares changed hands on that day. The *Hi*, *Lo*, and *Close* columns that come next show that on February 8, the price per share ranged from a high of \$50½ to a low of \$49⅞, and that it ended the day at \$50⅙. The final column—*Net Chg*—tells us that the price of a share of Gap stock increased by ⅙, or a little over 6 cents per share, from its price at the end of the *previous* day's trading.

In addition to reporting on individual stocks, the *Wall Street Journal* and other newspapers also report on changes in different stock market averages or indexes. These are meant to represent movements in stock prices as a whole, or movements in particular types of stocks. The most popular average is the **Dow Jones Industrial Average**, which tracks the prices of 30 of the largest companies in the United States, including AT&T, IBM, and Wal-Mart. Another popular average is the much broader **Standard & Poor's 500**, which tracks stock prices of 500 large corporations.

Dow Jones Industrial Average An index of the prices of stocks of 30 large U.S. firms.

Standard & Poor's 500 An index of the prices of stocks of 500 large U.S. firms.

Explaining Stock Prices. Glancing at the newspaper clipping in Figure 2, you can see that most stocks experience a price change on any given day. Why? Like all prices, stock prices are determined by supply and demand. However, our supply and demand curves require a bit of reinterpretation.

Figure 3 presents a supply and demand diagram for the shares of The Gap. Unlike most supply curves you've studied in this book—which tell you the quantity of something that suppliers want to *sell* over a given period of time—the supply curve in Figure 3 is somewhat different. It tells us the quantity of shares *in existence* at any moment in time. This is the number of shares that people are *actually* holding.

On any given day, the number of Gap shares in existence is just the number that The Gap has issued previously, up until that day. Therefore, no matter what happens to the price today, the number of shares remains unchanged. This is why the supply curve in the figure is a vertical line at 851 million, showing that there are 851 million shares in existence regardless of the price.

Now, just because 851 million shares of Gap stock actually exist, that does not mean that this is the number of shares that people *want* to hold. The desire to hold Gap shares is given by the downward-sloping demand curve. As you can see, the lower the price of the stock, the more shares of The Gap people will want to hold. Why is this?

As you've learned, the value of a share of stock to any owner is equal to the total present value of its future after-tax profits. However, individuals do not all cal-

Characterize the Market



THE MARKET FOR EXISTING SHARES OF THE GAP

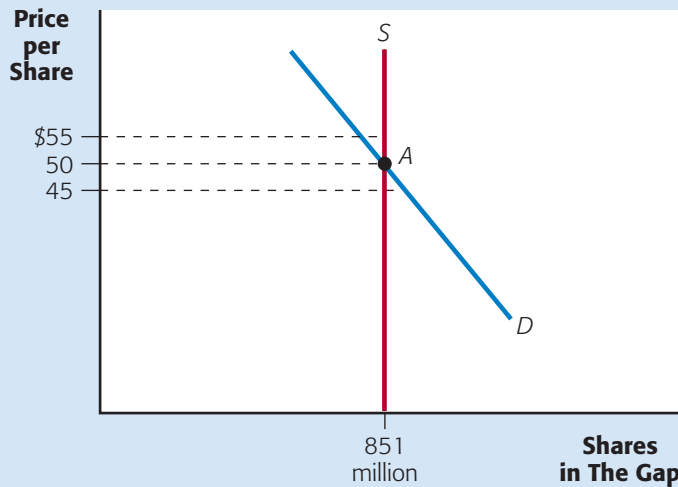


FIGURE 3

On a given day, the number of shares of a firm's stock is fixed. Here, the vertical supply curve shows that 851 million shares of stock in The Gap exist regardless of the price per share. The downward-sloping demand curve shows that different investors have different views of The Gap's prospects. The equilibrium price of \$50 per share is determined at point *A*, where the number of shares demanded equals the number in existence.

culate this total present value in the same way. Some may believe that The Gap's profits will continue to grow as rapidly as they have in the past, while others—more pessimistic—may believe that The Gap's best days are behind it, forecasting a much lower growth rate. Some investors may not mind risk much at all, while others may be especially risk averse, and use a higher discount rate that lowers the present value of each future year's profit.

Thus, at any given moment, there is an array of estimates of a stock's total present value. As the price of the stock comes down, it descends below more and more people's total present value estimates, and so more and more will find the stock to be a bargain, and want to hold it. This is what the downward-sloping demand curve tells us.

Now, looking at Figure 3, you can see that at any price other than \$50 per share, the number of shares people *are* holding (on the supply curve) will differ from the number they *want* to hold (on the demand curve). For example, at a price of \$45 per share, people would want to hold more shares than they are currently holding. Many would try to buy the stock, and the price would be bid up. At \$55 per share, the opposite occurs: People find themselves holding more shares than they want to hold, and they will try to get rid of the excess by selling them. The sudden sales would cause the price to drop. Only at the equilibrium price of \$50—where the supply and demand curves intersect—are people satisfied holding the number of shares they are *actually* holding.

Stocks achieve their equilibrium prices almost instantly. Legions of stock traders—both individuals and professional fund managers—sit poised at their computers, ready to buy or sell a particular firm's shares the minute they feel they have an excess supply or a shortage of those shares. Thus, we can have confidence that the price of a share at any time is the equilibrium price.

But why do stock prices *change* so often? Or, since stocks sell at their equilibrium prices at almost every instant, we can ask: Why do shares' *equilibrium* prices change so often?

Since a supply curve, like that in Figure 3, only shifts when there is an initial or secondary public offering, and these happen only occasionally and with great

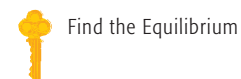
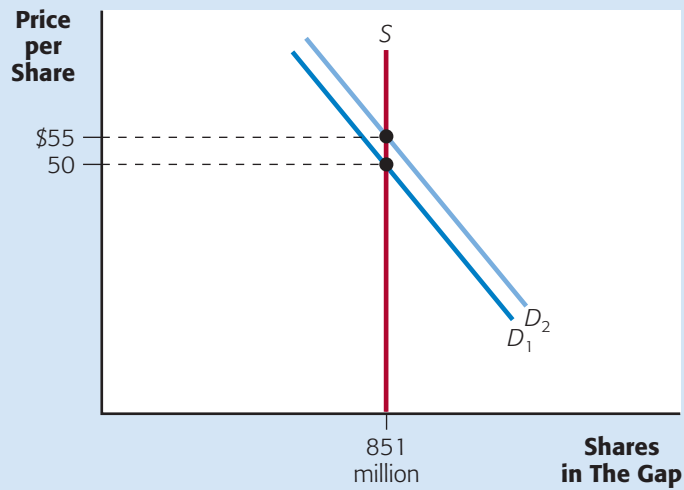


FIGURE 4

Demand for shares of The Gap will increase if (1) interest rates fall; (2) the perceived riskiness of the stock decreases; or (3) the firm's earning prospects brighten. In the figure, the demand curve shifts rightward, from D_1 to D_2 , driving up the price from \$50 to \$55 per share.

AN INCREASE IN DEMAND FOR SHARES OF THE GAP



What Happens When Things Change?



fanfare, the day-to-day changes in equilibrium prices cannot be caused by shifts in the supply curve. So they must be caused by shifts in *demand*. Figure 4 shows how a rightward shift in the demand curve for shares of The Gap could cause the price to rise to \$55 per share. Indeed, on rare occasions, the demand curve for a firm's shares has shifted so far rightward in a single day that the share price doubled or even tripled.

But what causes these sudden shifts in demand for a share of stock?

The logic of present value provides the answer. Anything that causes large groups of individuals to change their estimates of the total present value of future profits will shift the demand curve. For example, the discount rate used in *PV* calculations will *decrease* whenever there is a decrease in interest rates in the economy. It will also decrease if future earnings become more certain. By making the discount rate, d , smaller, these changes would increase the total present value of a share, and shift the demand curve to the right (people would want to buy more shares at any price). Similarly, an increase in current profit beyond what was expected, or an increase in the expected growth rate of profits will increase the total present value of shares, and shift the demand curve rightward.

When stock prices move dramatically, it is usually because some new information has become available. For example, suppose that The Gap were to announce that it is opening up 100 new stores in China. If people believe that these new stores will help increase The Gap's profit, their estimates of this stock's total present value will rise, shifting the demand curve to the right. As in Figure 4, this will increase the equilibrium price of the stock. On the other hand, it may be that one of The Gap's rivals announces a special sale. In that case, people may expect the rival's sales to increase at the expense of The Gap's, or that The Gap will have to lower prices in order to keep its market share. If so, they forecast lower profit per share, the demand curve shifts to the left, and the price falls (not shown in the figure).

THE ECONOMIC ROLE OF FINANCIAL MARKETS

Now that we've investigated some of the specific details regarding financial markets, it is worthwhile to back up and take a broader view. What functions do financial markets play? In this section, we will take an economist's viewpoint and try to pinpoint just exactly how financial markets make us all better off.

If there is a single word that resonates throughout this chapter, it is *time*. Markets for physical and human capital as well as financial markets reflect decisions made over time. When a firm purchases a capital asset, it makes an expenditure today in return for a machine or plant that generates benefits many years into the future. When an individual invests in human capital, something similar happens—costs are incurred today in exchange for future benefits.

In the absence of *capital* markets, we would all be constrained to live as if there were literally no tomorrow. We would have to forego the productivity advances embodied in new capital goods and the conceptual breakthroughs that arise from investment in education and training. Each of us—and society as a whole—would be much poorer.

We would also be poorer if there were no *financial* markets. Firms would be unable to become very large or grow very fast if they were constrained to fund their growth solely through retained earnings. Without capital markets, there would be no AT&T, no IBM, and no Microsoft, and we would not be able to enjoy the products these firms produce. All three of these firms—and indeed, most major corporations—turned to the stock and bond markets to obtain funds for their capital acquisitions.

Moreover, without financial markets, we would be constrained as individuals. We could save for retirement or for our children's education, but not very fruitfully, because we would not earn any interest or dividends on our savings. Without financial markets, banks would be little more than safe-houses, storing our cash until we needed it, and charging us a fee for the service instead of paying us interest.

All of the markets we have studied in this chapter enable us to save funds and earn a rate of return, and they enable firms to invest and grow. They help relax the economic constraints imposed by scarcity. And they certainly contribute to the high standard of living we enjoy. When savers and borrowers come together in financial markets, both sides benefit. Let's look more closely at some of the economic functions that financial markets play.

1. *Facilitating large-scale production.* The large industrial enterprises that are so common today are only about a century old. In the nineteenth century and earlier, it would have been extremely unusual to find any business employing a hundred workers, much less the thousands of employees that are common in today's firms. But as technology changed and innovations such as rail transportation, electricity, and the gasoline engine were introduced, it suddenly became possible to have firms that served national, rather than local markets. And as firms grew, so did their need to accumulate large sums of money.

To take just a single example, think about railroads. As the rail network spread across the United States, the railroads needed to (1) assemble large tracts of land for rail beds, stations, and other facilities; (2) purchase steel rails and hire the labor necessary to lay thousands of miles of track; and (3) invest in locomotives and rolling stock. These were huge tasks, requiring larger sums of money than firms could possibly accumulate from their retained earnings. Therefore, the new railroads turned to the bond market for funding, just as today's large enterprises look to both the stock and bond markets for cash to expand their operations. Without

smoothly functioning financial markets, little of the remarkable economic growth of the past century could have occurred.

2. *Reallocating spending across time.* We've seen that financial markets allow firms to invest in new projects today rather than waiting until the necessary funds accumulate from current operations. Something similar is true for individual households—financial markets allow them to reallocate their consumption over time. To see this, imagine that you wish to buy a new car. If there were no financial markets, then you would have to wait until you could save up the needed funds. With financial markets, you can take out a loan that enables you to increase your consumption now at the cost of reducing consumption later (as you repay the loan). Or imagine that you are concerned about income during retirement. The markets provide a way for you to accumulate the necessary funds. When you open a savings account, buy a bond, or invest in a mutual fund, you are reducing your consumption today in order to enjoy greater consumption in the future.

More generally, the financial markets reallocate funds from surplus units—mostly individuals who are not consuming their entire incomes today—to deficit units—mostly firms that desire to spend more than their current income today. In so doing, those markets help individuals, firms, and even the government, to achieve the best *intertemporal* utilization of resources.

3. *Reducing risk.* In the absence of financial markets, firms could still invest and households could still save. But doing so would be much riskier. Imagine, for example, that you had some extra money and you wanted it to grow. Then you could get in touch with a local business and offer to lend it money in return for future dividends or interest payments. If you are lucky, and the firm thrives, you will come out ahead. But what if the firm encounters hard times, or even goes out of business? With all your eggs in one basket, so to speak, your investment is quite risky.

Now let's replay the scenario—this time with financial markets. Again you have money to invest, but now you have many more options. Rather than putting all your funds into one firm, you can buy shares of stock in a variety of firms, or shares in a mutual fund. As we saw in our discussion of the CAPM, portfolio diversification is a good way to reduce risk. And such diversification is really only possible with well-functioning stock and bond markets.

4. *Disciplining management.* Every market determines a price, and financial markets are no exception. But the prices of stocks and bonds serve several important functions that are not obvious at first glance. One such function is providing instantaneous feedback that allows corporate managers to see how they are doing. The price of a share of Delta Airlines stock, for example, tells Delta's managers how the market perceives their policies. If the stock price increases, and there has been no change in the discount rate, it means that thousands of investors are—collectively—giving a vote of confidence in those policies. They believe that Delta's future earnings will be greater than they thought before the price rose. If the price decreases, that means investors are voting with their dollars against the way the firm is being managed.

Or imagine that you are the Chief Financial Officer at a new Internet startup company. Your firm has great prospects for growth, but you need to secure funds in the stock market. What kind of payoff can you expect from selling your stock? By checking the prices of your competitors' shares, you can form at least a rough estimate of what the market is willing to pay for your shares. And once you begin to participate in the stock market, the existing share price will give you an indication of how much money you can raise by selling additional shares.

CAN ANYONE PREDICT STOCK PRICES?

Every day, financial news programs, such as *Wall Street Week* or CNBC's *Squawk Box*, offer stock market advice to millions of television viewers. The stock market analysts interviewed on these shows tell us that they have done some careful research, or that they have a secret formula, and that by following their advice, you'll earn more dividends and capital gains than you could hope to earn on your own. Of course, for the *really* good predictions, you'll have to pay a price, and subscribe to their private newsletter or use them as your stockbroker.

It may surprise you to hear that the vast majority of economists don't believe them. Economists, as a rule, don't believe that *anyone*—no matter how smart, no matter how much research they do—can do much better than you, an introductory economics student, reading this book and finding out about the stock market for the first time. In fact, they don't believe that anyone can predict what will happen to stock prices much better than someone who has *never* taken economics, and who chooses which stocks to buy by throwing darts at the stock page.

How can this be? We'll answer this question by first considering the two different methods used by analysts to make their predictions. Both of these methods try to predict shifts in the demand curve for a stock, like the one in Figure 4 (p. 398). But the methods they use to make their predictions are very different.

PREDICTING STOCK PRICES: FUNDAMENTAL ANALYSIS

One widely practiced method for predicting stock prices is **fundamental analysis**. As its name suggests, fundamental analysis focuses on the *fundamental forces* driving a firm's future earnings, and the value placed on those earnings by stock market participants. Fundamental analysts study data on overall economic conditions, on specific industries, and on individual firms. To try to predict what will happen to share prices for specific firms, they consider the products made by a company, the future demand for these products, and the strategic moves of current or future competitors to the firm in question. They will also study the firm's top management and try to assess how smart and creative they are.

Using these methods, fundamental analysts try to determine whether a stock is undervalued or overvalued relative to the rest of the market. If it is undervalued in their view, they will recommend that their client or employer buy the stock. If it is overvalued, they will recommend that the stock be sold.

PREDICTING STOCK PRICES: TECHNICAL ANALYSIS

Another method for predicting stock prices—which seems to become more popular every year—is **technical analysis**. The basic idea is that you can graph the recent behavior of a stock's price and, based on certain patterns, predict whether the stock is going to increase or decrease in value over the near future. Technical analysts believe that everything you need to know to predict a stock's future price changes is contained in the stock's past behavior. Many technical analysts recommend that their clients or employers buy and sell particular firms' stocks based on past price movements, without even knowing what the firm produces!

Technical analysts believe that stocks move in trends. And they have numerous, colorful names for the patterns they claim to see. For example, there is a pattern



Fundamental analysis A method of predicting a stock's price based on the fundamental forces driving the firm's future earnings.

Technical analysis A method of predicting a stock's price based on that stock's past behavior.

called “head and shoulders” that appears when a stock’s price hits a high, then falls a little, then rises to a new high, falls again, then rises a third time. If this third rise—the right shoulder—fails to equal the previous rise, then many technical analysts expect a major decline in value.

Many people find technical analysis appealing because there are, indeed, elements of strategic behavior in stock market investing. When you attempt to forecast what will happen to your 50 shares of General Motors during the next six months, it is not enough to understand GM’s prospects and the demographic factors affecting the demand for automobiles. You also need to predict how other GM shareholders see things. If for some reason, many of them think GM shares will plunge in value, then they will sell their shares, thereby driving down their price. Your shares will decline in value as well. So, thinking about the stock market seems akin to the kind of game-theoretic reasoning we encountered in Chapter 10. It seems that each market participant has to determine what other participants are going to do. This is a daunting task that some people think can be handled by looking for patterns in stock prices, and they trust technical analysts to find those patterns.

THE ECONOMIST’S VIEW: EFFICIENT MARKETS THEORY

While economists believe that fundamental and technical analysis can often explain stock price movements in the *past*, they are extremely skeptical about anyone’s ability to *predict* stock price changes in the future. This is because economists tend to take the **efficient markets** view of the stock market. According to this view, the stock market digests new information that might affect stock prices *efficiently*—that is, rapidly and thoroughly.

The implications of the efficient markets view are startling. First, it means that you cannot, on average, beat the market by doing research and finding and buying underpriced (or selling overpriced) stocks. You cannot do this because any research that *you* do will also be done by others and is therefore already incorporated into the stock’s price. That means—if the goal is to outperform a broad stock market average like the Standard & Poor’s 500—both fundamental and technical analysis are largely a waste of time.

For example, fundamental analysis tells us that if a company comes up with a valuable new patent, the total present value of its future profits will rise. As a result, the demand curve for the firm’s stock will shift rightward, as in Figure 4 (p. 398), and the equilibrium price will rise. But who benefits from this price rise? If the patent announcement is a surprise, only those lucky enough to be holding the stock when the announcement is made can benefit. That’s because the demand shift and the adjustment to the new equilibrium will be virtually instantaneous. All those who hold the stock will immediately adjust their asking price upward, so no one will be able to buy at a price lower than the equilibrium.

But what if the announcement *isn’t* a surprise? What if, by doing careful research, you can predict which companies are *about* to come out with valuable new patents? Surely, then, your research would pay off, enabling you to buy a stock when its price is low, then sell it at a higher price when the announcement comes out and surprises everyone else. Right?

Sorry to say, but according to the efficient markets view, this is dead wrong. Because any research that *you* can do can and will also be done by *others*. Therefore, while you may be able to figure out which companies are likely to succeed, that information will already be reflected in the price of the stock. For example, the stock

Efficient market A market that instantaneously incorporates all available information relevant to a stock’s price.

of companies *more likely* to announce valuable patents will already have a higher price than the stock of companies less likely to do so.

According to the efficient markets view of the stock market, any information that can be used to predict a stock's future earnings will be incorporated into a stock's price as soon it becomes publicly available. Therefore, by the time a fundamental analyst predicts that a stock's price will rise or fall, it has already risen or fallen. Fundamental analysis cannot help you outperform the market.

What about technical analysis? After all, it's human beings who buy and sell stocks, and their decisions are based on human psychology. Surely, a brilliant technical analyst, who carefully studies buying and selling decisions of millions of people, and who can discover the secret psychological rules that govern stock trading by divining patterns amidst the chaos . . . surely *he* can outperform the market.

Sorry to say, but the efficient markets view is skeptical about this idea, too. Why? Imagine a very simple pattern: Because of exuberance or fatigue or superstition, people are more likely to buy stocks than to sell them on Friday, so on average, stock prices rise every Friday. Since everyone would anticipate this pattern, they would buy stocks on Thursday, hoping to profit from the Friday runup. But this would cause stocks to rise on Thursday, not Friday, so people would buy on Wednesday, and so on. Soon, there would be no patterns at all—Friday would be like any other day. While this is a very simple example, the logic applies to *any* pattern a technical analyst might uncover.

According to the efficient markets view of the stock market, any patterns in stock price movements that can be observed by a good technical analyst will be incorporated into stock prices as soon as they are discernable. Therefore, stock market patterns disappear as soon as anyone can discover them. Technical analysis cannot help you outperform the market.

Efficient markets theory tells us that the only information that affects the stock market is surprise information—a new announcement of a major technological breakthrough, or even new information that suggests a firm *might* achieve such a breakthrough. And the only people who benefit from this information when it is made public are those who are lucky enough to be holding the stock already, before the information was available at all.⁵

Moreover, efficient markets theory says that there are *no observable patterns* in stock price movements. This, in turn, means that individual stocks, and broad averages like the Dow Jones Industrial Average or the Standard & Poor's 500, will exhibit entirely *random* changes. If a stock—or the Dow Jones Industrial Average—has fallen three days in a row, the likelihood that it will fall again is no different than if it had fallen, risen, and then fallen.

Efficient markets may at first seem to be a gravity-defying theory of prices. Why do we spend so much effort learning how stock prices are determined, only to then learn that their changes are random? The reconciliation lies in understanding that it

⁵ Another group that can benefit from information is *insiders*—those with connections to the firm, and have access to information *before* it becomes public. They can buy or sell stock early, before information is reflected in the price of the stock. Profiting from insider information is illegal. Those who do so—if they are caught—pay stiff fines and sometimes even go to jail. However, enforcement of insider trading laws is difficult, since it is often hard to detect.

is *because* so much effort is put into figuring out what price stocks should sell for that price changes are random. Today's price reflects everything known today, by the market as a whole, about the stock. As a result, the price can change only if new information arrives. But information is new only if it was unexpected—that is, random. If we knew that the price would rise, it would already have risen.

The theory of efficient markets is one of the most exhaustively tested theories in all of economics. Thousands of studies have confirmed the efficiency of stock prices with respect to all sorts of information. You can't beat the market by buying stocks only in companies whose presidents went to MIT (or anywhere else). You can't beat the market by buying stock only in companies in growing industries. You can't beat it by buying stocks that have risen. You can't beat the market by buying stocks that have collapsed. You can't beat the market—period!

But wait. Every year, some fundamental and technical analysts seem to do remarkably well, and *do* outperform the market. And if you watch any television program on investing, you will see them being interviewed and making predictions further into the future. Doesn't this contradict efficient markets theory?

Not at all, and here's why. In any large group of people picking stocks, we would always expect some to be unusually lucky, just as we'd expect some to be unusually unlucky. In fact, we'd expect this even if no one in the group knew *anything* about the stock market—even if, say, they chose which stocks to buy by throwing darts at the stock page. Of course, the *unlucky* stock pickers will never be interviewed; only the lucky ones will get the attention. But the evidence shows that outperforming the market in one year—even by a lot—makes an analyst no more or less likely to outperform the market the next year.

Although the idea of efficient markets is sweeping and rules out a great many investment strategies as worthless, its implications for the investor who understands it can be quite valuable.

First, just because you can't outperform the market doesn't mean you shouldn't invest in the market at all. The average stock's price, over long periods of time, tends to rise. In fact, if dividends and capital gains are added together, stocks—over the long run—earn their holders a better yield than bonds. This is because stocks are more risky, and investors in the stock market must be compensated for bearing that risk.

Second, if someone asks you to pay for their stock-picking advice, *don't*. You can do just as well by picking stocks on your own, even if you pick them randomly. The stocks you pick will be as likely to rise or fall as stocks chosen by an expert.

Third, because you have to pay commissions when you trade stocks, you should trade as little as possible. By using a “buy and hold” strategy, you can participate in the long-run, higher-than-bonds rate of return at minimum expense.

Finally, choose a diversified portfolio with different stocks that tend not to rise and fall together. Such a portfolio will have less risk than an undiversified portfolio with the same expected rate of return. The investor who follows the implications of the efficient markets hypothesis will assemble a diversified set of stocks and then hold on to them, buying and selling only when new cash comes in or cash needs to be taken out.

S U M M A R Y

Physical capital, human capital, and financial assets all provide future benefits to their owners that can be bought outright by purchasing or hiring the asset that generates them. The principle of asset valuation tells us how firms and individ-

uals determine the value of any long-lived asset—as the total present value of all the future income the asset will generate.

A firm's demand for physical capital reflects the marginal revenue product of capital—the marginal product of capital

multiplied by the price of the firm's product. Because of diminishing marginal productivity, the marginal revenue product of capital generally declines as more capital is acquired. The value of an additional unit of physical capital is the *total present value* of all future years' marginal revenue products. This total present value will be smaller when interest rates are higher. Therefore, higher interest rates discourage investment in physical capital.

Human capital can be divided between general human capital—valuable at many firms—and specific human capital—mostly valuable at just one firm. While firms will generally pay for their workers to acquire specific human capital, it is up to individual workers to acquire their own general hu-

man capital. Higher interest rates discourage investment in human capital, just as they do for physical capital.

There are many types of financial markets, including those for bonds and corporate stock. The price of a bond will equal the total present value of its future payments. The value of a share of corporate stock is the total present value of the future after-tax profits of the firm, divided by the number of shares outstanding. This value depends on the firm's current profit, the expected growth rate of profits, the interest rate in the economy, and the risk associated with the firm's future profits. In an efficient market, the price of corporate shares will reflect all available information. There will be no predictable patterns in stock price movements that can be exploited for profit.

KEY TERMS

marginal revenue product of capital	general human capital	coupon payments	capital gain
present value	specific human capital	yield	Dow Jones Industrial Average
discounting	financial asset	primary market	Standard & Poor's 500
discount rate	bond	secondary market	fundamental analysis
investment	principal (face value)	share of stock	technical analysis
principle of asset valuation	maturity date	mutual fund	efficient market
	pure discount bond	dividend	

REVIEW QUESTIONS

1. What is the marginal revenue product of capital? How is it calculated, and how is it related to the demand for capital?
2. Why is \$100 received today more valuable than \$100 received one year from today?
3. What is the present value of \$1,000 to be received two years from today? Assume that the relevant interest rate is 10 percent per year.
4. What is the relationship between the present value of a future payment and (1) the size of that payment, (2) the interest rate, and (3) the date at which the payment will be received?
5. What is the principle of asset valuation? Give examples of how it would be used to value a piece of physical capital, general human capital, and a bond.
6. Give examples of general and specific human capital (other than those presented in the chapter).
7. Explain the relationship between:
 - a. a bond's price and its yield
 - b. a bond's price and the riskiness of the firm that issued it
 - c. a bond's price and its face value
8. Why would a corporation care about the price of its stock in the secondary market?
9. What are the economic roles played by financial markets? How do they help the economy operate more efficiently and grow more rapidly?
10. What is the efficient markets view of the stock market? What are its main implications?

PROBLEMS AND EXERCISES

1. You are considering buying a new laser printer to use in your part-time desktop publishing business. The printer will cost \$380, and you expect it to produce additional revenue of \$100 per year for each of the next five years. At the end of the fifth year, it will be worthless. Answer the following questions:
 - a. What is the value of the printer to you if the annual interest rate is 10 percent? Is the purchase of the printer justified?
 - b. Would your answer to part (a) change if the interest rate were 8 percent? Is the purchase justified in that case? Explain.

- c. Would your answer to part (a) change if the printer cost \$350? Is the purchase justified in that case?
- d. Would your answer to part (a) change if the printer could be sold for \$500 at the end of the fifth year? Is the purchase justified in that case? Explain.

What lessons can you derive from your answers to these questions?

2. Your firm is considering purchasing some computers. Each computer costs \$2,600, and each has an annual marginal revenue product. Because you plan to use the computers for different purposes, you have ranked those purposes in descending order or annual *MRP_k* as follows:

Computer	Annual <i>MRP_k</i>
1	\$3,000
2	\$2,000
3	\$1,000
4	\$ 500

- a. Assume that each computer has a useful life of three years, and no value thereafter. If the interest rate is 10 percent per year, how many computers should you purchase?
- b. If, before you purchased the computers, the interest rate decreased to 5 percent per year, how many computers would you purchase?
3. In each of the following cases, determine what would happen to the amount of human capital that individuals or firms would decide to invest in.

- a. State governments invest significant amounts of money in building new colleges and universities.
- b. New teaching methods increase the amount of knowledge that students accumulate in each course.
- c. The overall interest rate in the economy increases.
- d. Because employers are seeking a more skilled workforce, the average wage rate for college graduates increases.

4. Explain in what sense labor markets are similar to capital markets, and in what sense they are different.
5. Good news! Gold has just been discovered in your backyard. Mining engineers tell you that you can expect to extract five pounds of gold per year forever. Gold is currently selling for \$400 per ounce, and that price is not expected to change. If the interest rate is 5 percent per year, estimate the total value of your gold mine.
6. One year ago, you bought a two-year bond for \$900. The bond has a face value of \$1,000 and has one year left until maturity. It promises one additional interest payment of \$50 at the maturity date. If the current interest rate is 5 percent per year, what capital gain (or loss) can you expect if you sell the bond today?
7. Suppose that people are sure that a firm will earn annual profit of \$10 per share forever. If the interest rate is 10 percent, how much will people pay for a share of this firm's stock? Suppose now that people become uncertain about future profits, causing them to use a discount rate of 15 percent. How much will they pay now?

CHALLENGE QUESTIONS

1. Suppose you are thinking about attending medical school. Your medical education will cost \$15,000 per year, and you expect to receive your M.D. degree in four years. (The annual costs of \$15,000 are the opportunity cost of your education; they include such items as foregone wages and tuition, but do not include food and shelter, which you would consume in any case.) Suppose you expect that, as a result of becoming a physician, your earnings will be \$5,000 per year higher than they would have been had you gone to work immediately after earning your bachelor's degree. Assume that the interest rate is 10 percent per year and that your working life expectancy is 20 years. Is the decision to attend medical school justified as an economic investment? Identify the factors that could change the judgment about medical school as an investment.
2. The asset value formula can be modified to account for variable interest rates over time. For a three-year time horizon, the modified formula would be:

$$\text{Value} = \frac{Y_1}{(1+i_1)} + \frac{Y_2}{(1+i_1)(1+i_2)} + \frac{Y_3}{(1+i_1)(1+i_2)(1+i_3)}$$

where i_1 , i_2 , and i_3 are the interest rates in years 1, 2, and 3, respectively. Suppose a firm is considering two projects—A and B—with the following costs and revenues:

Project	Cost	Year 1 Revenue	Year 2 Revenue	Year 3 Revenue
A	50	20	20	20
B	33	20	30	40

Use this information to determine which of the projects should be undertaken if:

- a. The sequence of interest rates is $i_1 = 0.1$, $i_2 = 0.11$, $i_3 = 0.121$ (i.e., interest rates grow by 10 percent per year starting from an interest rate of 10 percent).

- b. The sequence of interest rates is $i_1 = 0.1$, $i_2 = 0.09$, $i_3 = 0.081$ (i.e., interest rates decline by 10 percent per year starting from an interest rate of 10 percent).
- c. What lesson can you derive from your answers in parts (a) and (b)?

EXPERIENTIAL EXERCISES

1. Go to Thomson Investors Network (<http://www.thomsoninvest.net/stocks/intro.sht>) and click on “Stocks.” Next, pick a stock that you find interesting and use the Stock Center to learn more about it. What factors, discussed in this chapter, are affecting the current value of this firm’s stock price?



2. The *Wall Street Journal* is an excellent source of information regarding U.S. financial markets. Most of the Money and Investing section is devoted to reporting on individual firms’ financial activities and on the stock and bond markets generally. This section is worth scanning every day, but today try to find an article related to an Initial Public Offering (IPO) of stock. What factors, discussed in this chapter, are influencing the price at which the new shares are being offered?



CHAPTER

14

ECONOMIC EFFICIENCY AND THE COMPETITIVE IDEAL

CHAPTER OUTLINE

The Meaning of Efficiency

Pareto Improvements

Side Payments and Pareto Improvements

The Elements of Efficiency

Productive Efficiency
Allocative Efficiency

Economic Efficiency and Perfect Competition: A Summary

The Inefficiency of Imperfect Competition

Where Do We Go from Here?

Using the Theory: The Collapse of Communism

In the late 1980s, a process of cataclysmic economic change began to sweep the world. An economic system under which more than 30 percent of the world's population lived began to unravel. The system—Soviet-style centrally planned socialism—had prevailed in the Soviet Union for more than 70 years and in China and Eastern Europe for 40 years. In Eastern Europe, the change was surprisingly abrupt, as country after country—Poland, Hungary, East Germany, and Czechoslovakia—dismantled the old economic order. In December 1991, the Soviet Union—the world's second most powerful country—ceased to exist, and 15 new countries—eager to abandon the old economic system—rose to take its place. In China, the economic change has been more gradual, but no less profound. At the beginning of the twenty-first century, China's 1.4 billion people were beginning to enjoy a Western-style freedom to start businesses, seek jobs, trade with foreigners, and own financial assets.

Many powerful forces combined to destroy and largely discredit Soviet-style central planning around the world. These included corruption within the highest levels of government, alienation and cynicism among the population, the universal desire for democracy and individual freedom, and strong nationalism within the Soviet republics. But there was an additional reason for the system's demise—a purely *economic* reason: It was deeply inefficient.

In this chapter, we take a close look at the concept of efficiency. You will learn that there is more to this concept than appears at first glance. You will also learn why economists believe that the market system is able to achieve a higher level of efficiency than any other economic system devised so far.

In many situations, however, the market does not work properly. Thus, in order to achieve the highest possible level of efficiency, the government must play an active role in the economy. Economics has much to say about the need for government intervention, and the type of intervention needed, to remedy different types of problems in the private economy. In the next chapter, we focus on the role of government in enabling and improving economic efficiency.

THE MEANING OF EFFICIENCY

What, exactly, do we mean by the word *efficiency*? We all use this word, or its opposite, in our everyday conversation: “I wish I could organize my time more

efficiently,” “He’s such an inefficient worker,” “Our office is organized very efficiently,” and so on. In each of these cases, we use the word *inefficient* to mean “wasteful” and *efficient* to mean “the absence of waste.”

In economics, too, efficiency means the absence of waste—although a very specific kind of waste: *the waste of an opportunity to make one person better off without making anyone else worse off*. More specifically,

Economic efficiency is achieved when there is no way to rearrange the production or allocation of goods in a way that makes one person better off without making anybody else worse off.

Notice that economic efficiency is a limited concept. While it is an important goal for a society, it is not the only goal. Most of us would list fairness as another important social goal. But an economy can be efficient even if most people are poor and a few are extraordinarily rich—a situation that many of us would regard as unfair.

An efficient economy is not necessarily a fair economy.

Why, then, do economists put so much stress on efficiency, rather than on issues of fairness? Largely because one’s definition of fairness depends on ethical and moral values, about which there is considerable disagreement in our society. Issues of fairness must therefore be resolved politically.

But virtually all of us would agree that if we fail to take actions that would make some people in our society better off *without harming anyone*—that is, if we fail to achieve economic efficiency—we have wasted a valuable opportunity. Economics—by helping us understand the pre-conditions for economic efficiency, and teaching us how we can bring about those pre-conditions—can make a major contribution to our material well-being.

PARETO IMPROVEMENTS

Imagine the following scenario: A boy and a girl are having lunch in elementary school. The boy frowns at a peanut butter and jelly sandwich, which, on this particular day, makes the girl’s mouth water. She says, “Wanna trade?” The boy looks at her chicken sandwich, considers a moment, and says, “Okay.”

This little scene, which is played out thousands of times every day in schools around the country, is an example of a trade in which both parties are made better off, and no one is harmed. And as simple as it seems, such trading is at the core of the concept of economic efficiency. It is an example of a *Pareto* (pronounced puh-RAY-toe) *improvement*, named after the Italian economist, Vilfredo Pareto (1848–1923), who first systematically explored the issue of economic efficiency.

A *Pareto improvement* is any action that makes at least one person better off, and harms no one.

Pareto improvement An action that makes at least one person better off, and harms no one.

In a market economy such as that in the United States, where trading is voluntary, literally hundreds of millions of Pareto improvements take place every day. Indeed, every purchase is an example of a Pareto improvement. If you pay \$30 for a pair of jeans, then the jeans must be worth more to you than the \$30 that you parted with, or you wouldn’t have bought them. Thus, you are better off after making the purchase. On the other side, the owner of the store must have valued your \$30 more

highly than he valued the jeans, or he wouldn't have sold them to you. So he is better off, too. Your purchase of the jeans, like virtually every purchase made by every consumer every day, is an example of a Pareto improvement.

The notion of a Pareto improvement helps us arrive at a formal definition of economic efficiency:

Economic efficiency A situation in which every Pareto improvement has occurred.

Economic efficiency is achieved when every possible Pareto improvement is exploited.

This definition can be applied to an individual market or to the economy as a whole. For example, suppose we look at the market for laser printers and cannot identify a single Pareto improvement in that market that has not already been exploited. No matter how hard we look, we cannot find a change in price or output level, or any other change for that matter, that would make some producer or some consumer better off without harming anyone. Then we would say that the market for laser printers is economically efficient.

Alternatively, we can look at the economy as a whole. If we discover remaining Pareto improvements that are not occurring—say, a change in the price of some good or a change in the quantity of a good produced—then we would deem the economy economically *inefficient*.

Of course, no economy can exploit *every* Pareto improvement, so no society can ever be *completely* economically efficient according to our definition. But achieving something close to economic efficiency is an important goal. When we look at real-world markets and real-world economies, it is best to view economic efficiency as a continuum. At one end of the continuum are economies in which, in most markets, most opportunities for Pareto improvements are exploited. At the other end of the continuum are economies in which many markets are economically inefficient—where many opportunities for mutual gain remain unexploited. As you will see in this chapter, perfectly competitive markets tend to be economically efficient, and market economies tend to lie closer to the economically effi-

cient end of the spectrum than other types of economies.



Deciding what is and what is not a Pareto improvement can often be confusing. For example, suppose you are in the desert, about to die of thirst, and you come upon a stand that sells bottled water. “How much for the bottle?” you ask. “Let me see your wallet,” says the owner of the stand. You hand over your wallet, and the owner quickly assesses its contents: \$200. “That’ll be \$200 per bottle.” You are so desperate for the water that you agree.

Was this a Pareto improvement? Absolutely. The water was worth much more than \$200 to you (without it, you would have died), so you are definitely better off. The seller benefited as well, since he has presumably realized quite a large profit. And no one was harmed by this transaction.

But wait . . . didn't the seller of the water take advantage of you? How can such a clear example of exploitation be considered a desirable Pareto improvement? To understand why it is desirable, remember that characterizing an action as a Pareto improvement only means that both sides benefit from the action; it doesn't tell us whether the total benefit is *distributed* between the two parties in a manner we would consider *fair*. In this example, both parties are better off if they trade, rather than not trade. Thus, it is a Pareto improvement. The lesson to remember is that a Pareto improvement is not necessarily fair or equitable, but both parties are always better off for making it.

SIDE PAYMENTS AND PARETO IMPROVEMENTS

The examples of Pareto improvements we have considered so far involve easily arranged transactions, in which one person trades with another and both come out ahead. Since both parties benefit, they have every incentive to find each other and trade.

But there are more complicated situations, involving groups of people, in which a Pareto improvement will come about only if one side makes a special kind of payment to the

other, which we call a *side payment*. Here's an example: Suppose a dry cleaning shop sets up on the ground floor of an apartment building. Everyone who lives in the building suffers from fumes and loud noise, and they want the dry cleaner to move. But because of a zoning law that prevents the entry of additional dry cleaners into the area, the dry cleaner is making an economic profit. He does not want to move. Thus, we seem to be at an impasse: If the dry cleaner moves to another location (say, in the competitive business district of the city), the tenants would gain, but the dry cleaner would be harmed—he would lose his economic profit. Is it possible to make at least one party—the dry cleaner or the tenants, or both—better off without simultaneously harming anyone? Let's see.

Suppose there are 100 tenants, and each would gladly pay an extra \$50 per month—a total of $100 \times \$50 = \$5,000$ —to get the dry cleaner out. Suppose, too, that the dry cleaner's economic profit is \$3,000 per month. The tenants might get together and agree to pay the dry cleaner \$4,000 per month to move from the building. If the deal is struck, the tenants are better off, since they gain benefits that are worth at least \$5,000 per month to them, but actually pay only \$4,000 per month. The dry cleaner is better off, since he loses \$3,000 in monthly profit, but gains \$4,000 in monthly payments from the tenants. In other words, by arranging a proper side payment from the tenants to the dry cleaner—compensation to leave the building—everyone can be made better off. As you can see, there are many side payments that would do the trick: Tenants could pay any amount between \$3,000 and \$5,000 per month, and both parties would still gain.

We'll be looking at quite a number of Pareto improvements in this chapter and the next one. To help keep track, we'll illustrate each of them with a scorecard, to show that nobody involved in the deal comes out behind. For our apartment tenants and dry cleaner, the scorecard looks like this:

Action: Tenants pay dry cleaner \$4,000 per month to move out of building.

Dry cleaner	Gains payment of:	\$4,000 per month
	Loses profit of:	\$3,000 per month
	Comes out ahead by:	\$1,000 per month
<hr/>		
Tenants	Gain benefits worth:	\$5,000 per month
	Pay:	\$4,000 per month
	Come out ahead by:	\$1,000 per month

Notice that, without the side payment, making the dry cleaner leave the building would *not* be a Pareto improvement, since the dry cleaner would be harmed. But the side payment converts an action that would harm someone into an action that harms no one—a Pareto improvement.

Some actions that—by themselves—would not be Pareto improvements can be converted into Pareto improvements if accompanied by an appropriate side payment.

THE ELEMENTS OF EFFICIENCY

Economic efficiency can be broken down into two components: *productive efficiency* and *allocative efficiency*. As the names suggest, productive efficiency involves arranging production to get the maximum possible output from our available

resources. Allocative efficiency, on the other hand, deals with *which* goods and services the economy should produce. For the next several pages, we'll concentrate exclusively on productive efficiency and then turn to an analysis of allocative efficiency.

PRODUCTIVE EFFICIENCY

Productive efficiency has to do with how well we use our resources in *producing* goods and services.

Productive efficiency When it is impossible to produce more of one good without producing less of some other good.

*An economy is **productively efficient** when it is impossible to produce more of one good without producing less of some other good.*

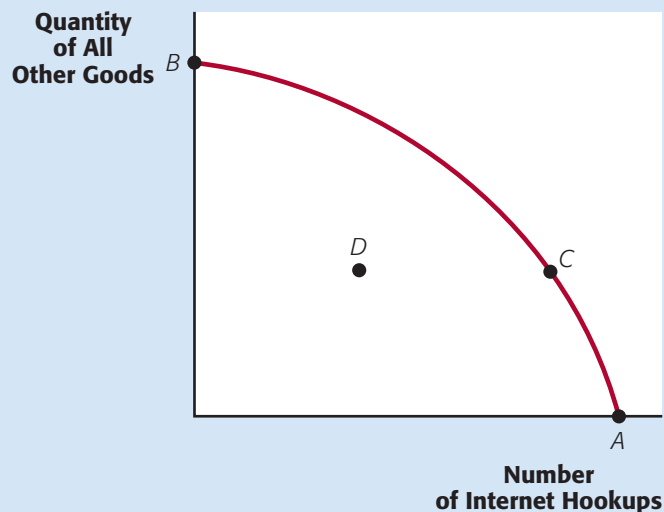
To understand this definition, let's consider its opposite: productive *inefficiency*. We first looked at this concept in Chapter 2—in the section titled, “The Search for a Free Lunch.” There, you learned that a productively *inefficient* economy could produce more of one thing *without* producing any less of something else. Clearly, such a society is not producing all that it could. It is wasting opportunities to produce more goods and services. A productively efficient economy, by contrast, does *not* waste any opportunities to produce more output.

Let's review what you learned about productive efficiency in Chapter 2. Figure 1 shows a production possibilities frontier for the economy. The horizontal axis measures the number of households hooked up to a high-speed Internet connection during the year, while the vertical axis measures the production of all other goods and services. Point *A* represents the maximum number of high-speed hookups we could produce each year if we threw *all* our resources—our land, labor, and capital—into the production of the necessary goods and services for high-speed hookups. Of course, since at point *A* we are using all of our resources for hookups to the Internet, we would produce nothing else. Point *B*, by contrast, shows the maximum possible production of other goods and services, but implies that *none* of our resources are used to provide high-speed Internet hookups.

FIGURE 1

PRODUCTION POSSIBILITIES BETWEEN INTERNET CONNECTIONS AND OTHER GOODS

The PPF illustrates the trade-off between the production of two types of goods. Along the curve, we can produce more of one type of good only by producing less of the other. To be productively efficient, the economy must operate *along* its PPF. Any point inside the frontier is inefficient. For example, at point *D*, it is possible to produce more high-speed Internet hookups without producing less of other goods—by moving to *C*.



Movements along the *PPF* illustrate the trade-off that exists between the production of these two categories of goods. For example, if the economy is initially operating at point *B*, then in order to move to point *C* (to produce more high-speed hookups), resources must be pulled out of producing other goods: growing wheat, producing television shows, making furniture, and so on, and put to use installing fiber optic cable, producing modems, and providing help lines for confused customers. Thus, moving along the curve, we see that producing more high-speed hookups means producing less of other goods, and vice versa.

What has this got to do with productive efficiency? Quite a bit.

In order to be productively efficient, an economy must be operating on its PPF.

For example, suppose we are located at point *D*, *inside* the *PPF*. Then, for the given amount of other goods, the economy is *not* producing the maximum quantity of high-speed hookups it could produce. By moving to a point like *C*, we could hook up more people each year without producing less of anything else. Therefore, at *D*, the economy is productively *inefficient*.

Productive efficiency is crucial for achieving the highest possible standard of living. Resources are scarce—there is not enough land, labor, and capital to produce all of the goods and services that people dream of. If there is productive inefficiency, then somebody—or possibly everybody—can enjoy a higher standard of living by correcting the inefficiency.

Three Requirements for Productive Efficiency. We can understand productive efficiency better by looking at the conditions that must be satisfied for an economy to achieve it. The three conditions for productive efficiency are:

1. The economy must use all of its available resources (full employment).
2. Each firm must produce the maximum amount possible from the resources available to it.
3. The allocation of inputs among firms must produce the maximum possible amount of output.

Let's consider each of these conditions in turn.

Full Employment of Resources

To be productively efficient, the overall economy must be operating at full employment, making use of all resources offered by resource owners.

Unemployed resources are the most obvious form of inefficiency. When people who want to work are not working, when buildings and land are unused, or when machines are idle, there is an opportunity to produce more output by putting those resources to work. Even well-organized economies like that of the United States have unemployed resources. Many school buildings remain completely empty in the summer, and many actors remain completely unemployed for months at a time, even though the buildings and actors could be used for theatrical productions that would have value to the public. (In many cities, entrepreneurs or government officials have found a way to make this happen, but not in all cities.)

Occasionally, a *recession* hits the economy. Unemployment of workers and other resources rises. A recession represents a wasted opportunity to produce more output and moves the economy farther away from productive efficiency. We can see this in Figure 1: A recession puts an economy inside its *PPF*, at a point like *D*. If we



Because of local regulations, this cab will have to return to Virginia without any passengers—an example of productive inefficiency.

could end the recession and put the unemployed to work, we could produce more of some things without having to produce less of anything else. The causes of recessions, and what the government can and cannot do to end them, are macroeconomic problems, and ones we will not address in this chapter.

Maximum Production from Given Inputs

Productive efficiency requires that every firm in the economy produce the maximum possible output from the resources it is using.

A few decades ago, union contracts required that typesetters set type for all newspaper advertisements even when the firms running the ads provided their own camera-ready copy. The “bogus” type of the typesetters would be destroyed, and the customer’s original copy was actually used. This practice is a perfect example of inefficiency within a firm. The newspaper could have used those typesetters to make a better newspaper that would sell to more people, but instead had to use them completely unproductively. This kind of practice has tended to disappear as the economy has been streamlined and made more efficient.

But we can still find examples. Virginia taxi drivers who take people to Washington, DC are not allowed to pick up passengers in Washington, and Washington taxi drivers are not allowed to pick up passengers in Virginia after dropping off passengers at the two Virginia airports. As a result, hundreds of taxis make return trips without passengers every day—wasting fuel and labor time that could have been used to provide valuable taxi service. Clearly, we are not producing the maximum amount of transportation services from the resources we are devoting to transportation.

Efficient Allocation of Inputs Among Firms. The third requirement for productive efficiency is that inputs be allocated among firms for the maximum possible production. More specifically,

Productive efficiency requires that resources be allocated among firms in such a way that the economy cannot increase the production of one good without decreasing the production of some other good.

This requirement for efficiency is the subtlest of the three in our list. Unemployed resources and inefficient practices *within* firms may be glaring, but the misallocation of inputs *among* firms is unlikely to be conspicuous.

Here is a simple example: Imagine that a warehouse and an auto repair shop are located next door to each other. The auto repair shop has no rest room, so its mechanics have to walk more than a block to use the public rest room several times a day, but it has an extra personal computer that it rarely uses. The warehouse has a rest room that is underused, but it has no computer for tracking inventory, and so its employees must spend hours figuring out the next day’s orders. One day, the owners of the two businesses get together and realize that if the auto shop traded its extra computer for the use of the warehouse’s rest room, then the warehouse could ship more orders each day and the auto shop could fix more cars. Since this would increase the output of both firms’ goods without decreasing the production of anything else, the economy was not productively efficient before the trade. On a *PPF*, the trade would be represented as a movement from a point inside the curve to a point *on* the curve.

Of course, this example was constructed to be easy, not realistic. But we can imagine more realistic cases where a rearrangement of inputs among firms would get us more of some good with no less of any other. Suppose we found that a hospi-

tal in Minneapolis had six technicians for every radiologist, so that the technicians were underused and the radiologists were stretched thin, while another hospital in St. Paul had only two technicians for every radiologist. Shifting radiologists to St. Paul and technicians to Minneapolis could increase the volume of X-rays produced in both hospitals.

Perfect Competition and Productive Efficiency. It is one thing to describe the conditions for achieving productive efficiency and another to understand how they come to be met in practice. Next, we will see how perfect competition achieves productive efficiency.

In Chapter 8, we analyzed perfectly competitive product markets, and in Chapters 11 and 13, we looked at perfectly competitive markets for labor and financial assets. In these cases, markets were competitive because they satisfied three conditions:

1. There are large numbers of buyers and sellers, each of whom buys or sells only a tiny fraction of the total quantity in the market.
2. Sellers offer a standardized product or resource.
3. Sellers can easily enter into or exit from the market.

One reason that economists devote so much time to analyzing the market structure of perfect competition is because of an important conclusion:

Perfectly competitive markets tend to be productively efficient.

That is, when markets are perfectly competitive, we cannot produce more of one thing without producing less of something else. Let's see why by examining how competitive markets help satisfy all three conditions for productive efficiency.

Profit Maximization and Full Employment of Resources. Profit-maximizing firms will try their best to make use of any unemployed resources. Think again about how many school buildings are left empty in the summer. This waste of resources would be unlikely to occur if the buildings were privately owned and the owners were free to seek the maximum profit in using them. The profit motive is a strong force for putting unused resources to work if it is not counteracted by regulation, government ownership, or similar forces. (Of course, this does not mean that regulation or government ownership is a bad thing; remember that productive efficiency—and even economic efficiency—is not the only goal of society.)

Perfectly competitive firms strive for maximum profit, but so do other kinds of firms. Thus,

An economy in which firms are free to seek the maximum profit—whether perfectly competitive or not—will tend to have full employment of resources.¹

Profit Maximization and Maximum Production with Given Inputs. Profit-maximizing firms will also strive to avoid inefficiencies in their internal operations—they produce the most they can from the inputs they use. Again, the profit motive is a strong inducement to this kind of efficiency. If more output can be obtained from the same inputs, the owner of a firm will earn more revenue without



Characterize the Market

¹ Periods of high unemployment during recessions are an important—but temporary—exception to this statement.

any increase in cost. Who could resist making that change, which generates profit that goes straight into the owner's pocket?

To see this another way, suppose that some firm did *not* satisfy the “maximum production” condition. That is, suppose it could find a way to produce the same output using fewer resources. Then it could have lower costs with the same total revenue. Since profit is total revenue minus total cost, this change in production would always increase profit. Thus, in a market economy in which all firms strive to maximize profit, each firm will produce the maximum possible output from the resources it uses:

An economy in which firms are free to seek the maximum profit—perfectly competitive or not—firms will tend to produce the maximum output possible from the inputs they use.

Perfect Competition and the Best Allocation of Inputs Among Firms. To see that a perfectly competitive economy is productively efficient, we have to check our last condition—that the allocation of inputs *among* firms is the best possible. That is, we must be sure that there is no way to move inputs from one firm to another that raises the output of one product without lowering the output of another product.

Consider the case of Anita and Bob, who are both onion growers. Each owns her or his own land (a fixed input) and hires labor (a variable input) during the growing season. On each farm, adding another worker will increase total output by the marginal product of labor (*MPL*) on that farm.

The efficient division of labor—the one that gives the highest level of output—occurs where the marginal products of labor at the two farms are equal. Or, more generally,

When labor is allocated efficiently between two firms producing the same kind of output, the marginal product of labor will be the same at both firms.

To see why, suppose the marginal products were not equal. In particular, suppose that on Anita's farm, $MPL^A = 2$ —that is, an additional worker produces 2 tons of onions—while on Bob's farm, $MPL^B = 3$. Then if we were to shift one worker from Anita's farm to Bob's, Anita would lose 2 tons of onions, while Bob would gain 3 tons. Total output of onions would rise by 1 ton per week. Thus, whenever marginal products of labor are different for any two firms, the economy is productively inefficient.

To see how this productive inefficiency implies economic inefficiency, we need only demonstrate that shifting one worker from Anita's farm to Bob's is a Pareto improvement. To do so, we'll suppose that onions sell for \$100 per ton and each farm worker earns \$240. We'll also assume that workers are indifferent between working in the two farms, as long as they are paid the same wage. Thus, only Anita and Bob are affected by the shift. Here's the scorecard for this Pareto improvement:

Action: One worker is shifted from Anita's farm to Bob's farm.

Bob	Gains revenue from 3 tons of onions:	\$300
	Pays wage of:	\$240
	Comes out ahead by:	\$ 60
<hr/>		
Anita	Gains from saved wage payments:	\$240
	Loses revenue from 2 tons of onions:	\$200
	Comes out ahead by:	\$ 40
<hr/>		

As labor is shifted from Anita's farm to Bob's, the marginal product of labor will change. Because of the law of diminishing marginal productivity, Bob's marginal product of labor will fall, and Anita's will rise. But as long as the two MPL s are not equal, further Pareto improvements are possible.

Now let's see how competition helps to bring the economy toward the efficient allocation of labor. Suppose that the two farmers hire workers in a competitive labor market and sell their onions in a competitive product market. Then each pays the wage, \$240 per week, that prevails in the market. Because the labor market is competitive, neither farmer can affect that wage. And because the onion market is competitive, both farmers sell at a price, \$100, that they can't affect. Recall from Chapter 11² that, to maximize profit, Anita will hire labor up to the point at which $P \times MPL^A = W$, which we can rearrange to $MPL^A = W/P$. Similarly, Bob will hire labor until $P \times MPL^B = W$, rearranged to $MPL^B = W/P$. Since MPL^A and MPL^B are both equal to W/P , they must be equal to each other, so we have $MPL^A = MPL^B$. In the process of maximizing profit, competitive firms end up with equal marginal products of labor. As a result, they allocate labor efficiently.

What is true of this type of labor will be true of any input—other types of labor, capital, land, oil, and anything else. In the end, we know that when input and product markets are perfectly competitive, it is impossible to transfer any input from one firm to another and have production increase. Thus, the allocation of inputs among competitive firms is productively efficient.

Productive Efficiency in Perspective. Productive efficiency, in and of itself, is good for the economy. Any society that wants to achieve a high standard of living will strive for productive efficiency; otherwise it will not make the best of its limited resources. A more specific way of stating this is the following:

Productive efficiency is necessary for economic efficiency.

How do we know this? Imagine an economy that is productively *inefficient*. From our definition of productive efficiency, we know that in this economy, there is an opportunity to produce more of some good—say, high-speed Internet connections—without producing less of anything else. As long as *some* member of society could be made better off by getting an additional high-speed connection, and as long as no other production would suffer—so no one would be harmed—then we *must* produce the additional connection in order to be economically efficient. You can see that moving toward productive efficiency by producing more high-speed hookups would also be a move toward *economic* efficiency as well. Unless we achieve productive efficiency, our citizens are not as well off as they can be, so we are not economically efficient either.

But productive efficiency *does not guarantee* economic efficiency. That is, a productively efficient economy could still be economically *inefficient*. How?

Imagine, for example, an economy with fully employed resources, in which each firm is producing the maximum possible output with the resources it is using, and resources are distributed among firms so as to get the maximum total output of goods and services. This economy would be productively efficient: It would be impossible to produce more of one good without producing less of some

² In Chapter 11, the condition for profit-maximizing employment is that $MRP = W$. But Chapter 11 also shows that, when a firm sells its output in a perfectly competitive market, MRP is the same as $P \times MPL$. Thus, we can rewrite the condition for profit-maximizing employment as $P \times MPL = W$.

other good. But suppose that most of the resources were being used to produce goods and services that no one wanted? (Homes built entirely of glass? Underwear woven from steel wool? Pasta with pinecones?) Or suppose this economy was producing goods that people did, in fact, desire, but they ended up going to the wrong households? (Rock fans get George Straight CDs, while Straight fans end up with Alanis Morissette CDs.) Even though such a society would be productively efficient, it would still be *wasting* something important: the opportunity to make people as well off as they can be, given the resources available *and given their preferences*. In our examples, we could make everyone better off if we changed the mix of goods produced to better suit people's tastes or changed the way any given collection of goods was distributed among the population. Could we really call such a society "efficient" when it is wasting the opportunities to make its citizens better off?

Productive efficiency is only part of the story of efficiency. A productively efficient society might be wasting opportunities to make its citizens better off because it is producing the "wrong" goods or because goods are not distributed to those who value them the most.

Economic efficiency requires not only productive efficiency, but also another kind of efficiency—*allocative efficiency*—which is the subject of the next section.

ALLOCATIVE EFFICIENCY

As we've just seen, productive efficiency does not *guarantee* economic efficiency. In addition to *producing* goods efficiently, economic efficiency requires that households get the *right* goods in the *right* amounts. That is, economic efficiency requires *allocative* efficiency:

An economy is allocatively efficient when there is no change in the quantity consumed of any good by any consumer that would be a Pareto improvement.

Allocative efficiency When there is no change in quantity consumed of any good by any consumer that would be a Pareto improvement.

To explore the concept of allocative efficiency, let's return to a familiar tool—supply and demand.

Another View of Supply and Demand Curves. Figure 2 shows Angela's demand curve for oranges. This demand curve shows us the quantity she demands at each price. For example, at \$0.28 per orange, Angela buys five oranges per week.

But we can also view this demand curve in a different way: It tells us how much each additional orange is *worth* to Angela. For example, suppose we want to know the value, to Angela, of the sixth orange. We know that when the price of oranges is \$0.28, Angela chooses *not* to buy the sixth orange. But if the price falls to \$0.27, she *does* buy the sixth orange. Therefore, the sixth orange must be worth \$0.27 to her.

We can conclude that the height of the demand curve at each quantity indicates the additional value or extra benefit a consumer would obtain by consuming that last orange. The *marginal benefit* of the sixth orange to Angela (at point A) is \$0.27. Similarly, the marginal benefit of the sixteenth orange (at point B) is \$0.17. This is why, in the figure, the demand curve has also been labeled as the *marginal benefit curve*.

What's true for Angela will also be true for every other consumer of oranges, so when we turn our attention to the *market* demand curve for oranges—like the one

ANGELA'S MARGINAL BENEFIT FROM CONSUMING ORANGES

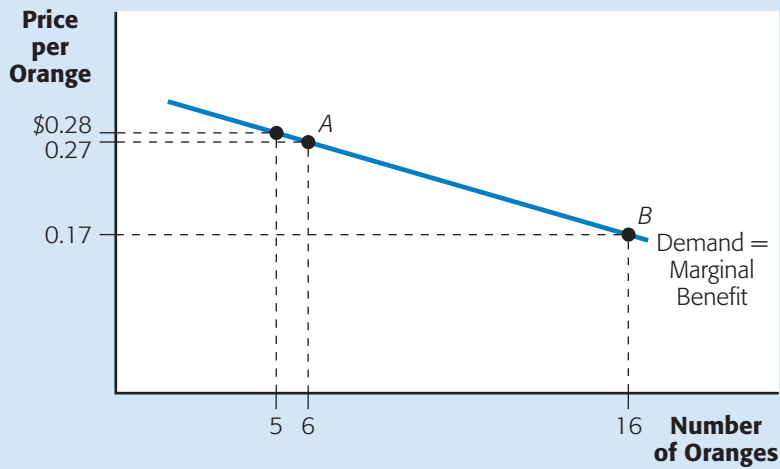


FIGURE 2

Angela's demand curve shows the marginal benefit she receives from each additional orange. At \$0.28 each, she consumes 5 oranges. If the price falls to \$0.27, she consumes 6 oranges. Her marginal benefit from that sixth orange must be \$0.27. In a similar way, the height of the demand curve at point *B* shows her marginal benefit from the sixteenth orange—\$0.17.

EFFICIENCY IN THE MARKET FOR ORANGES

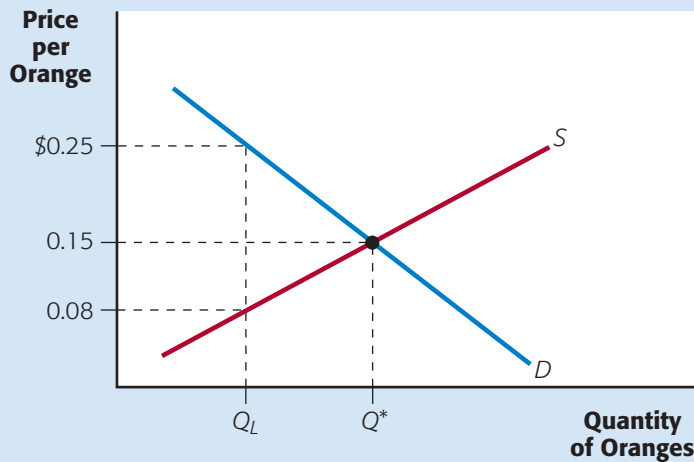


FIGURE 3

Quantity Q^* , where the demand and supply curves cross, is the economically efficient quantity. The marginal benefit to some consumer from the last orange consumed just equals the marginal cost of producing it—\$0.15. At a lower quantity, such as Q_L , the marginal benefit (\$0.25) exceeds the marginal cost (\$0.08). That is inefficient.

in Figure 3—the height of the curve tells us the value that *some* consumer will get from the last orange consumed. More generally,

the height of the market demand curve at any quantity shows us the marginal benefit—to someone—of the last unit of the good consumed.

Now that we've reinterpreted the market demand curve as the marginal benefit curve, let's take another look at its counterpart: the market supply curve. In Chapter 5, you learned that a competitive firm's supply curve is also its marginal cost curve. And in Chapter 7, you learned that the market supply curve tells us the marginal cost of producing an additional unit of output at *some* firm. That is,

the height of the market supply curve at any quantity measures the marginal cost—to some firm—of the last unit produced.

Find the Equilibrium



The Efficient Quantity of a Good. What is the efficient quantity of a good—that is, the quantity that takes advantage of all possible Pareto improvements in the market? It will be the quantity at which the supply and demand curves intersect—represented by Q^* in Figure 3. At this quantity, the marginal benefit to some consumer from the last orange consumed—\$0.15—just equals the marginal cost to some firm of picking, shipping, and selling that orange.

But why is this the economically efficient quantity? To see why, let's consider other levels of output and show that a Pareto improvement occurs as consumption moves closer to Q^* . At Q_L , a level below the efficient level, the demand (marginal benefit) curve lies above the supply (marginal cost) curve. From the marginal benefit curve, we know that some consumer would be willing to pay as much as \$0.25 for an additional orange. The marginal cost curve tells us that some firm would be willing to supply that orange for \$0.08. Both parties would be better off if the orange were produced and sold at any price between \$0.08 and \$0.25. For example, a Pareto improvement occurs if a consumer buys another orange for \$0.15, as you can see in the following scorecard:

Action: One more orange is produced and sold for \$0.15.

Some consumer	Gains benefits worth:	\$0.25
	Pays:	\$0.15
	Comes out ahead by:	\$0.10
Some producer	Gains revenue of:	\$0.15
	Cost:	\$0.08
	Comes out ahead by:	\$0.07

In a similar way, you should be able to show that a consumption level above Q^* is too high to be efficient. (Describe a Pareto improvement that involves producing one less orange.) The only level of consumption at which no Pareto improvement is possible is Q^* , which is the efficient level:

The efficient level of production of any good is where the demand, or marginal benefit, curve crosses the supply, or marginal cost, curve. At any other level of output, a Pareto improvement is possible by changing production.

Now we can pull together everything we've learned about supply and demand and efficiency:

In a perfectly competitive economy, the marginal cost of a good is given by the market supply curve, and the marginal benefit of the good is given by the market demand curve. Thus, the equilibrium quantity—where the supply and demand curves intersect—is also the efficient quantity—where marginal benefit and marginal cost are equal.

Let's consider this last statement carefully. It tells us that, if we leave producers and consumers alone to trade with each other as they wish, then—as long as the mar-

ket is perfectly competitive—the quantity bought and sold will automatically be the economically efficient quantity.

Now we can see why the market never produces goods such as pasta with pinecones. Suppose it costs \$5.00 to make a plate of this delicacy, but no consumer ever places a marginal value on it of more than \$0.20. The efficient amount to produce would be zero—it should not be produced. And that is just what the market

will do, since there is no price at which the supply and demand curves intersect. (Try drawing possible supply and demand curves for this good to prove this to yourself.)

How does the market achieve these remarkable results—not producing the wrong goods, and producing the right goods in the right quantities? The market's job is to establish the equilibrium price of the good—the price at which quantity demanded and quantity supplied are equal. Once that price is determined, each consumer adjusts consumption until his *marginal benefit just equals the price*. Each firm continues to adjust its production until its *marginal cost equals the price*. Thus, each consumer will end up consuming until his *marginal benefit is equal to the marginal cost* to some firm. Thus, in the market as a whole, the last unit produced will provide a marginal benefit to some consumer equal to its marginal cost to some firm.

Notice that this result comes about automatically. Consumers don't have to approach firms and ask them to produce the economically efficient quantity of each good. Rather, firms maximize profit and consumers maximize their well-being, and—as a consequence—the economy produces the efficient amount of the good. There are no remaining opportunities for Pareto improvement in the market.



Here is an important reminder: Don't confuse efficiency with fairness. Producing the quantity of a good where the demand and supply curves intersect will be *efficient*, but it may not be *fair*.

To see why, remember that the demand—or marginal benefit—curve tells us how much income some consumer would give up to buy another unit of a good. But this, in turn, depends on how much income the consumer *has*. A very poor person might want food very badly, but if she has no income, her desire would not register at all on the demand curve in Figure 2. Thus, in principle, an efficient level of food production could be one in which many people starve, and just a few—those with income—have food.

More generally, the market demand curve for any good will depend on the distribution of income and wealth in the society. If that distribution is regarded as unfair, then the quantities of goods produced and consumed will be unfair as well, even though they may be efficient.

ECONOMIC EFFICIENCY AND PERFECT COMPETITION: A SUMMARY

Figure 4 summarizes the four conditions we've discussed—three for productive efficiency and one more for allocative efficiency. At the bottom, the table notes the conditions under which the corresponding kind of efficiency will occur. Notice that all of the conditions are satisfied under perfect competition.

Perfectly competitive markets tend to be economically efficient—that is, both productively and allocatively efficient.

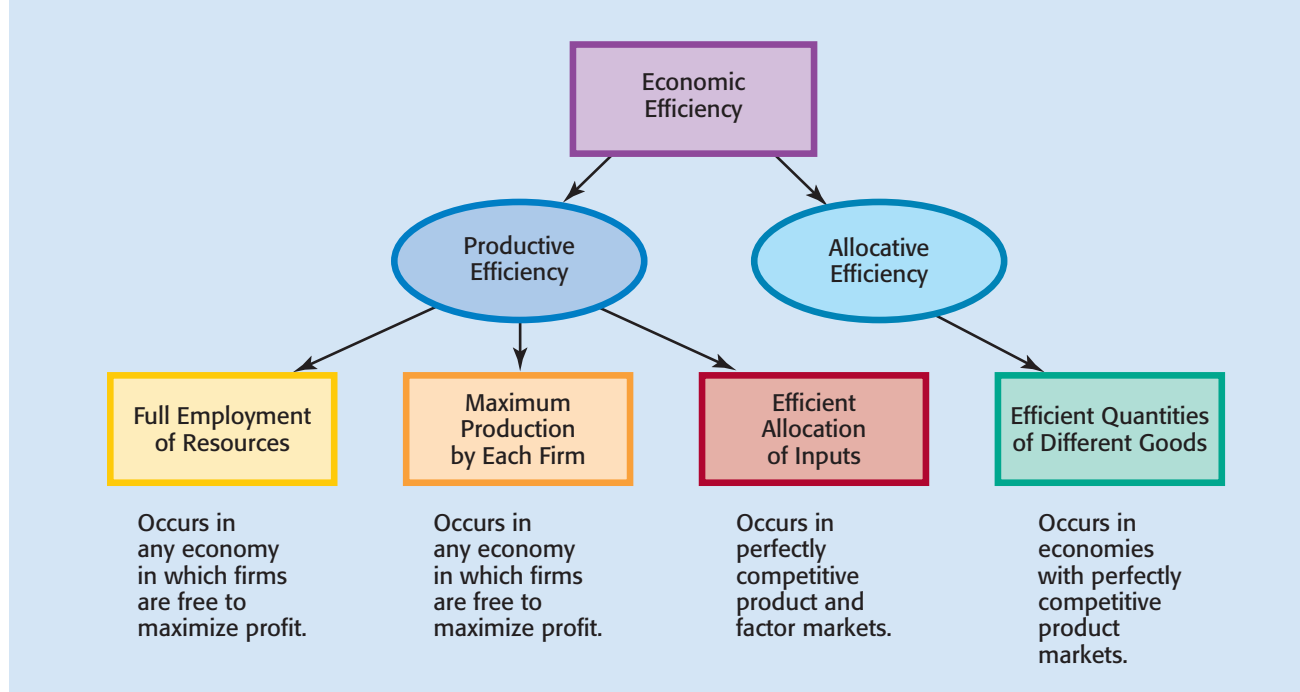
The notion that perfect competition—where many buyers and sellers each try to do the best for themselves—actually delivers an efficient economy is one of the most important ideas in economics. The great British economist of the eighteenth century, Adam Smith, coined the term *invisible hand* to describe the force that leads a competitive economy relentlessly and automatically toward economic efficiency:



Can smuggling contribute to economic efficiency? See Sheila Campbell's "Smuggling Smokes, Eh?" at http://www.dismal.com/thoughts/th_sc_013100.stm.

FIGURE 4

TYPES OF ECONOMIC EFFICIENCY AND CONDITIONS UNDER WHICH THEY OCCUR



[The individual] neither intends to promote the public interest, nor knows how much he is promoting it. . . he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was not part of his intention. Nor is it always the worse for the society that it was no part of it. By pursuing his own interest, he frequently promotes that of society more effectually than when he really intends to promote it. [Emphasis added.]

One implication of Smith's insight is that in many cases the results we get from the invisible hand of competition are better than we would get from the *visible* hand of regulation or government operation of the economy. But remember that the invisible hand works best in an economy in which markets are working well and where they are perfectly competitive. As we will see in the next section and also in the next chapter, when there is imperfect competition or when markets fail to function in other ways, then the invisible hand may not work. In those cases, government action may be needed to bring about economic efficiency.

THE INEFFICIENCY OF IMPERFECT COMPETITION

We've seen that perfect competition delivers the efficient quantities of goods to the consumer. What about other market structures? Here we will consider an example of the inefficiency of imperfect competition.

Let's consider the market for cornflakes shown in Figure 5. There is imperfect competition in this market when each producer of cornflakes—Kellogg, for example—faces a downward-sloping demand curve for its product, like the one in the figure. As you first learned in Chapter 7, when the demand curve facing the firm slopes downward, marginal revenue at each output level will be less than the price. This is

Characterize the Market



why, in the figure, the marginal revenue curve is drawn *below* the demand curve. Finally, you've learned that the firm will maximize profit by equating marginal revenue to marginal cost. In the figure, this occurs when the firm produces the output level q^* .

Now consider a crucial feature of this market:

In an imperfectly competitive market, the equilibrium price exceeds the firm's marginal cost of production.



Identify Goals and Constraints



Find the Equilibrium

In Figure 5, when the firm is maximizing profit at output level q^* , the price—and the marginal benefit to some consumer—is \$3.00 per box of cornflakes, while marginal cost is just \$1.00. The result is *economic inefficiency*—more specifically, an output level of cornflakes that is smaller than the efficient level. Why? Because when output is q^* , we can find a Pareto improvement. For example, suppose a consumer buys one more box of cornflakes and pays \$2.00. Here is a scorecard for that transaction, showing that it would yield a Pareto improvement:

Action: One more box of cornflakes is produced and sold.

Some consumer	Gains benefits worth:	\$3.00
	Pays:	\$2.00
	Comes out ahead by:	\$1.00
Cornflakes company	Gains revenue of:	\$2.00
	Marginal cost:	\$1.00
	Comes out ahead by:	\$1.00

The additional consumption is beneficial, because the marginal benefit to the consumer exceeds the marginal cost to the producer. If Kellogg does *not* produce an additional box, then the market for cornflakes is inefficient.

As you can see, in monopoly or imperfectly competitive markets, the profit-maximizing quantity for firms—and the equilibrium quantity in the market—is inefficient. It is possible to produce more of the good and make both producers and

THE INEFFICIENCY OF IMPERFECT COMPETITION

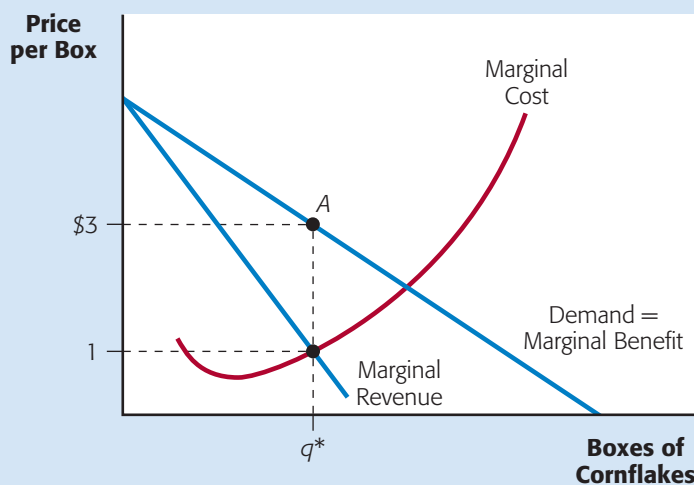


FIGURE 5

An imperfectly competitive firm, such as Kellogg, faces a downward-sloping demand curve and maximizes profit by producing q^* boxes of cornflakes. At that output, the benefit of another box to some consumer (\$3) exceeds the marginal cost of producing it (\$1). That is economically inefficient.

consumers better off—a Pareto improvement. However, this will not happen as long as firms behave as shown in Figure 5. Thus, from the point of view of efficiency,

monopoly and imperfectly competitive markets, in which firms charge a price greater than marginal cost, produce too little output at too high a price.

This conclusion applies to any market structure—monopolistic competition, monopoly, or oligopoly—in which we expect price to exceed marginal cost.

But wait, you might object, if both sides gain from the kind of transaction we've just described, what stops them from carrying it out? The answer can be found in Figure 5. The demand curve for cornflakes slopes downward. To sell additional boxes, Kellogg must charge a lower price—on *all* boxes. That is why marginal revenue is less than price. (If you need a refresher on this point, look back at Chapter 9 or Chapter 10.) If the firm could sell an additional box at \$2.00—and *keep charging \$3.00 on all other boxes—then it would, indeed, make the transaction.* As you learned in Chapter 9, a price-discriminating firm can do just that. But you also learned that not every firm can price discriminate, and even a price-discriminating firm may not be able to charge enough different prices to take advantage of *every* Pareto improvement.

In imperfect competition, it is the inability of firms to make separate side deals through price discrimination that prevents Pareto improvements from being carried out.

WHERE DO WE GO FROM HERE?

In this chapter, we've explored the concept of economic efficiency, focusing on three central issues: what economic efficiency means, what it requires, and why an economy with *well-functioning, perfectly competitive markets* tends to achieve it. But notice the italicized words in that last sentence. As you've just seen, when markets are *imperfectly* competitive, economic efficiency will not be achieved. In addition, markets may not function well for other reasons, besides imperfect competition. What are these other ways that markets can fail to perform? And what are we to do when an economy with such markets—left to itself—will not achieve economic efficiency? We take up these questions in the next chapter.

USING THE THEORY: THE COLLAPSE OF COMMUNISM

Using the THEORY



Economic efficiency can be a powerful tool to help us understand the economic changes that shook the world in the late 1980s and early 1990s. It can also help us understand the changes that are continuing to take place in China in the year 2000 and beyond, and may—in the future—take place in such holdouts as Cuba, Belarus, and North Korea as well. Simply put, the system that these nations had in place for decades—centrally planned socialism—was economically inefficient. True, the Soviet-inspired system of resource allocation by command and resource ownership by the state had its advantages. It enabled both the Soviet Union and China to become superpowers. But the system was plagued by so much inefficiency that—far from achieving its goals of beating living standards in market economies—it collapsed.

In this section, we'll look at the former Soviet economy through the lens of economic efficiency. Much of what we say, though, applies to other countries that based their economies—in whole or in part—on the Soviet model.

How did the Soviet economy actually work? Here were some of its key features:

- *Resource ownership*: With few exceptions, the state owned all factories, land, and capital equipment.
- *Resource allocation*: By command. Planners in Moscow set thousands of output targets and allocated the resources that the state felt were needed to achieve them. Since the output of one firm was the input of another, firms were heavily interdependent, and state planners took their output targets very seriously.
- *Prices of consumer goods*: Set by the state. In part, planners attempted to equate the quantity demanded for each good with the quantity supplied (the state's output target). But in practice, this usually proved impossible. Shortages and surpluses were regular occurrences.
- *Price of raw materials and resources*: Set by the state; often deviating widely above and below marginal cost.
- *Wages*: Set by the state, with heavy emphasis on equity.

Running the economy from Moscow never worked the way the authorities wanted it to. One reason was sheer complexity: Thousands of plants needed to be told what to produce, how to produce it, in what quantities, with how much of each input, and which firms and stores to deliver the output to. When one plant failed to meet its output target, shortages cascaded throughout the economy, paralyzing sectors dependent on that plant.

A famous example occurred in 1989, when the plant that produced locomotives for the entire Soviet economy reached only 75 percent of its target. When Moscow planners investigated, they found out why: The plant that made *engines* for locomotives had reached only 75 percent of *its* target. Why? Because the engine plant couldn't get enough deliveries of raw materials to meet *its* target. Why? Because the train system was running so poorly, in part because of a shortage of locomotives!

To prevent gridlock like this, Moscow did everything it could to ensure that output targets were satisfied, often regardless of cost. Managers faced serious consequences if they fell short of their output target—at some times in Soviet history, they faced long prison terms—but they faced little penalty for any other poor management practices, since the state was so heavily focused on the output targets.

As a result, the economy was productively inefficient. The system did not harness the powerful incentive of profit maximization, so plants did not necessarily produce the maximum output from a given amount of inputs. (See Figure 4.) If a manager ran out of inputs, he just asked the state planning agency to give him more. There was no way for the planning agency to verify whether the manager of, say, a radio factory truly *needed* more copper wire, or whether he was doing a poor job of managing, or even whether he was selling copper wire on the side.

Further, if a manager actually was able to produce his target output one year by using *less* than his allocated amounts of inputs, the planning agency had a habit of ratcheting his input allocation down the next year. For example, if an automobile factory was able to produce its target number of cars with 10 percent less steel than it was allocated, the next year it would likely find that its steel allocation was 10 percent lower. This gave the manager of the auto factory an incentive to *make sure* all allocated steel was used up, even if it meant wasting some steel to do so. Hotels would burn off heating fuel on hot days to make sure that they weren't caught with extra gas or coal at the end of the year.

In other words, compared to market economies, the Soviet economy did a poor job of solving the principal-agent problem discussed in Chapter 7. The backup systems that help in the U.S. economy if managers perform poorly—shareholder revolts and takeovers—had no counterparts in the Soviet system. Soviet managers—as agents of the planners—could continue to do whatever they wanted, as long as they satisfied their output targets and made it seem to the distant central planners that they were doing a good job.

In addition to not making the best use of inputs within firms, inputs were allocated *among* firms in an inefficient manner. For example, every plant manager had an incentive to *hoard* labor—to keep more on hand than was needed and to keep every worker that was allocated to the firm doing *something*—even if just standing around. Since firms had no private owners, there was no one to complain about the extra cost of the unneeded labor. But having extra workers made it easier to take care of emergency production—such as when the state suddenly asked the firm to increase its production because some other firm needed more inputs. As a result, some plants had excess labor, while others suffered severe shortages. Shifting workers from one firm to another could have increased the production of some goods, without decreasing the production of any other. But such shifts rarely took place, because managers with excess labor had no incentive to give it up to some other firm.

Finally, aside from its productive inefficiency, the Soviet economy also failed to achieve allocative efficiency—to provide efficient amounts of different goods, taking individuals' preferences into account. While households were free to buy whatever goods they wished with their state-determined incomes, many goods were unavailable. When anything went wrong in the plan—say, the electricity industry fell short of its monthly output target—it was always the consumer who suffered. The state would ensure that all *industries* that needed electricity got it. If those industries didn't get it, the shortage could threaten the entire structure of the plan. So instead, the state would simply make less available to households.

The same occurred with other products that served as both consumer goods and inputs for other firms—pencils, paper, wood, gasoline, cooking oil, sugar, and more. In a market economy, if the supply of some good decreases, its price will rise, and consumers will economize on its use, trying to find substitutes instead. But in the Soviet economy, all prices were set by the state. A decrease in supply simply meant a shortage. Most Soviets carried collapsible shopping bags all of the time just in case something to buy became available.

While the Soviet system suffered shortages of many consumer goods, it also had surpluses of others, especially shoes, shirts, suits, and dresses that no one wanted because they were shabby or out of style. The central planners had enough trouble coming up with a *consistent* plan—one where some firms produced enough inputs to enable other firms to produce enough outputs. They spent little time trying to make their plan coincide with consumer preferences.

In sum, the Soviets tried to rely on a *visible* hand, rather than relying on the automatic mechanisms of the *invisible* hand. In the end, this proved too daunting a task, and the system was extremely inefficient. Many important Pareto improvements that *would* have taken place in a market economy did not—or could not—take place in the Soviet economy. Since resources were scarce (as in *any* country), foregoing so many opportunities to make people better off led to a much lower standard of living than the Soviet Union should have had. The nation was rich in natural resources and was only about 20 percent behind the United States in capital per worker and in education. But production per worker in the USSR was only about a third of the level in the United States.

The standard of living was not only low relative to the West, but—beginning in the 1980s—it began to *grow* much more slowly than the West's as well. It became clear to Soviet leaders that their country not only was behind the West, but it was falling farther and farther behind with each passing year. As fax and copying machines, videotapes, and other modern methods of communication made it possible to learn about life in other, better organized countries, Soviet citizens became more and more disaffected and cynical about their own system.

Now Russia and the other countries that made up the Soviet Union are trying to convert to market economies. This task, as well, is proving difficult. One of the reasons for the difficulty was discussed in Chapter 6's "Using the Theory" section: Under the Soviet system, it made sense to build huge factories that could produce enough to satisfy the entire country's need for particular products. But now, with the transition to a market system, the owners of these plants are monopolists, charging inefficiently high prices.

Moreover, in market economies, the invisible hand operates within an infrastructure of laws and institutions provided by the government, which you will study in the next chapter. While many Eastern European nations have had great success in building this infrastructure, the nations of the former Soviet Union—including Russia—are still struggling. It is a sad fact that 10 years after the collapse of the Soviet Union, output in Russia was still below the already low level it had attained under central planning.

S U M M A R Y

A market or an economy is economically efficient when there is no way to reallocate production or consumption in a way that makes at least one person better off without making anyone else worse off. Behind that definition are several specific conditions that must be satisfied.

Productive efficiency occurs when it is impossible to produce more of one good without producing less of another. It requires that all productive resources be fully employed, that each firm produce the maximum output possible from the resources it uses, and that resources be allocated among firms to produce the maximum possible output. These three conditions are satisfied in perfectly competitive output and input markets.

Productive efficiency is necessary for, but does not guarantee, economic efficiency. Allocative efficiency is also a necessary condition for economic efficiency. Allocative efficiency is achieved when there are no changes in the quantity consumed of any good that would yield a Pareto improvement. This condition is automatically satisfied in perfectly competitive markets where the marginal benefit of the last unit consumed equals the marginal cost of producing it. Imperfectly competitive markets are not efficient because price exceeds firms' marginal costs of production.

K E Y T E R M S

Pareto improvement

economic efficiency

productive efficiency

allocative efficiency

R E V I E W Q U E S T I O N S

- Briefly define productive, allocative, and economic efficiency. What is the relationship between productive efficiency and economic efficiency? What is the relationship between allocative efficiency and economic efficiency?
- Explain why imperfect competition leads to inefficient outcomes.
- Briefly describe each of the following characteristics of efficiency and identify the conditions under which they will occur:
 - Full employment of inputs
 - Maximum production within each firm
 - Efficient allocation of inputs among firms
 - Efficient levels of production and consumption of different goods

4. Which of the following actions would be a Pareto improvement? Which *could* become a Pareto improvement if the right side payment were included?
 - a. You buy a Coke at the airport restaurant, where it costs \$2.50.
 - b. You and a friend go to a movie and compromise on which one to see.
 - c. An acquaintance who values your tennis racket more than you do borrows it and never returns it.
5. Give an example of a situation that is productively efficient but not economically efficient.

P R O B L E M S A N D E X E R C I S E S

1. In each of the following situations identify a Pareto improvement that is unexploited. Explain what can be done to permit the Pareto improvement to be realized. In each case, would a side payment be involved? Why might a government want to prevent such a Pareto improvement?
 - a. You are a low-income individual who receives food stamps from the government. Food stamps cannot be used to purchase nonfood items (e.g., paper towels). You wish to buy some paper towels.
 - b. In some cities, the government limits the rent that can be charged for apartments. You wish to rent a rent-controlled apartment. The controlled rent is below the equilibrium rent, and many other people also desire to rent this apartment. You have an appointment with the superintendent of the building to see the apartment.
 2. There are 30 students in an economics class. Each student likes doughnuts—all types, and the more the better. But they differ in their preferences. Half of the students prefer chocolate doughnuts to plain. The other half of the students prefer plain to chocolate. The instructor wishes to give away all the doughnuts he has. Explain whether the actions in parts (a) through (c) result in a situation that is economically efficient.
 - a. The instructor brings in 60 doughnuts (all plain) and gives them to a single student; no other student receives any doughnuts.
 - b. The instructor brings in 60 doughnuts (all plain) and gives two to each student.
 - c. The instructor brings in 60 doughnuts (half plain, half chocolate) and gives two (one of each kind) to each student.
- For any allocation you identify as inefficient, describe a Pareto-improving trade.
3. Look back at Figure 3 (p. 419). Suppose the government imposes price controls in the market for oranges by setting a minimum price of \$0.25 per orange. What would the new price and quantity be? Is that result efficient? If not, describe the nature of the inefficiency by identifying a Pareto improvement that is not being exploited.
 4. Chapter 9 discussed perfect price discrimination. Recall that a perfectly price-discriminating monopolist produces up to the point where its marginal cost curve intersects the market demand curve. Is that level of output efficient? Explain your answer.
 5. Figure 5 shows an inefficient level of output produced by a monopolistically competitive supplier of cornflakes. Suppose the government imposed a tax of \$2 per box of cornflakes. Would the tax affect the level of output produced by that firm? Would the result still be inefficient?

C H A L L E N G E Q U E S T I O N

A monopoly supplier of electricity faces a demand curve given by $P = 15 - Q$ where P is price in cents per kilowatt-hour of electricity and Q is thousands of kilowatt-hours produced and sold. The marginal revenue (MR) curve is $MR = 15 - 2Q$, and the marginal cost of producing a kilowatt-hour of electricity is constant at $MC = 5$ (i.e., \$0.05 per kilowatt-hour).

- a. What are the equilibrium price and quantity?
- b. The city government wishes to negotiate a special price at which an additional 2,000 kilowatt-hours of

electricity will be sold to low-income households. The special price will have no effect on the price charged to existing customers. What is the maximum price per kilowatt-hour the utility can charge and still expect to sell the extra electricity? What is the minimum price it would be willing to accept?

- c. Would moving to this two-tiered pricing system be a Pareto improvement? Explain why or why not. Develop a scorecard to illustrate your conclusion.

GOVERNMENT'S ROLE IN ECONOMIC EFFICIENCY

The U.S. economy relies heavily on markets. Yet even in the United States, market activity is supported by government in two crucial ways. First, the government provides the infrastructure that permits markets to function. Part of the infrastructure is physical—roads, bridges, airports, waterways, and buildings. Equally important is the market system's *institutional infrastructure*—laws, courts, and regulatory agencies. Although maintaining the institutional infrastructure uses only a small fraction of the nation's resources, the market economy would collapse without it.

The second way government supports market activity is by stepping in when markets are not working properly—when they leave Pareto improvements unexploited and therefore fail to achieve economic efficiency. The government's tools for making markets more efficient include regulation, antitrust law, and taxation. In this chapter, you will see how these tools work.

This is not the first time we've discussed government involvement in markets. But our earlier discussions focused on situations in which government *interfered* with the workings of a market economy. These situations included price ceilings and price floors, as well as government-created barriers to entry in product and labor markets. In many cases, the result was problems for the economy that could have been avoided by better policies. In this chapter, our focus is entirely different: We will look at how government *contributes* to economic efficiency by helping us achieve Pareto improvements that would not otherwise occur.

We begin by examining the institutional infrastructure of a market economy and then—in more detail—two important parts of that infrastructure: the legal and regulatory systems. Then we'll turn our attention to markets that fail to work efficiently and what the government can do about them.

THE INSTITUTIONAL INFRASTRUCTURE OF A MARKET ECONOMY

Americans take their institutional infrastructure almost completely for granted. The best way to appreciate the infrastructure of the United States is to visit another country that has a poor one. In many countries, the police are more likely to *steal*

CHAPTER OUTLINE

The Institutional Infrastructure of a Market Economy

- The Legal System
- Regulation
- Law and Regulation in Perspective
- Taxation

Market Failures

- Monopoly and Imperfect Competition
- Externalities
- Public Goods

Efficiency and Government in Perspective

Using the Theory: Case Studies of Antitrust and Regulation

- Breaking Up a Monopoly: Alcoa
- Regulation and Deregulation:
 - The Airlines
- Preserving Competition:
 - Soft Drinks
- An Ongoing Challenge:
 - Mighty Microsoft

from citizens than to protect them from thievery. In some nations, the people have no effective rights to their own property—somebody can start building a shack on their land, and the government won't stop him. If a person is injured by a drunk driver, there may be no system for compensating her or punishing the driver. Many nations suffer from powerful mafias that extort protection money by threatening to shut down businesses or physically harm their owners.

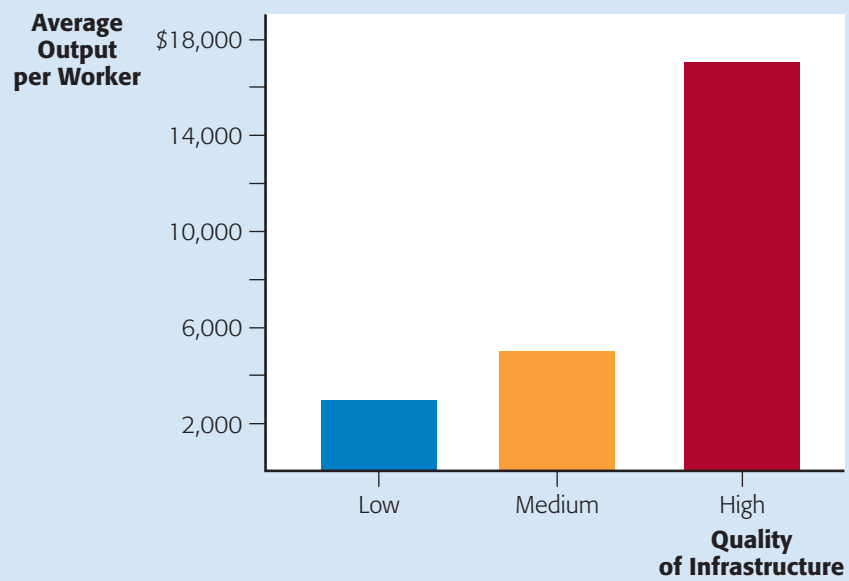
All too many nations suffer from problems such as these. Figure 1 illustrates one important result: When countries are divided into three groups, according to the quality of their institutional infrastructure, there is a strong relation between infrastructure and output per worker. The countries on the left—the ones with the lowest-quality infrastructures—were able to produce only about \$3,000 in output per worker-year in 1988. These are the nations where property rights are weak, contracts are not enforced, and the government is more often predator than protector of economic activity. In the middle of the figure are countries with medium-quality infrastructures, and these countries average about \$5,500 in output per worker per year. On the right are the best-organized countries, averaging \$17,000 in output per worker, and those countries with the very best infrastructures—such as the United States—achieved levels of output more than double that average.

The U.S. type of infrastructure is successful in supporting a thriving market economy, but it is not the only kind of infrastructure that works. Japan's market economy generates about 80 percent as much output per worker as does the United States. But Japan relies much less than the United States on the legal system and much more on relationships within networks of firms. In both countries, however, the government plays a major role in providing and supporting the institutional infrastructure that makes markets run more smoothly.

FIGURE 1

Countries with low-quality infrastructures produced an average of only \$3,000 per worker per year in 1988. These countries tend to have corrupt governments, poor enforcement of contracts, and weak property rights. Countries with higher-quality infrastructures, including the United States, produced an average of \$17,000 per worker per year.

GOVERNMENT INFRASTRUCTURE AND OUTPUT PER WORKER



THE LEGAL SYSTEM

The backbone of a market economy's institutional infrastructure is the legal system. Of course, the legal system is also important for noneconomic reasons. The law protects us from physical and emotional harm, and guarantees us freedom of speech and other vital civil liberties. Here, we will focus on the purely economic role of the legal system—that is, on the ways that it supports markets and helps us achieve economic efficiency. We'll look at five very broad categories: criminal law, property law, contract law, tort law, and antitrust law.

Criminal Law. While criminal law has important moral and ethical dimensions, its central *economic* function is to limit exchanges to voluntary ones. Since both parties agree to a voluntary exchange, they must each benefit from it. Therefore, as long as no third party is harmed, such an exchange will always be a Pareto improvement. But an involuntary exchange—robbery, for example—always harms one side.

By making most involuntary exchanges illegal, criminal law helps to channel our energies into exchanges and productive activities that benefit all parties involved. In this way, criminal law contributes to economic efficiency.

Of course, to be effective, it is not enough to merely *define* which activities are harmful; the criminal code must also be enforced, with penalties serious enough and certain enough to dissuade people from committing harmful crimes. In some cases, it has proven much easier to draft a criminal code than to provide for enforcement.

Russia, for example, enacted a sophisticated new criminal code in the mid-1990s, but has been unable to enforce it, due to massive corruption in local governments and police forces. As a result, a disproportionate number of Russian citizens pursue activities that harm others, such as running protection rackets that victimize small businesses, or eliminating business competitors through threats and even assassinations. The prevalence of economic crime is one of the main reasons why Russia—in spite of transforming itself from a centrally planned to a market economy—remained mired in economic inefficiency as it entered the new millennium.

Property Law. Property law gives people precisely defined, enforceable rights over the things they own. Without property law, you would spend a good part of your time dealing with people who claimed to own your house, your farm, or your factory. In the United States and other advanced countries, highly secure systems keep track of who owns land, cars, shares of stock, airplanes, patents, and other important pieces of property. Disputes about property ownership are rare because the system works so well.

When property rights are poorly defined, much time and energy are wasted in disputes about ownership, and people spend time trying to capture resources from others, time that could have been spent producing valuable goods and services. As a result,

countries with poorly defined property rights do not produce as much output from their resources as they could with better-defined property rights. Greater output could make some people better off without harming anyone—a Pareto improvement. Thus, countries with poorly defined property rights are economically inefficient.

Contract Law. In 1991, a lawyer named John Mackall had a great idea—to start a mail-order company to sell batteries for laptop computers. He had the money to start the business, but not the time. So he made a deal with a (very lucky) business school student, Ken Hawk, to start 1-800-BATTERIES.

As part of their deal, Mackall and Hawk signed a contract that gave Hawk a 75 percent interest in the company. Mackall invested \$50,000 and received the remaining 25 percent of the company. It turned out that the company was a success: By 1998, it has become the world's largest supplier of batteries and accessories for laptops and cell phones, with annual sales of about \$30 million.

But what guaranteed that Hawk would give Mackall his 25 percent share of the profits? In countries in which contract law is less well defined or less strictly enforced, somebody in Mackall's position would worry that he would not be able to collect his share later. In the United States, that worry would not arise, because contracts can be enforced.

A contract is a mutual promise. Often, as in the example of the battery business, one person does something first (Mackall provided his idea and \$50,000 in cash), and the other person promises to do something later (Hawk promised to run the business and pay Mackall 25 percent of the profit). As long as no third parties are harmed, the exchange that occurs under a contract is always a Pareto improvement—it's a voluntary deal that won't happen unless both sides are made better off.

Contracts play a special role in a market economy. Without them, the only Pareto improvements that could take place would be those involving simultaneous exchange. But contracts enable us to make exchanges in which one person goes first. That person has to be able to rely on the other person to make good on the promise later. Without this assurance, whoever goes first would not be willing to make the deal in the first place. Thus, contracts make it possible to form new companies and to hire the services of experts who specialize in such things as auto repair, plumbing, roof repair, dentistry, and legal services—all cases in which someone goes first:

Contracts enable us to make exchanges that take place over time and in which one person must act first. In this way, contracts help society enjoy the full benefits of specialization and exchange.

It is important to note that legal enforcement of contracts is not the only force that makes people keep promises. First, parents, religious organizations, and schools teach people that keeping promises is a moral obligation. Second, a reputation for failing to keep promises would be harmful to a business or a person. The Internet retailer *barnesandnoble.com* could not stay in business if the company developed a reputation for taking people's money, but only *sometimes* delivering the books that they ordered.

Still, while socialization and concern over reputation are important, contracts and the infrastructure for enforcing them play a vital role in making the economy more efficient. For example, we know that despite the efforts of parents, religious leaders, and schools, there are enough would-be cheaters to create problems. Contract law provides an effective way to deal with them. Moreover, contract law makes it easier for new businesses to enter an industry and grow. After all, it takes time to develop a reputation for making good on promises. But because of contract law, people are more willing to take a chance with a new business, since they know that they have the law behind them if the new business reneges on a deal.

Tort Law. Contract law deals with people or businesses that are economically involved with each other, such as suppliers and their customers, or partners in a busi-

ness deal. Tort law, on the other hand, deals with interactions among strangers or people not linked by contracts.

More specifically, a **tort** is a wrongful act—such as manufacturing an unsafe product—that causes harm to someone, and for which the injured person can seek remedy in court. Tort *law* defines the types of harm for which someone can seek legal remedy, and what sorts of compensation the injured person can expect.

When people and businesses are held responsible for injuries they cause, they act more carefully. Tort law in the United States provides incentives for drivers to drive carefully, for doctors to examine their patients more completely, and for manufacturers of products such as power mowers to control hazards through proper design.

Tort law also protects against *fraud*, in which a seller of something—a product, a business, shares of stock—lies to the buyer in order to make the sale. In some countries, fraud is such a pervasive problem that you can't trust the claims made by the sellers of anything. In the United States, by contrast, sellers are extraordinarily careful about their claims. You would be extremely surprised if you bought a down vest and found later that it was stuffed with polyester. Similarly, the information that is released by a company when it sells stock to the public is scrutinized minutely by the company's lawyers, to be absolutely sure of its accuracy. If a firm says it owns 17 million barrels of proven oil reserves or movie theaters with 125 screens, you can be almost completely confident that it does. The penalties for lying about a subject like that are severe.

Antitrust Law. Antitrust law is designed to prevent businesses from making agreements or engaging in other behavior that limits competition and harms consumers. More specifically, antitrust law operates in three areas:

1. *Agreements among competitors.* U.S. antitrust law—expressed in Section 1 of the Sherman Act—prohibits “contracts, combinations, or conspiracies” among competing firms that would harm consumers by raising prices. The most flagrant agreements prohibited by this law are those that directly fix prices. But the law also prohibits agreements that raise prices *indirectly*, by limiting competition among sellers. An agreement by firms to allocate markets among them—so that one seller serves one group of customers exclusively, while other sellers are assigned their own groups of exclusive customers—may violate the law. For example, in the mid-1990s, the only two important sellers of review courses for the bar exam taken by prospective lawyers divided up their territory to avoid competition. The courts outlawed their agreement because it reduced competition.

2. *Monopolization.* Section 2 of the Sherman Act makes it illegal to monopolize or attempt to monopolize a market. As the law is now interpreted, it is illegal for one seller to harm a rival by interfering with its operations or hobbling the rival in certain ways. For example, it is illegal for a company to spread false information about a rival's product as part of an attempt to drive that rival out of the market. But the law does not prohibit monopoly or harm to competitors. Rather, it prohibits *certain steps* to acquire or maintain a monopoly or to harm competitors. A firm that harms its rivals by selling a better product, thus taking business away from them, is not in violation of the law.

3. *Mergers.* In a merger, two firms combine to form one new firm. The result is to increase the danger of higher prices from oligopoly or monopoly. For example, if the largest firm in a market has a 40 percent share of total sales, and the second-largest has a 30 percent share, we can expect that the rivalry between them will benefit consumers. But if they merge to form a single firm with a 70 percent share, the rivalry would disappear, and prices would rise. Mergers of this type are often

Tort A wrongful act that harms someone.

blocked by the U.S. government based on Section 7 of the Clayton Act. We'll look at mergers more carefully later in this chapter.

REGULATION

Regulation is another important part of the institutional infrastructure that supports a market economy. Under regulation, a government agency—such as the Food and Drug Administration (FDA), the Environmental Protection Agency (EPA), or a state public utilities commission—has the power to direct businesses to take specific actions. The EPA has detailed control over what substances a business can release into the atmosphere or into the water. Public utilities commissions set the prices for electricity, gas, and telephone service. Often, regulators must approve business actions before they are undertaken, as in the case of the FDA's approval of new drugs. In addition to protecting public safety and health, regulation is also used to help markets function more efficiently, as we will discuss more thoroughly later in the chapter.

Regulation differs from the use of legal procedures in a fundamental way: Regulators reach deep into the operations of businesses to tell them what to do, while legal procedures typically result in fines or other penalties if businesses do something wrong. To help see the distinction, consider the different ways in which regional and long-distance telephone companies are treated. Because they are regulated, regional telephone companies (such as Bell South or Cincinnati Bell) are *told* what price to charge. Long-distance phone companies, by contrast, are largely unregulated, so they can charge whatever price they wish. But if long-distance companies are caught breaking the law in setting prices (such as, by entering into illegal agreements to restrict competition), they will have to pay fines, and their managers may even have to go to jail.

LAW AND REGULATION IN PERSPECTIVE

The invisible hand of the market system cannot operate on its own. The legal system, along with our regulatory agencies, creates an environment in which the invisible hand can do its job. Almost every Pareto improvement that we can think of relies on the legal and regulatory infrastructure. Recall the last time you bought a meal in a restaurant. If you paid cash, the criminal law against counterfeiting enabled the restaurant to more readily accept your paper currency. If you paid by credit card, contract law assured the restaurant that it would eventually be paid by the credit card company. The restaurant itself couldn't function without contracts with its suppliers, landlord, and employees. You could be reasonably confident that the food was not contaminated, in part because of inspections by local regulatory agencies and also because tort law provides legal disincentives for harmful products.

But what about cases where law and regulation don't seem to be working perfectly? After all, we still have crimes against people and property. Unsafe products like poorly designed automobiles or tainted frozen dinners *are* produced and only sometimes recalled before someone is harmed. Businesses *do* fix prices and are only sometimes caught. Do these and countless other examples mean that our institutional infrastructure is failing us?

Yes . . . and no. While instances like these are never welcome, our society has *chosen* not to eliminate them entirely. We could, if we wanted to, eliminate all crime, all unsafe products, and all other detriments to economic life by enacting more stringent laws and regulations and enforcing them to the hilt. But doing so would require even larger expenditures on legal and regulatory enforcement than

we currently make. In deciding whether to make these expenditures, we must balance the benefits—safer products, reduced crime, and the like—against the costs.

For example, in part because of our strong tort law, the United States is one of the safest countries of the world, and is growing safer. By 1990, the death rate for children under age 12 had fallen to less than half its 1960 level. Adult on-the-job death rates have fallen dramatically as well. But even the United States has chosen not to *completely eliminate* safety hazards: Each year, 20 people out of every 100,000 die from accidents. Why do we accept this? Because the complete (or almost complete) elimination of fatal accidents would require too many of our resources to be diverted from other uses. Most of us would think it is simply not worth it. For example, to eliminate all preventable fatal accidents, we would have to require that every passenger aircraft be inspected dozens—perhaps even hundreds—of times after each flight; that drivers enroll in a refresher course each year, perhaps each month, updating and reinforcing their driving skills and safety consciousness; that all restaurants inspect every meal for *E. coli* contamination before serving it; and that all floors and shoes be manufactured out of special materials to make slipping impossible. Moreover, all of these requirements would have to be strictly enforced, requiring more police and inspectors to catch violators and more courts and jails to prosecute and penalize them. In such a world, our standard of living would plummet, and we'd all agree that we'd be better off taking on some additional risk of accidents in order to free up resources for increased production.

A legal and regulatory system that ensured the complete elimination of crime, unsafe products, and other unwelcome activities would be less efficient than a system that tolerated some amount of these activities. An efficient infrastructure must consider the costs, as well as the benefits, of achieving our legal and regulatory goals.

TAXATION

The legal and regulatory infrastructure provided by government is not free. It takes *resources* to provide these institutions—the labor of police officers, judges, and inspectors; the capital of police vehicles, court buildings, and computers; and the land on which government buildings sit. In a market economy, the government does not commandeer these resources; it *buys* them in the marketplace, just as would anyone in the private sector.

But since the government rarely *sells* its services to the public, it needs a source of funding for all of its purchases. Enter *taxes*.

The main types of taxes in the United States are:

- excise taxes on particular products, such as gasoline
- income taxes paid by individuals and corporations
- payroll taxes, used to finance Social Security
- property taxes
- sales taxes

Taxes have an important effect on the efficiency of the economy. On the one hand, by providing the funds for the legal and regulatory infrastructure of the economy, taxes help to *increase* efficiency. Moreover, as you'll see, specific taxes can sometimes help to bring about economic efficiency in a market.

But taxes can—and do—create *inefficiencies* as well.

Taxes increase efficiency by providing the funds for the social infrastructure of a market economy. But they can make specific markets more or less efficient, depending on the nature of the tax and the initial conditions in the market.

In the rest of this section, we'll look closely at how two types of taxes—excise taxes and income taxes—affect economic efficiency.

What Happens When
Things Change?



Excise Taxes. In Chapter 4, you learned about excise taxes, which are taxes on specific goods or services. You saw in that chapter how an excise tax affected the market for airline tickets: an increase in the equilibrium price and a decrease in the equilibrium quantity of tickets sold. In our discussion in Chapter 4, we described *what* happens after an excise tax is imposed, but we did not *assess* excise taxes in terms of economic efficiency. We'll do that now.

Figure 2 looks at the market for international air travel, the same market analyzed earlier, in Figure 4 of Chapter 4 (p. 87). We assume that D is the public's demand curve for airline tickets and that there is a \$100-per-ticket excise tax imposed in this market. Notice that there are two supply curves in the figure. S is the familiar market supply curve that tells us the price that airlines must receive in order to supply each quantity of tickets per day, *with no excise tax*. But with a \$100 excise tax, an airline would need to receive \$100 more in order to get the same amount per ticket as before the tax. Thus, the price needed to induce airlines to supply each number of tickets will be \$100 greater than before. This is why the supply curve S' lies \$100 above the original supply curve S .

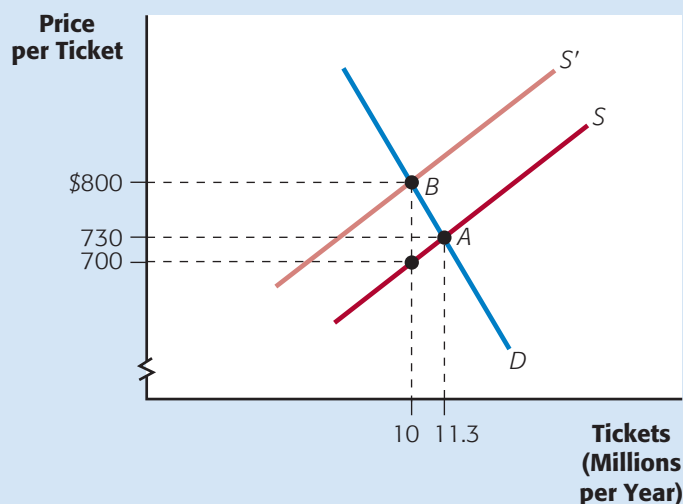
With no excise tax, the market would be in equilibrium at point A , with 11.3 million tickets sold and a price of \$730. But with the \$100 excise tax, the equilibrium moves to point B , with 10 million tickets sold. Travelers pay \$800 per ticket. Of that, the airline receives \$700 per ticket, and the government gets the difference—\$100.

Now we can see how an excise tax can create inefficiency in a market. Since we are using supply and demand curves, we have assumed that the airline industry is

FIGURE 2

EFFECT OF AN EXCISE TAX

At point A , the market is in equilibrium with 11.3 million tickets sold and a price of \$730 per ticket. If the government now imposes an excise tax of \$100 per ticket, the supply curve shifts up to S' . Equilibrium moves to point B where travelers pay \$800 per ticket and the airline keeps \$700. This is inefficient because some travelers could benefit from additional tickets by more than it would cost the airlines to provide them. Selling more tickets would enable the airlines, passengers, and the government to gain additional benefits.



competitive, so that the supply curve S is also the marginal cost curve. Thus, at 10 million tickets, the cost to some airline to provide one more international ticket would be \$700. But, according to Chapter 14's interpretation of the demand curve as the marginal benefit curve, some traveler would gain a benefit of \$800 from that ticket. So, at point B , a traveler could get a benefit of \$800 for an additional trip that would cost only \$700 to produce. But that ticket is *not* being sold with the excise tax in place. There is room for a Pareto improvement.

The following is just one example—among many possible examples—of a Pareto improvement: The airline could charge, say, \$750 for one more trip, including \$10 in excise tax. The government would receive an additional \$10 in tax revenue for the trip, and the airline would keep \$740. The following scorecard shows how each of the three parties would be affected:

Action: An airline sells an additional ticket for \$750, with \$10 in excise tax.

Traveler	Gains benefits worth:	\$800
	Pays:	\$750
	Comes out ahead by:	\$ 50
Airline	Receives:	\$750
	Pays tax of:	\$ 10
	Marginal cost:	\$700
	Comes out ahead by:	\$ 40
Government	Receives:	\$ 10
	Comes out ahead by:	\$ 10

As you can see, in our example, selling the additional ticket would be a Pareto improvement: No one is harmed, and the traveler, the airline, and the government all gain. But this Pareto improvement would require the government to accept less than its usual \$100 in taxes from the extra trip. Ordinarily, governments won't cut such deals. If they did, they would be pressured to extend the tax reduction to all travelers, and government tax policy would unravel. But as long as the government insists on getting its full \$100, there is no room for a Pareto improvement between the traveler and the airline.

Raising tax revenue with an excise tax on a good has a distorting effect similar to imperfect competition. It raises the price of the good and causes people to consume too little of it—less than the efficient quantity. In the presence of an excise tax, some mutually beneficial increases in production and consumption will not take place:

Raising general tax revenue with excise taxes is inefficient, since it creates a situation in which the marginal benefit for a consumer is greater than the marginal cost to some producer. Hence, too little of the taxed goods will be produced and consumed.

So far, we've considered an excise tax used to raise general tax revenue. But now let's consider another reason for an excise tax: as a charge for a specific government service. In particular, suppose our excise tax is used to pay for air traffic controllers, airport repair, and other goods and services that benefit travelers when they fly. Suppose, too, that each ticket sold costs the government \$100 in additional services. In this case, when the government charges an excise tax of \$100, and the market reaches equilibrium at point B in Figure 2, no further Pareto improvements are possible—the market is efficient. How do we know? Consider one more trip. The government cannot come out even unless it receives \$100 in revenue to cover the

costs of the additional services it provides. So when we try to find a Pareto improvement from point B , we must include \$100 for the government. Further, airlines would have to receive at least \$700 for the ticket to come out even, since that is the marginal cost of another ticket. So let's see what would happen if the airline sold another ticket for \$800 and gave \$100 to the government:

Action: An airline sells an additional ticket for \$800, with \$100 in excise tax.

Traveler	Gains benefits worth:	\$800
	Pays:	\$800
	Comes out ahead by:	(comes out even) \$ 0
Airline	Receives:	\$800
	Pays tax of:	\$100
	Marginal cost:	\$700
	Comes out ahead by:	(comes out even) \$ 0
Government	Receives:	\$100
	Incurs costs of:	\$100
	Comes out ahead by:	(comes out even) \$ 0

As you can see, everybody comes out just even. Indeed, as long as the government must receive at least \$100, and the airline must receive at least \$700, selling another ticket cannot make *anyone* better off unless someone is harmed. From point B , no Pareto improvements are possible; thus, the equilibrium at point B is an efficient equilibrium.

When a government service is used along with a market good, it is efficient for the government to charge an excise tax on a good equal to the marginal cost of providing a government service used with that good.

What Happens When
Things Change?



The Income Tax. In the last section, you saw that raising *general* revenue with an excise tax (rather than paying for a specific government service) is inefficient. You might think that it would be efficient to raise general tax revenues from the income tax. After all, the income tax is applied to *all* of our income, not to some specific good. But in this section, you'll see that income taxes, too, can create inefficiency.

Figure 3 shows why, using the labor demand–labor supply diagram from Chapter 11. On the vertical axis is the daily wage paid by employers. Their market demand curve for workers is L^D . The workers' supply curve, without an income tax, is L_1^S . After an income tax at a rate of 25 percent is applied to wage income, the labor-supply curve becomes L_2^S . Each point on L_2^S is higher than the point below it on L_1^S .

Why the shift? Consider point C on L_1^S . This point tells us that, when workers are paid \$150 per day with no income tax, 80 million people will want to work. But when the government collects an income tax, workers choose whether or not to supply labor on the basis of what they will be able to keep *after* taxes. Once the 25 percent income tax is introduced, workers must be paid \$200 in order to take home \$150, so it will now take a daily wage rate of \$200 (point B on L_2^S) to get 80 million people to supply their labor. And the same is true of every other point on the labor supply curve: Whatever wage was required before to get any given number of people to work, now it will take a higher wage.

With no income tax, the equilibrium would be at point A , with a wage of \$175 per day. With the income tax, however, the market equilibrium occurs at point B ,

LABOR MARKET EFFECT OF AN INCOME TAX

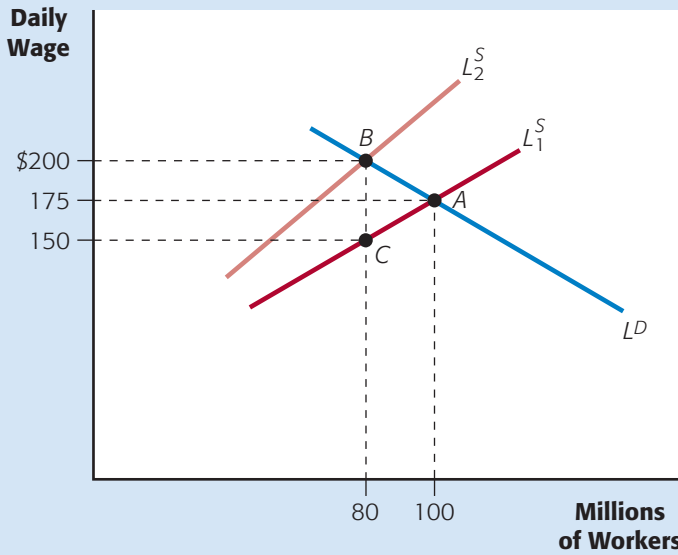


FIGURE 3

At point A, the labor market is in equilibrium with an hourly wage of \$175. Now the government imposes an income tax at a rate of 25 percent on all earned income. The labor supply curve rotates upward to L_2^S . It shows that workers must now be paid more in order to take home the same wage as before the tax. The new equilibrium at point B is inefficient because hiring one more worker would provide some firm with more revenue (\$200 per day) than it would take for the worker to willingly provide the work (\$150 per day).

where the wage paid by employers is \$200 and the wage received by workers, after paying income tax, is \$150. Is the new equilibrium at point B efficient? Not at all, since we can easily come up with a Pareto improvement from that point.

Suppose, for example, that we are at point B, and an additional worker is hired. We know that worker’s marginal revenue product to some firm is \$200 per day, because at point B, each firm hires workers until the *MRP* of labor is equal to the market wage rate (see Chapter 11). Thus, hiring one more worker would give some firm \$200 in additional revenue. We also know that, at point B, there is some worker who would be just indifferent between working and not working at a take-home wage of \$150. So here is our Pareto improvement: Some firm hires one more worker and pays her *more* than \$150 per day, say, \$170. Moreover, we’ll get the government to agree to tax this specific worker only \$10 (instead of the usual 25 percent, which would be $0.25 \times \$170 = \42.50), so the worker will take home \$160 per hour after taxes. The following scorecard shows that everyone benefits:

Action: A firm hires another worker for \$170 per day, and the government collects an additional \$10 per day in taxes.

Worker	Receives wage of:	\$ 170
	Pays income tax of:	\$ 10
	Gives up time worth:	\$150
	Comes out ahead by:	\$ 10
Employer	Gains revenue of:	\$200
	Pays:	\$170
	Comes out ahead by:	\$ 30
Government	Receives taxes of:	\$ 10
	Comes out ahead by:	\$ 10

As the scorecard shows, the worker is better off: She would have been willing to work for a take-home pay of \$150 per day, but she gets \$160 after taxes. The employer's marginal benefit from the added worker, read off the demand curve D , is \$200 at point B , but the employer pays only \$170—an improvement. And the government gets an additional \$10 in taxes that it would not have collected without this move. Everyone comes out ahead.

Unfortunately, governments can't dicker over income taxes for individual employment decisions any more than they do over the taxes for airline tickets. Therefore, our proposed Pareto improvement will not take place. The income tax creates inefficiency in the labor market.

Now, you've seen that two types of taxes—excise taxes (when used for general revenue) and the income tax—are both inefficient. A similar analysis of other taxes commonly used in market economies, such as payroll taxes or general sales taxes, would come to the same conclusion: They are inefficient.

But is there *any* efficient tax we could use to support general government activities? If you look back at Figures 2 (p. 436) and 3, you will see that the taxes we've analyzed create inefficiency by *changing a market price*, such as the market price of airline tickets or the market price of labor. As a result, the equilibrium changes from the *efficient* quantity at point A to the *inefficient* quantity at point B .

An efficient tax, by contrast, would be one that did not affect any price in any market, and therefore would not move the quantity in any market away from the efficient level. For example, imagine if everyone had to pay in taxes the same amount—say, \$2,000—regardless of their income or wealth or how much they bought or sold of any good. This type of tax—called a *lump-sum* tax—would not directly affect the price or quantity in any market, so it would not create any inefficiency. Most of us, however, regard lump-sum taxes as unfair because they require the poor to pay just as much as the rich. As you can imagine, such taxes—even though efficient—are unlikely to be chosen by any democracy. In a society concerned about fairness, taxes will be based on income or other measures that involve some inefficiency:

Since the government can't function without taxes, and the only efficient taxes are unfair, we tolerate the inefficiency that taxes cause.

MARKET FAILURES

Market failure A market equilibrium that fails to take advantage of every Pareto improvement.

A **market failure** occurs whenever a market—left to itself—is inefficient. That is, a market failure occurs whenever the market participants fail to take advantage of every Pareto improvement. In this section, we'll look at three different types of market failures to which economists have devoted a lot of attention: monopoly and imperfect competition, externalities, and public goods. As you'll see, government involvement can often help deal with, and even cure, a market failure. But government involvement has costs as well as benefits, and dealing with market failures remains one of the more controversial aspects of economic policy.

MONOPOLY AND IMPERFECT COMPETITION

In Chapter 14, we saw that a purely competitive market will produce the economically efficient level of output—all Pareto improvements will be made. But we also saw that when competition is less than perfect, firms produce less than the efficient

quantity of output and leave opportunities for mutual gain unexploited. Therefore, market structures other than perfect competition can be regarded as *market failures*.

The most extreme departure from perfect competition is monopoly—a market with only one seller and no close substitutes. A monopolist—like any other less-than-perfect competitor—faces a downward-sloping demand curve. In Chapter 14, we saw that such a firm will produce *less* than the efficient level of output, leaving Pareto improvements unexploited. (You may want to flip back to the previous chapter and review this result now.)

What can the government do? One option is to break up the monopoly into two or more smaller firms that will have to compete with each other. The government has done this on more than one occasion (and you’ll learn about a famous example in the “Using the Theory” section at the end of this chapter). But there is one situation in which breaking up the firm would not make sense: when the monopoly is a *natural* monopoly.

The Special Case of the Natural Monopoly. In Chapter 9, you learned that a *natural monopoly* exists when, due to economies of scale, one firm can produce for the entire market at a lower cost per unit than can two or more firms. If the government steps aside, such a market will naturally tend toward monopoly. Figure 4 illustrates an example of a natural monopoly: an electric utility company in a typical town or city. Because the utility must produce and maintain electric wiring to every neighborhood in the city, the cost per unit (*LRATC*) is very high at low levels of output. However, producing an *additional* unit of electricity is very inexpensive—a constant \$0.15 per kilowatt-hour (kwh) in our example—because it involves only fuel and labor costs. Thus, cost per unit drops as output increases, because the cost of the wiring can be spread among more and more units of output.¹

In the absence of any government intervention, we know what the natural monopolist will do: It will try to make the highest possible profit. Its constraints are familiar: It faces a demand curve for electricity, it uses some particular technology of production, and it must pay for its inputs. All these constraints are illustrated by the demand and cost curves in Figure 4. Following the rule of marginal decision making, the firm will produce 5 million kwh of electricity, where the *MR* and *MC* curves cross. The firm will charge a price of \$0.60—the highest price it can charge in order to sell 5 million kwh per day. This puts the market at point *A* on the monopolist’s demand curve. The firm’s total profit is indicated by the blue rectangle.

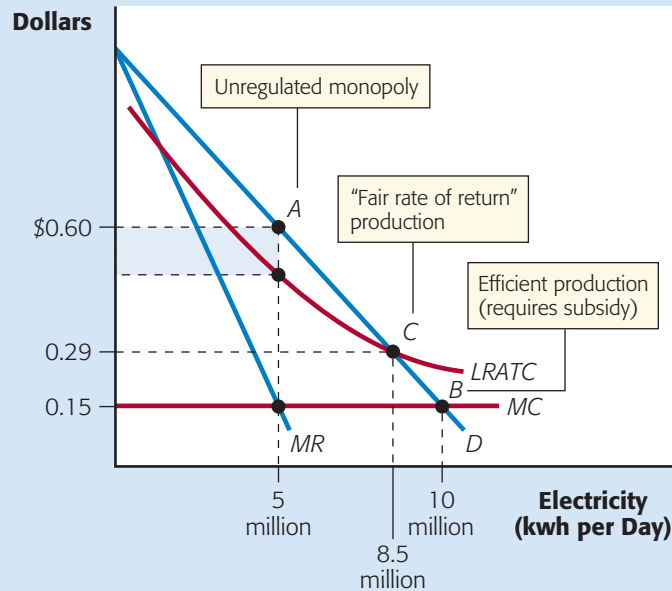
But point *A* represents an *inefficient* level of output. An additional kwh of electricity would be valued by some consumer at \$0.60, but it would only cost the firm \$0.15 to produce. In fact, the efficient level of output in the figure is 10 million kwh per day, *where the MC curve crosses the demand curve*. Once output has risen to this level, a further increase is worth \$0.15 per kwh to some consumer, which is *just* what it would cost the firm to produce it. No Pareto improvement is possible.

However, unless the firm can price discriminate—charging some customers less than \$0.60 while maintaining the price at \$0.60 for all of its current customers—it will not increase its output beyond 5 million kwh, so the market fails to give us the economically efficient output level—a market failure.

¹ Interestingly, the market for electricity—long a standard textbook example of a natural monopoly—has become competitive in some states. In 1998, Californians were given the opportunity to choose among several suppliers of electricity—including those in other regions of the state. By 2000, this choice had spread to several other states. However, the entry of new competitors in this case did not involve any duplication of costs like wiring. Instead, it has come about from a widening of the market to include utilities in other localities. This was made possible by technological advances in information processing and energy transfer.

FIGURE 4

REGULATING A NATURAL MONOPOLY



A natural monopoly has a downward sloping $LRATC$ curve throughout the entire range of market demand. Left unregulated, the monopoly would produce 5 million kilowatt-hours, where marginal cost equals marginal revenue, and earn a profit shown by the shaded rectangle. This quantity is inefficient because the value of the 5 millionth gallon to some consumer (\$0.60) exceeds the marginal cost of producing it (\$0.15).

Government regulators could achieve the efficient outcome by mandating a price of \$0.15 per kwh. Then the monopolist would produce the efficient quantity of 10 million kwh at point B . However, with price less than $LRATC$ at that quantity, the firm would have to be subsidized or it will go out of business.

An alternative is to set price equal to average cost at point C , so that the firm earns a "fair rate of return." The resulting quantity of 8.5 million kwh is still inefficient, but not as inefficient as the unregulated quantity of 5 million kwh would be.

What can the government do?

One policy that would *not* be very effective would be to break up the natural monopolist into several competing firms. Why? Several firms would each supply for only a *part* of the market, so their costs per unit could never be as low as that of a single firm. (*Think:* Each power company would have to provide and maintain its own network of wiring to households, so each would have very high fixed costs, but be unable to spread them among customers in the entire market.) Such a policy could never bring us to point B in Figure 4, since with more than one firm, none of them could ever charge a price of \$0.15 per kwh and stay in business.

If breaking up a natural monopoly is not advisable, what can a government do to bring us closer to economic efficiency? There are two other options: (1) regulation and (2) public ownership. Let's consider each in turn.

At the beginning of this chapter, you learned that under regulation, a government agency digs deep into the operations of a business and takes some of the firm's decisions under its own control. In the case of a natural monopoly, regulators are interested in achieving economic efficiency, which they do by telling the firm what *price* it can charge.

At first glance, you might think that natural monopoly regulators have an easy job. For example, in Figure 4, we know that the efficient quantity is 10 million kwh

per day, and we know that if the price is \$0.15, that is just the quantity consumers will buy. Therefore, all the regulators have to do is set the official price at \$0.15 and—voilà—an efficient market.

Unfortunately, it's not that easy. First, there is the matter of information: The regulators must be able to trace out the firm's *MC* curve as well as the market demand curve. This job is especially difficult when the monopoly's managers—hoping for a higher price—have an incentive to overstate costs. Even with a cooperative monopoly, the job is extremely complex, and the best regulators can hope for is a crude approximation to the actual curves.

More importantly, even with perfect information about the monopolist's cost and demand curves, regulators have a serious problem. If you look again at Figure 4, you'll notice that the *MC* curve lies everywhere *below* the *LRATC* curve. This must be the case for a natural monopoly, since economies of scale—the reason for the natural monopoly—means that the *LRATC* curve slopes downward, and this can only occur when marginal cost is less than average cost. (See Chapter 6 on the marginal-average relationship if you've forgotten why.)

Now you can see the problem for regulators: If they set the efficient price of \$0.15, so that buyers demand the efficient quantity of 10 million kwh per day, the firm's cost per unit is *greater* than \$0.15 per kwh. The firm will suffer a loss. In the long run, it will go out of business.

This problem leaves the regulator with two alternatives. First, it can set price equal to *MC* (\$0.15 per kwh in our example) and *subsidize* the monopoly from the general budget, to make up for the loss. But this would require taxpayers in general—rather than just the monopoly's customers—to help pay for the product. A slight improvement would be to charge all customers a *user fee* for participating in the market. The user fee—which becomes revenue for the monopoly—could be set to just eliminate the monopoly's loss and keep it in business.

In practice, however, regulators in market economies around the world have usually chosen a different solution. The regulators determine a price that gives owners a “fair rate of return” for funds they've put into the monopoly. This fair rate of return is designed to be the same rate of return they *could* have earned in a similar, alternative investment. In other words, the fair rate of return should give the monopoly what economists call *normal profit*—a profit just high enough to cover all of the owners' opportunity costs, including the foregone interest on their own funds.

What price will accomplish this? Remember that we've included *all* costs into our cost curves, including the opportunity cost of owners' funds. Thus, a fair rate of return is already built into the *LRATC* curve in Figure 4. If the firm charges a price equal to *average cost*, it will cover all the costs of the operation, including the fair rate of return for owners. You can see that at point *C*—with a price of \$0.29 per kwh—the firm is charging the lowest possible price that prevents it from suffering a loss. This strategy—called **average cost pricing**—is the most common solution chosen by regulators of natural monopolies. More generally,

with average cost pricing, regulators strive to set the price equal to cost per unit where the LRATC curve crosses the demand curve. At this price, the natural monopoly makes zero economic profit, which provides its owners with a fair rate of return, and keeps the monopoly in business.

Average cost pricing is not a perfect solution. For one thing, it does not quite make the market efficient. For example, notice that in Figure 4, only 8.5 million units are produced, instead of the efficient quantity of 10 million. Nevertheless,



Electric utilities have long been natural monopolies that present a challenge to government regulators.

Average cost pricing The regulatory strategy of setting price equal to a natural monopolist's long-run average total cost.

compared to no regulation at all, average cost pricing lowers the price to consumers, and increases the quantity they buy, bringing us closer to the efficient level.

Averch-Johnson effect The tendency of regulated natural monopolies to overinvest in capital.

Another problem with average cost pricing is that it provides little or no incentive for the natural monopoly to economize on capital. That is, the monopoly can grow larger and larger—taking in more and more new owners by issuing stock and using the proceeds to buy machinery and capital—confident that the regulators will always ensure that the price will be adjusted to assure normal profit. The tendency of regulated natural monopolies to overinvest in capital is known as the **Averch-Johnson effect**, after the two economists who first explained it.² The Averch-Johnson effect is a specific example of a more general idea: that when a firm is not striving to maximize profit (in this case, because the government is guaranteeing a specific rate of return), the firm need not economize on costs.

Other Cases of Less-Than-Perfect Competition. Remember that a natural monopoly is just *one* example of how a breakdown in perfect competition causes a market failure. *Any* market that is not perfectly competitive is, technically, a market failure. In some cases, the departure from perfect competition is so great that the government deems that action is needed. In the “Using the Theory” section at the end of this chapter, you’ll see some examples of how antitrust policy has dealt with market failures in the case of (nonnatural) monopoly and oligopoly.

But a close scrutiny of markets could find less-than-perfect competition—and therefore, a market failure—almost everywhere in the economy, in markets for books, clothing, automobiles, movies, bicycles, computers, and more. What should the government do about all these cases?

Let’s take an example: the less than perfectly competitive market for movie popcorn. Suppose it would cost \$0.50 to make another box of popcorn at a movie theater, but popcorn sells for \$2.00 because the theater owner is the only convenient seller. We know that customers derive a marginal benefit of \$2.00 or more when they choose to buy a box of popcorn. Otherwise, they would not make the choice to buy it at that price. But since price at \$2.00 exceeds marginal cost at \$0.50, there is room for a Pareto improvement. If a customer receives another box and pays, say, \$1.50, the theater makes an extra \$1.00 (\$1.50 received less \$0.50 cost), while the customer gains \$0.50 (\$2.00 marginal benefit minus the \$1.50 payment). But this mutual gain will not occur, unless the movie theater can price discriminate (which is doubtful—see Chapter 9). Lowering the price on one box of popcorn would require lowering the price on *all* boxes. And this would make the theater’s overall profit decrease. Therefore, the additional popcorn will not be produced—the Pareto improvement will not take place.

Should the government use antitrust law to change the imperfectly competitive market for popcorn at the movies into a competitive one? Or should it use some other correction—such as regulation to lower the price of popcorn—to remedy this market failure? Most economists and policy makers would answer both questions with an unqualified “no.”

First, none of the specific acts forbidden by antitrust law causes imperfect competition in the theater popcorn market. It’s a safe assumption that the theater’s owner has not conspired with other owners to set high popcorn prices. Instead, a theater is just not a very promising place for competition to occur. We can’t expect owners to set up competing popcorn stands inside their theaters.

² Harvey Averch and Leland Johnson, “Behavior of the Firm under Regulatory Constraint,” *American Economic Review*, December 1962, pp. 1052–1069.

What about *regulating* the price of popcorn? We could imagine a state theater popcorn commission. It would have a staff to gather data on the costs of making popcorn in theaters and determine, say, once a year, the maximum price that theaters could charge for popcorn, based on average-cost pricing. Would we want our government to set up such a commission? Most likely not. The costs of a government bureaucracy to regulate the price of popcorn would probably exceed the benefits.

The government generally ignores market failures due to imperfect competition when there is no violation of antitrust law, and the product is not important enough to affect our economic welfare in a serious way. Instead the government concentrates on applying its antitrust and regulatory efforts to products that account for larger fractions of most families' budgets and for which an unregulated price would be far above marginal cost, as in the case of electric utilities and local phone service.

EXTERNALITIES

If you live in a dormitory, you have no doubt had the unpleasant experience of trying to study while the stereo in the next room is blasting through your walls—and usually not your choice of music. This may not sound like an economic problem, but it *is* one. The problem is that your neighbor, in deciding to listen to loud music, is considering only the *private* costs (the sacrifice of his own time) and *private* benefits (the enjoyment of music) of his action. He is not considering the harm it causes to you. Indeed, the harm you suffer from not being able to study might be greater than the cost to him of turning down the volume. In this case, his turning down the volume could be a Pareto improvement, with an appropriate side payment. And unless he does turn down the volume, the situation remains inefficient.

When a private action has side effects that affect other people in important ways, we have the problem of *externalities*:

An externality is a by-product of a good or activity that affects someone not immediately involved in the transaction.

Externality A by-product of a good or activity that affects someone not immediately involved in the transaction.

For example, the by-product of your neighbor blasting his stereo is the noise coming into your room. We call this a *negative* externality, because the by-product is harmful. But notice that the definition of externality is not limited to harmful effects. When the by-product is *beneficial* to a third party, it is a *positive* externality. We'll consider examples of positive externalities a bit later.

Negative externalities often arise in social situations. If you are doing a group project, someone who likes to hear himself talk may dominate the conversation and prevent others in the group from making progress. The talker is considering the private costs and benefits of his action (to continue speaking), but not the extra time costs imposed on the group as a whole. In cases like these, rules, social conventions of politeness, or side payments can often solve the problem. In the group project, a reminder might be enough to keep everyone's comments short and to the point. In the case of the loud stereo, a request to turn down the volume might be sufficient, or the dorm might establish a rule forbidding the blasting of a stereo between certain hours. If all else fails, you might offer an implicit side payment: telling your neighbor that you'd be happy to lend him some of your CDs if he would just keep the volume down.

But when negative externalities arise in *markets* with large numbers of affected people, the problem can rarely be solved with social conventions, simple rules, or side payments. In the previous chapter, we saw an example of an externality—the dry cleaner whose fumes annoyed the tenants in an apartment building. We showed that getting the dry cleaner to move, with the appropriate side payment, would be a

Pareto improvement. But chances are, where business is concerned, no social convention or universally accepted rule will get the dry cleaner out. And side payments to the dry cleaner may be problematic. Just because the apartment dwellers would be *willing* to compensate the dry cleaner for his loss doesn't mean that they can actually *arrange* the necessary side payment. In order for this to happen, somebody must have the idea of collective action, negotiate with the parties involved, and then collect money from each tenant. With 100 tenants, that will be quite a difficult undertaking. Some tenants may try to get a free ride—refusing to pay, reasoning that if most others pay up, the dry cleaner will move anyway, and the nonpaying tenants get the benefit for free. This “*free rider*” problem stands in the way of many Pareto improvements. This is why we typically turn to government to deal with externalities that are important and affect many people.

What Happens When
Things Change?



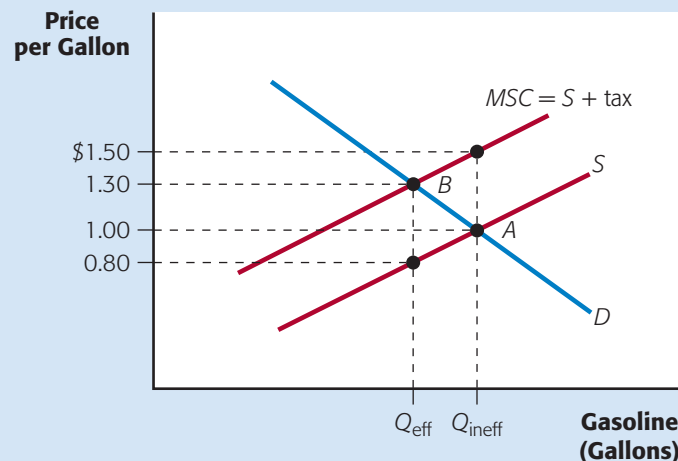
Dealing with a Negative Externality. Many negative externalities result from some kind of *pollution*. The blaring stereo is noise pollution—the addition of undesirable noise to your environment. Cities pollute rivers and lakes with sewage, and industries pollute them with chemicals. Cars and power plants pollute the atmosphere. As you are about to see, pollution—like other negative externalities—creates inefficiency.

Figure 5 illustrates an inefficiency that might result from the production and use of gasoline, which pollutes the air with carbon monoxide and soot, dust and other visible and microscopic solids. In the figure, we assume that the market for gasoline is perfectly competitive. The supply curve reflects the marginal costs of producing gasoline to some firm. We can call this the *marginal private cost*, since it ignores any costs to the general public, such as the health and environmental damage caused by pollution.

The demand curve, D , reflects the *marginal private benefit* of the good. It tells us the value that consumers place on the good when they consider the benefits to

FIGURE 5

A NEGATIVE EXTERNALITY



At point A , the market for gasoline is in equilibrium. However, the use of gasoline creates pollution, imposing a cost on society of \$0.50 per gallon. The equilibrium at point A is inefficient because the marginal social cost (MSC) of \$1.50 per gallon exceeds the marginal private benefit of \$1.00. The government could remedy this externality by imposing a tax of \$0.50 per gallon, equal to the cost of the pollution created. This would shift the supply curve to $MSC = S + \text{tax}$. In the new equilibrium at point B , marginal social cost equals marginal benefit, so point B is efficient.

themselves only. Without any control of pollution, competitive market equilibrium occurs at point *A*, where the supply curve *S* and the demand curve intersect. At this point, the private benefit of the last unit produced is equal to its private cost of production. If there were no externality, this point would be efficient, as we discussed earlier in this chapter.

But this is not the economically efficient output level. Why? Because each gallon of gasoline sold causes harm in the form of pollution, and that is not being considered in the market. Let's suppose that each gallon of gas imposes a cost on the economy of \$0.50. The curve labeled *MSC* tells us the *marginal social cost* of gasoline, which is equal to marginal private cost plus \$0.50. Notice that the *MSC* curve lies *above* the market supply curve. Since there are no important benefits to the public other than those enjoyed by consumers of gasoline, we can assume that the *marginal social benefit*—the marginal private benefit plus the marginal benefits to the general public—is the same as the marginal private benefit curve to consumers—the market demand curve.

Once we draw the separate *MSC* curve in Figure 5, we discover a problem: At the equilibrium output level (point *A*), the marginal social cost of \$1.50 is greater than the marginal social benefit (and marginal private benefit) of \$1.00. That is, the last gallon of gasoline produced in this market costs society more than it benefits society. As you are about to see, this is inefficient.

A market with a negative externality associated with producing or consuming a good will be inefficient. In market equilibrium, the marginal costs to all parties exceeds the marginal benefit to all parties.

Let's demonstrate this general conclusion in the market for gasoline. We'll do this by coming up with a Pareto improvement—a change in output away from point *A*—that does not occur. Suppose some firm were to produce one less gallon of gasoline. The consumer who gives up this gallon would lose a marginal benefit of \$1.00. But now suppose he receives a side payment of \$0.80 from the gasoline producer and \$0.40 from society at large, for a total of \$1.20. The firm pays \$0.80 of the side payment, but avoids a marginal production cost of \$1.00. And society in general pays the remaining \$0.40 of the side payment, while gaining a marginal benefit of \$0.50 from the reduced pollution. We can fill out our usual Pareto improvement scorecard as follows:

Action: A consumer uses one less gallon of gasoline and is given a side payment of \$0.80 from a gasoline producer, and \$0.40 from society in general

Gasoline Producer	Cost saving from producing 1 less gallon:	\$1.00
	Pays share of side payment:	<u>\$0.80</u>
	Comes out ahead by:	\$0.20
Gasoline user	Receives side payment of:	\$1.20
	Loses gasoline benefits worth:	<u>\$1.00</u>
	Comes out ahead by:	\$0.20
Humankind	Gains benefits from less pollution worth:	\$0.50
	Pays share of side payment:	<u>\$0.40</u>
	Comes out ahead by:	\$0.10

We've just demonstrated that, in the market for gasoline, point *A* is inefficient. But now consider point *B*. At this point, the marginal private cost is \$0.80 per gallon, and the negative externality costs society \$0.50 per gallon, so the marginal



Jeffrey Frankel's "Greenhouse Gas Emissions" is an interesting analysis of an important negative externality. You can find it at <http://www.brook.edu/comm/PolicyBriefs/pb052/pb52.htm>.

social cost of producing gasoline equals the marginal benefit of \$1.30 per gallon. This point is efficient.

To see why, let's consider what would happen if there were a *further* decrease in production of one gallon. This would deprive a consumer of gasoline he valued at \$1.30. It would release resources at some gasoline-producing firm worth only \$0.80, so that is the maximum side payment the firm could make without being harmed. Finally, the maximum side payment that could come from the general public—without causing them a loss—would be \$0.50. Thus, the maximum possible side payment that could be paid to the consumer is $\$0.80 + \$0.50 = \$1.30$, which is just enough to compensate the consumer for the loss of the gasoline. No one comes out ahead in this move, so there is no reason to make it! At point *B*, all Pareto improvements have been exploited—the market has reached its efficient point.

Now, how can we get this efficient result? Only through government action. It would take too much time and trouble for individual gasoline producers and consumers to arrange the appropriate side payments and production cutbacks, and to monitor the arrangement after everyone agreed.

One method government could use to move the gasoline market to point *B* would be a tax. In Figure 5, suppose the government imposed a tax equal to \$0.50—the harm caused by each additional gallon of gasoline. Then, in addition to paying for its other inputs, each firm in this market would have to pay an additional \$0.50 to the government. This would raise each firm's marginal cost of production by \$0.50—that is, it would make the marginal *private* cost equal to the marginal *social* cost. As a result, the market supply curve would shift upward from the curve labeled *S* in the figure to the curve labeled $MSC = S + \text{tax}$. Notice that, as a result of the tax, the new market supply curve intersects the demand curve at the efficient point *B*. Once the tax is imposed, the market will *automatically* reach the economically efficient output level—point *B*.

A tax equal to the difference between marginal social cost and marginal private cost can correct a negative externality and make a market efficient.

A tax is not the only way to correct a negative externality. The government could instead use regulation to move the gasoline market to the efficient point. Regulators could tell car owners how much they could drive, or tell car producers how much pollution their vehicles are allowed to create. Indeed, this last regulation—pollution restrictions on new automobiles—has been the method of choice for reducing automobile pollution in most states. But whether taxes, regulation, or some other government policy is used, the conclusion remains the same: In the presence of a negative externality, the market, by itself, will produce “too much” output—too much to be efficient. Government intervention is needed to decrease output to the efficient level.

What Happens When
Things Change?



Dealing with a Positive Externality. What about the case of a positive externality, in which the by-product of an activity or a service *benefits* other parties, rather than harms them? Once again, the market will not arrive at the economically efficient output level—but in this case, output will be *too low*. To see why, consider the market for a college education. Each of us, in deciding whether to go to college, takes account of the private costs (tuition, room and board, what we could have earned instead of going to college) and the private benefits (a higher-paying and more interesting job in the future, the enjoyment of learning). But by becoming educated, you also benefit other members of society in many ways. For example, you will be a more informed voter and thereby help to steer the government in direc-

tions that benefit many people besides you. If you major in chemistry, biology, or mechanical engineering, you may invent something that benefits society at large more than it benefits you. Or you may learn concepts and skills that make you a more responsible member of your community. Thus, the market for college education involves a positive externality.

Let's see why a competitive market in college education—with no government interference—would not produce the economically efficient amount of education. Figure 6 shows the market for bachelor's degrees. Without a policy to correct for the externality, the market will be in equilibrium at point A, where the marginal private benefit curve (demand curve) intersects the marginal private cost curve (supply curve). In this equilibrium, the demand curve tells us that the last student who buys a college education values it at \$50,000.

But this is not the economically efficient output level. Why? Because each time a student goes to college, the general public benefits in ways that are not being considered in the market. Let's suppose that we can measure these benefits to the general public, and that they amount to \$15,000 per additional bachelor's degree. In Figure 6, the curve labeled *MSB* tells us the *marginal social benefit* of another bachelor's degree, which is equal to the marginal private benefit plus \$15,000. We'll assume that the marginal social cost of an additional bachelor's degree is the same as the marginal private cost. That is, there are no negative externalities in this market—only positive externalities. Thus, the market supply curve is also the marginal social cost curve.

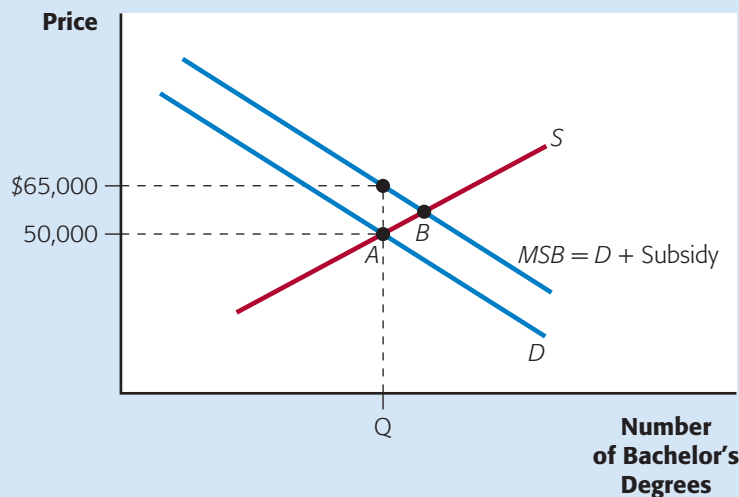
Now you can see the problem: At the equilibrium (point A), the marginal social benefit is greater than the marginal social cost. At point A, there are additional students who—if they went to college—would provide additional benefits (to themselves



In "Government's Role in Primary and Secondary Education," (http://www.dallasfed.org/html/pubs/pdfs/er/er_99_01.pdf), Lori Taylor explores the rationale for government funding of education.

A POSITIVE EXTERNALITY

FIGURE 6



At point A, the market for college education is in equilibrium. Education, however, creates a positive externality for other members of society. Point A is not efficient because the marginal social benefit (*MSB*) of \$65,000 per degree exceeds the marginal cost of production, equal to \$50,000. Government could remedy the externality by subsidizing higher education. This would shift the demand curve upward to $MSB = D + \text{Subsidy}$. The new equilibrium at point B would be efficient, since marginal social benefit equals marginal production cost there.

and society) greater than the additional costs. Unless these students go to college, we are not taking advantage of a potential Pareto improvement. Point A is not efficient.

More generally,

a market with a positive externality associated with producing or consuming a good will be inefficient. In market equilibrium, the marginal benefit to all parties exceeds the marginal cost to all parties.

If you are still not convinced, here is just one example of a Pareto improvement that is not occurring when we are at point A: One more student goes to college accompanied by a \$10,000 side payment from the people of the United States. The side payment is made to the college, which passes half of it on to the student in the form of a reduction in tuition from \$50,000 to \$45,000. The following scorecard shows how everyone could gain in this example.

Action: One more student goes to college with a \$10,000 side payment from society to the college.

College student	Gains benefits worth:	\$50,000
	Pays:	<u>\$45,000</u>
	Comes out ahead by:	\$5,000
College	Receives tuition from student of:	\$45,000
	Receives side payment of:	\$10,000
	Incurs cost of:	<u>\$50,000</u>
	Comes out ahead by:	\$5,000
Citizens of the United States	Gain benefits worth:	\$15,000
	Pay side payment of:	<u>\$10,000</u>
	Come out ahead by:	\$5,000

Now that we know that the market equilibrium at point A is inefficient, what point represents the *efficient* output level? The answer is: point B, where the marginal social benefit of a college education and the marginal social (and private) cost are equal. From point B, there are no Pareto improvements left to make—no changes in output for which we could find a side payment that would make one person better off and harm no one.

How can we move a market with a positive externality to a more efficient outcome, such as at point B?

One answer: The government could subsidize college education by \$15,000 per bachelor's degree. The subsidy could be provided to the student or the school; the effect will be the same in either case. In our diagram, we suppose that the subsidy goes to the student. (As an exercise, see if you can draw the diagram for the case where the subsidy goes to the college. *Hint:* In this case, it will affect the market supply curve, not the market demand curve.)

Once students receive a subsidy of \$15,000, the market demand curve will shift upward by \$15,000. Whatever price led to a certain number of degrees *before* the subsidy, now that price will be \$15,000 higher. That is, if 2 million bachelor's degrees per year were demanded without the subsidy at a price of \$50,000 each, then, after the subsidy, the same 2 million degrees would be demanded at a price of \$65,000. Of that price, buyers would only be paying \$50,000 out of their own pockets, and the rest would be paid by the government. Notice that, as a result of the subsidy, the new market demand curve intersects the supply curve at the efficient point B.

A subsidy equal to the difference between marginal social benefit and marginal private benefit can correct a positive externality and make a market efficient.

PUBLIC GOODS

One of the major roles of government in the economy is to provide **public goods**. These are goods that the market *cannot* provide, and *should not* provide, because of their unique characteristics. It is left to the *government* to provide public goods in the efficient quantities, usually free of charge.

To understand what makes a good public, rather than private, let's begin by noting two features of **private goods**, those supplied by private firms in the marketplace. First, a private good is characterized by **rivalry** in consumption—if one person consumes it, someone else cannot. If you rent an apartment, then someone else will *not* be able to rent that apartment. The same applies to virtually all goods that you buy in the marketplace: food, computers, furniture, and so on. Rivalry also applies to privately provided services. For example, the time you spend with your doctor, lawyer, or therapist is time that someone else will *not* spend with that professional.

Most of the goods and services we've considered so far in this text are rival goods. By allowing the market to provide rival goods at a price, we ensure that people will properly take account of the costs of their decisions to use these goods. If these goods were provided free of charge, people would tend to use them even if their value were less than the value of the resources used to producing them. Moreover, offering a rival good free of charge enables some people who don't value the good very highly to grab up all available supplies, depriving others who might value the goods even more. Thus, leaving such goods to the market—which will charge a price reflecting their marginal cost—tends to promote economic efficiency. We conclude:

If there is rivalry in consumption of a good, the private market should provide it.

A second feature of a private good is **excludability**, the ability to exclude those who do not pay for a good from consuming it. When you go to the supermarket, you are not permitted to eat frozen yogurt unless you pay for it. The same is true when you go to the movies, purchase a car, or attend college. But imagine a situation in which firms could not prevent nonpayers from consuming a good. Then the market would be *unable* to provide the good, because no firm will willingly offer it for sale. Without excludability, no customer would pay for the good, since it can be consumed with or without paying.

Private goods have two characteristics: rivalry and excludability. Rivalry suggests that the market should provide the good, and excludability suggests that the market will provide the good.

But not all goods have these two characteristics. Consider, for example, an urban park located in an area where many people pass by during the day. People will enjoy walking by the park, just because it is pretty to look at. But to provide and maintain it requires many resources and raw materials: the labor of landscape architects and gardeners, gardening tools, fertilizer, flower bulbs, and so on. However,

Public good A good that is non-rivalrous and non-excludable; the market cannot, and should not, provide such goods.

Private good A good that is rival and excludable, and is supplied by private firms in the marketplace.

Rivalry A situation in which one person's consumption of a good or service means that no one else can consume it.

Excludability The ability to exclude those who do not pay for a good from consuming it.

a walk-by park is, essentially, *nonexcludable*. If a private firm provided the park, the firm could not limit the benefits of walking by to those who paid for it. (Yes, it could construct a giant fence, but that would prevent *everyone* who walked by from seeing and enjoying the park.) For this reason, a private firm would have difficulty surviving by creating and maintaining an urban, walk-by park.

“But wait,” you may think. “Couldn’t the firm *ask* people to contribute according to the importance they place on the park?” Yes, but then each individual would have an incentive to downplay its importance and pay nothing. This is the *free rider problem* mentioned earlier in this chapter:

When a good is nonexcludable, people have an incentive to become free riders—to let others pay for the good, so they can enjoy it without paying.

Privately provided urban parks would face an extremely serious free-rider problem. People would reason that their own contribution would make such a small difference to the park fund that, other than moral obligation or a sense of social responsibility, they would have no reason to pay at all. There would be so many free riders that those who did pay would share a very heavy burden—too heavy to bear. Thus, a private firm would be unable to provide this service at all—it would not be able to stay in business.

When a good is nonexcludable, the private sector will not provide it. If we want such a good, government must provide it.

In addition to being nonexcludable, urban parks are *nonrival*: One person can consume or enjoy passing by the park without anyone else consuming or enjoying less of it. Moreover, it uses up *no more of society’s resources* when the benefits of the park are extended to an additional person. For this reason, even if the private sector *could* somehow charge us according to our consumption of the view as we walk by the park, it *should not* charge us. Why not? Because by charging a price each time we walk by, it would force each of us to consider a personal cost that does *not* correspond to any opportunity cost for society. Each time an additional person sees the park, a Pareto improvement takes place: That person gains, and no one loses. Thus, to be economically efficient, *everyone* who places *any value at all* on seeing the park should be able to see it. But this will only happen if the price of seeing the park is *zero*.

This leads us to an important conclusion: Since the economically efficient price for the park is zero, private firms—which would have to charge a positive price—should not be the ones to provide it. That is, even if a firm *could* exclude those who do not pay, it *should not* do so. By charging a positive price, the number of people deciding to pay and enjoy the park would be below the economically efficient level.

When a good or service is nonrival, the market cannot provide it efficiently. Rather, to achieve economic efficiency, the good or service would have to be provided free of charge.

We’ve now seen that goods can be either rival or nonrival, and either excludable or nonexcludable. This leads to four possible combinations of characteristics, as shown in the following table:

	Excludability	Nonexcludability
Rivalry	Private Good: Market should and will provide.	Mixed Good: Market will not provide at all.
Nonrivalry	Mixed Good: Market will provide too little.	Public Good: Market should not and will not provide.

At the upper left are private goods—the types of goods we’ve been discussing in earlier chapters. These include bed frames, carwashes, soybeans, extermination services, electricity, and so on. The market, possibly assisted by corrections for externalities, will provide them efficiently.

The lower left is a mixed category—a good that is neither purely private nor purely public. This type of good is becoming increasingly important, because it includes most information products. Consider the software produced by Microsoft. Microsoft has the power to exclude you from using its software, at least in principle. Yet your use of that software costs Microsoft hardly anything at all—software is so easy to copy that it is effectively a nonrivalrous good. The same reasoning that has led us to make weather reports freely available argues for allowing unlimited free copying of software. But in fact, copying software is a violation of federal copyright law. And we do not generally subsidize the writing of software. Thus, public policy deliberately fails to distribute software efficiently among computer users. Why?

The answer concerns incentives. Unless Microsoft can charge for the use of its software, it would have no incentive to develop new products. The same principle applies to many other information products, such as patented ideas, paintings, books, movies, and trademarks. In the absence of patent protection, no new products would be developed by private firms.

In the upper-right quadrant is another mixed category—a good that is neither purely public nor purely private. In this category, goods are rivalrous—so they *should* be sold for a price—but excludability is difficult or impossible, so the market *will not* provide them. City streets, urban parks, and some important natural resources fall into this category. Economists use the term **tragedy of the commons** to describe the problem that develops when people can’t be excluded from using rivalrous goods. In a traditional English village, the commons was an area freely available to all families for grazing their animals. Grazing rights are a rivalrous good—if one cow eats the grass, another cannot. But the commons had no method of exclusion, so it was overgrazed, causing harm to *all* families.

In modern life, the most important example of the tragedy of the commons is in the use of roads. With few exceptions, government provides us with roads, and we can use them as much as we want, free of charge; they are regarded as largely nonexcludable. Yet space on the road is completely rivalrous—two cars can’t use the same place on the road at the same time. As a result, we have the tragedy of the commons: traffic jams at peak times, as commuters overuse roads because they are not taking into account the delays they cause other drivers. On some roads, government or private firms have excluded some users by charging tolls, but these are the exceptions to the rule.

A newer example of the tragedy of the commons began developing late in 1999 and early in 2000, when streaming media came into common use over the Internet. Streaming video, for example, allows the viewer to see an uninterrupted movie

Tragedy of the commons The problem of overuse when a good is rival but nonexcludable.

image on her computer. Streaming audio does the same for music and other sounds. The problem is that these streaming data packets are so large that they often cause congestion on the Web. Moreover, streaming data—in order to provide uninterrupted sounds or images—always gives itself priority over other Internet traffic, violating the rules that govern all other Internet data. As a result, if we are all free to view a live broadcast or a film clip on our computers whenever we want, we may be creating huge delays for more normal sorts of Internet traffic—delays that make us all worse off when the system as a whole is considered. As streaming becomes a bigger issue, the government and private institutions that regulate the Internet must evolve to charge for these streams, just as users of the telephone get charged for long phone calls. Unless this occurs, a true tragedy of the commons may develop.

Finally, at the lower right corner in the table, we find public goods: those that are nonrival—so the private market *should not* provide them—and nonexcludable—so the private market *will not* provide them. Most of us agree that public goods should be provided by government, and governments around the world do so. In addition to parks, public goods include national defense, police and fire protection, and the legal and regulatory infrastructure we discussed earlier in this chapter. (See if you can explain why this infrastructure is both nonexcludable and nonrival.)

Keep in mind, though, that just because government provides a good does not automatically make it a *public good* in the economic sense. Some governments in other countries own banks, manufacturing firms, and media companies. These governments provide goods and services, but economists would call them private goods because they are rivalrous and excludable. We categorize goods into public and private based on their *characteristics*, not which sector ends up providing them.

EFFICIENCY AND GOVERNMENT IN PERSPECTIVE

In this chapter, you've seen that an economy with *well-functioning, perfectly competitive markets* tends to be economically efficient. But notice the italicized words. As you've seen in this chapter, many types of government involvement are needed to ensure that markets function well and to deal with market failures. The government helps markets to function by providing a legal and regulatory infrastructure. In extreme cases of imperfect competition, government antitrust action or regulation may be needed. The government imposes taxes and subsidies to deal with externalities. And the government itself steps in to provide goods and services that are nonrival, nonexcludable, or both.

These cases of government involvement are not without controversy. In fact, most of the controversies that pit Democrats against Republicans in the United States (or Conservatives against Labourites in Britain, or Social Democrats against Christian Democrats in Germany) relate to when, and to what extent, the government should be involved in the economy. Debates about public education, Social Security, international trade, and immigration center on questions of the proper role for government. Some of the disagreement is over the government's role in bringing about a more *fair* economy, but there is also debate about the government's role in bringing about economic efficiency.

These controversies are so heated, and so varied, that it is easy to forget how much *agreement* there is about the role of government. Anyone studying the role of government in the economies of the United States, Canada, Mexico, France, Germany, Britain, Japan, and the vast majority of other developed economies, is struck by one glaring fact: Most economic activity is carried out among private individu-

als. In all of these countries, there is widespread agreement that—while government intervention is often necessary—the most powerful forces that exploit Pareto improvements and drive the economy toward efficiency are the actions of individual producers and consumers.

CASE STUDIES OF ANTITRUST AND REGULATION

In this chapter, we described two tools—antitrust law and regulation—that can help solve an important problem: the inefficiency caused by imperfect competition. Now we'll consider some examples in which the U.S. government has stepped into a market and used those tools to try to make the market more efficient. As you'll see, the government has often, but not always, succeeded.

BREAKING UP A MONOPOLY: ALCOA

The Aluminum Company of America (Alcoa) invented modern aluminum production and was the only seller of aluminum in the United States for many years. (The only reason why Alcoa was not a *pure* monopoly was that some aluminum was imported from abroad.) In 1937, Alcoa became the target of antitrust action by the U.S. government. With price about 60 percent above marginal cost, the aluminum market was distinctly imperfectly competitive. The government's case was that Alcoa had maintained its near monopoly by preventing new producers from entering the market. For example, aluminum production requires huge amounts of electricity. Alcoa had signed agreements with electric power producers that prohibited them from selling power to any other aluminum maker.

Because of actions like these, the Supreme Court determined that Alcoa had violated antitrust law and ordered the company's partial breakup. Alcoa's Canadian branch was turned into a separate, competing company. At the same time, the government also established two new competitors—Reynolds and Kaiser Aluminum. This unusual step—in which the government actually created new private firms—was possible because the government had built aluminum plants during World War II. The decision to set up the two new firms rather than sell the plants to Alcoa reflected a concern about the efficiency of the aluminum market and a desire to make it more competitive.

As a result of the two policies—the antitrust action and the government-created firms—the aluminum market became much more competitive. The experience with Alcoa is one example of a more general conclusion:

Government policy can make a market more competitive, with lower prices, by transforming a monopoly market into a market with competing companies.

REGULATION AND DEREGULATION: THE AIRLINES

The airline industry provides many examples of actual and proposed policies for improving efficiency and correcting market failures. Early in the development of air travel—in the 1930s—the federal government decided that the airline market was sufficiently important, and was performing sufficiently poorly, that regulation was needed. A regulatory agency, the Civil Aeronautics Board (CAB), was created to determine which airlines would be allowed to operate, where and when they could fly,

Using the
THEORY



and how much they could charge. The CAB continued to regulate the industry until 1978, when it was abolished.

One of the concerns about the airline business—both at the time the CAB was created and today—is that airlines might charge *too little* for their seats as they competed for customers. This might lead bargain airlines to cut corners on safety. Another concern is more subtle, but merits serious study: Low prices may be a way to keep potential rivals out of a market or even to drive an existing rival out of business. Subsequent experience has shown that the CAB created more inefficiency than it cured.

Why? First, the CAB essentially prohibited any new airlines from starting up. Second, it set high fares and restricted routes and schedules for the airlines that it permitted to operate. The strongest evidence of the inefficiency caused by regulation came from markets for travel within California and Texas, where the CAB had no regulatory power. Each of these states had a vigorous, successful *unregulated* airline that offered far lower fares and superior service to the regulated airlines. One of those airlines, Southwest, has extended those benefits to many other states since it became free to operate everywhere it chooses.

The CAB's regulation of the airline industry was unsuccessful and was ended in 1978. After deregulation, many new airlines started up. Although all carriers were free to charge fares as high as they wanted, the average level of fares fell substantially, in comparison to the prices of other goods and services. As a result, the volume of air travel has risen sharply.

This kind of practical experience with regulation in the airline business has made many observers cautious in recommending that other industries be regulated. Indeed, the airline industry has been transformed by deregulation. Hundreds of new airlines have started up, and many existing carriers, especially Southwest, have grown enormously.

In practice, regulation often has the undesirable effect of preventing the entry of new competitors into the market. When that occurs, the removal of regulation will result in increased competition and lower prices.

What accounted for the failure of a regulatory agency like the CAB to deliver efficiently low prices? What lessons can we learn about regulation in general? The most important lesson is this: *Regulatory agencies often fall under the influence of the firms they regulate.* The airlines that benefited from the CAB's protection persuaded the agency to oppose competition and low prices. The flying public was less effective in lobbying the CAB to provide these benefits.

Although deregulation *improved* efficiency in the airline industry, significant inefficiency remains. The deregulated airline industry is nowhere close to purely competitive. Despite conditions that might seem favorable to competition—many firms either selling in each market or capable of entering the market easily—there are signs of imperfect competition. The strongest sign is the extremely high price of unrestricted tickets. If you don't book ahead and stay over Saturday, a transcontinental trip may cost you close to \$2,000, about 40 cents a mile. But the marginal cost of supplying air travel is only around 8 cents a mile, and bargain fares are generally around this level.

Another sign of limited competition is that on some routes there is only a single airline flying nonstop, at very high fares. Consider flights to Minneapolis. There are 6 nonstop flights a day from Boston, 14 from Detroit, 6 from Los Angeles, 7 from Memphis, 9 from New York, 7 from San Francisco, and 14 from Washington, DC-Baltimore. But *all but six of them are flown by Northwest Airlines.* And fares on these routes are relatively high.

Why don't other airlines enter these markets and compete with Northwest? After all, there are airlines that already operate at both airports. While they currently have no flight connecting the two airports, they could add one in a few days. So, couldn't they share in some of Northwest's profit? What stops them?

In the strategic interaction between Northwest and its rivals, what matters to an airline thinking about entering one of these markets is not what Northwest is doing *now*, but what it might do *later*—after a rival airline has entered the market. Northwest can make a substantial profit flying passengers from New York to Minneapolis at relatively high fares, but can discourage rivals by convincing them that it would set *low* fares should one of these rivals choose to enter the New York-Minneapolis market.

But the story of competition in the airline industry is still unfolding. And in late 1999, the government gave a strong boost to competition in the industry. That year, the U.S. Department of Justice started an antitrust case against American Airlines over its response to other airlines that tried to compete with American on routes to Dallas-Fort Worth. The government believes that American violated the antitrust law by cutting fares and expanding service when rivals entered the market, and then raising fares and cutting service when the rivals found they could not make money.

The government's action against American may have already paid off for the consumer, even though—as this is being written—the case had not yet been heard in court. Merely bringing the antitrust suit seems to have emboldened at least some competitors to enter routes from which they had formerly stayed away. For example, in May of 1999, shortly after the American Airlines case was started, a small, low-fare airline, Sun Country, started competing with Northwest on a number of its previous monopoly routes. (In fact, Sun Country flies all six of the flights to Minneapolis that Northwest does not fly. See the list above.) It appears that one factor in Sun Country's decision to compete on these routes was a belief that—in light of the antitrust suit against American—Northwest would not dare react aggressively.

PRESERVING COMPETITION: SOFT DRINKS

In the mid-1980s, Pepsi announced its intention to buy 7-Up, and Coca-Cola suggested it might buy Dr Pepper. Mergers between rivals may reduce competition, and the result may be higher prices and greater inefficiency in the market. But policy makers must consider two other factors before concluding that a merger is harmful. First, a merger may result in reduced costs—say, because a larger firm could take better advantage of economies of scale. In this case, the merger would make a contribution to productive efficiency, which might even outweigh the reduction in efficiency from reduced competition. Second, the impact of a merger may be so small that it is not worth trying to prevent it, just as policy makers choose to do nothing about overpriced popcorn.

The Antitrust Division of the Justice Department has published the criteria it uses to screen mergers for possible challenge. The criteria look at market shares—the fraction of total market sales accounted for by each seller. Here are the market shares of the leading soft drink manufacturers at the time of the proposed mergers:

Seller	Share of Soft-Drink Market (percent)
Coca-Cola	39
Pepsi-Cola	28
Dr Pepper	7
7-Up	6
RJ Reynolds (Canada Dry and Sunkist)	5
All Others	15

Herfindahl-Hirschman Index The sum of squared market shares of all firms in an industry.

If there were a large number of sellers, each with a small share, then the merger of a pair of sellers would probably have little effect on competition and price. But in this case, two of the companies proposing to enlarge themselves through mergers—Coca-Cola and Pepsi—had shares around a third of the market.

In practice, the Justice Department decides its stance toward mergers based on the **Herfindahl-Hirschman Index (HHI)**. That index is the sum of the squared percent market shares of all the sellers in the market. If the market is a monopoly, in which one seller has a 100 percent market share, the HHI has its highest possible value of $100^2 = 10,000$. If there were 100 sellers, each with a share of 1 percent—probably a quite competitive market—the HHI would be $1^2 + 1^2 + 1^2 + 1^2 + \dots = 100$. In general, *the higher the HHI, the more concentrated (and less competitive) is the industry*. The screening rules of the Justice Department are:

If the HHI is	And the change in the HHI from a proposed merger is	Then
Less than 1,000	—	Don't challenge the merger
Between 1,000 and 1,800	Less than 100	Don't challenge the merger
Between 1,000 and 1,800	Above 100	Consider challenging the merger
Above 1,800	Less than 50	Don't challenge the merger
Above 1,800	Above 50	Consider challenging the merger

Before the proposed mergers, the HHI in the soft drink industry was about $39^2 + 28^2 + 7^2 + 6^2 + 5^2 + 15(1)^2 = 2,430$. (As an approximation, we account for the last 15 percent of the market by assuming that 15 small firms each make about 1 percent of total sales.) Thus, the industry falls into the “above 1800” category of the rules. As you can calculate on your own, Pepsi's acquisition of 7-Up would have raised the HHI by 336 points. (Subtract 28^2 and 6^2 , and add $(28 + 6)^2$.) If Coca-Cola bought Dr Pepper, the HHI would have risen by another 546 points. Each of these mergers would have raised the HHI by more than 100 points, and—not surprisingly—the government announced that it would oppose both acquisitions. Faced with the prospect of a time-consuming and costly investigative process, and a likely court battle, both companies eventually abandoned their plans.

The discouragement of the soft drink merger is just one example of a more general conclusion:

Antitrust policy can sustain competition by preventing mergers between large competitors.

The effectiveness of antitrust policy in preserving competition goes far beyond the mergers actively discouraged by the Justice Department. Many mergers that might otherwise occur never even reach the planning stage, because the government's negative response can be easily predicted.

AN ONGOING CHALLENGE: MIGHTY MICROSOFT

The goods and services produced by the giant Microsoft Corporation touch all our lives on a daily basis. With almost any computer you use (other than an Apple product), you encounter Microsoft each time you boot up into Windows. Even with an Apple computer, you are likely to use Microsoft Word or Excel. And even if you don't use a computer, you may have frequent dealings with people who do. For example, when you book a flight, visit your doctor, or buy textbooks, the businesses

you deal with may keep track of their accounts using Microsoft products. Your economics instructor may even keep track of student grades using Microsoft software.

The products that Microsoft sells can be divided into three broad categories:

1. Windows operating system software, leased to computer makers and sold to users.
2. Applications software, such as Word, Excel, and Internet Explorer.
3. Information about current and future versions of Windows that would be helpful to developers of other applications, such as word processors, spreadsheets, and Internet browsers, as well as to developers of other operating systems that could run Windows applications.

All of these products are close to nonrivalrous. Software is basically information—it doesn't reduce one user's benefit from software if another user has the same software. The only cost of equipping another user is the minimal cost of copying the software. The efficient provision of nonrivalrous products requires a price of close to zero.

Yet Microsoft charges a lot more than zero for its software. Windows costs \$90 at retail, Windows NT hundreds of dollars, and Word and other applications \$100 or more. Information intended for software developers comes on a CD-ROM that costs several hundred dollars.

Some of the other problems that have been identified by Microsoft's critics are:

1. Microsoft's contracts with computer makers and Web sites made it difficult for other software suppliers to displace Microsoft, even when the rivals offered better terms to the computer makers. The contracts did not allow computer makers to reduce their payments to Microsoft even if they equipped some of their machines with non-Microsoft operating systems.
2. In one applications market—electronic checkbook software—Microsoft tried to reduce competition by merging with its leading rival, Quicken.
3. Microsoft has used the threat of low prices to keep rivals out of its markets and has actually used low prices, including giving away software for free, to displace existing software. For example, the government has complained that Microsoft gives away Internet Explorer as part of Windows.
4. Microsoft does not reveal all of the secrets about Windows to outside developers. This gives Microsoft advantages in applications markets and makes it difficult for anyone else to develop an operating system that will run Windows programs. For example, WordPerfect may lag behind Microsoft Word in the word-processing market because software developers for WordPerfect don't have the same information about new features of Windows that Microsoft developers have.

Are there solutions for these problems based on the tools described in this chapter? No easy ones. Let's take the problem of efficient pricing first. Hardly anyone thinks that Microsoft should be required to set efficient prices—close to zero—for its software. At such prices, Microsoft would have no incentive to develop software in the future. And no one believes that software should be developed and supplied free of charge by the government, like weather reports. Given the practical necessity for software to be written by private companies like Microsoft, it is necessary to give those companies the opportunity to earn profits to pay for development costs. The only practical source of these profits is the right to sell the software for well above the cost of copying it.

What about the anticompetitive behavior that Microsoft allegedly engages in? The U.S. Department of Justice has been investigating various claims of illegal behavior by Microsoft since 1994 and has taken three important actions. First, it reached an agreement in which Microsoft—without admitting anticompetitive behavior—would

change its contracts with computer makers so that rivals could have a chance to sell their alternative operating systems. (So far, that agreement has had little effect, because there are no effective rivals in the market.) Second, the government blocked Microsoft's proposed merger with Intuit, the company that makes Quicken. Since the government took that step, competition between Microsoft's checkbook software and Quicken has intensified, to the benefit of the consumer.

Third, the government took Microsoft to court in 1998 for a number of practices that limited the chances that a rival to Microsoft Windows would emerge. According to the government's case, if Microsoft had obeyed the antitrust laws, there is some chance that we would be using advanced Web browsers to do computer work that instead we still have to do with Windows.

Notice that the first two steps actually taken by the Justice Department had a surgical character—they dealt with very specific problems with equally specific solutions. They did not try to deal with bigger issues, such as efficient pricing. These larger issues are under active consideration as the government's case against Microsoft progresses.

The most controversial problem is the last one on the list—Microsoft's policies about providing outsiders information about Windows. Although Microsoft provides developers with a bewildering volume of information, there is even more that is known only inside the company. And people inside Microsoft learn about forthcoming improvements to Windows before those on the outside. Many people in the software business—with some support from economists—have called for much more aggressive use of antitrust laws to give outside software developers a better chance. They propose that outside developers of applications have all the information that Microsoft's applications developers have, in order to permit the outsiders to compete effectively with Microsoft.

Another proposal is that Microsoft should be required to publish the computer code for Windows, so that outsiders could figure out how it works. Finally, some have proposed that Microsoft be divided into separate companies for Windows and for applications software, so that all applications software developers would have the same information.

All of the proposals to limit Microsoft's activities may cause more harm than good. For example, splitting Microsoft into Windows and applications companies would eliminate the advantage that Microsoft's applications developers now have. But it would also deny consumers the benefit of product improvements that are possible at Microsoft because the same teams work on both the operating system and the other software.

Would regulating Microsoft make sense? Proposals for regulation have focused on giving outsiders a larger role in setting standards and on providing information more quickly about the new operating system features that Microsoft is developing. But the uneven success of regulation in other industries makes many people skeptical of the wisdom of setting up any kind of regulation in software.

The Microsoft Corporation, a dominant firm in a new industry, presents special challenges to the government. Although it is close to a monopoly, applying the standard tools of antitrust law or regulation very broadly would have costs as well as benefits.

Microsoft continues to be an active developer of new and better software—it has not been held back significantly by the measures taken so far. But it is important to keep in mind that Microsoft's behavior has been strongly influenced by the U.S. economic and legal system in which it operates—a system that provides strong incentives to behave as an honest competitor.

S U M M A R Y

When markets fail to achieve economic efficiency—when they leave potential Pareto improvements unexploited—government can often step in and help. Governments also have a role in providing the institutional infrastructure that helps markets thrive. This chapter is about government’s role in economic efficiency.

The legal system run by governments is a key element of institutional infrastructure. Criminal law limits exchanges to voluntary ones. Property law contributes to enforceable property rights. Contract law helps improve the efficiency of exchange when one party must go first, while tort law affects interactions among strangers.

Finally, antitrust law attempts to prevent harm to consumers from limited competition. In addition to the legal system, the government’s regulatory system affects many aspects of economic life.

To finance their operations, governments rely mainly on tax revenues. But taxes have other economic effects. In some

cases, they create inefficiencies and prevent Pareto improvements from taking place. In other cases, they can be used to improve economic inefficiency. Two important types of taxes are the excise tax—a sales tax on a particular product—and the income tax.

A market failure occurs when a market, left to itself, fails to achieve economic efficiency. Imperfect competition, externalities, and public goods are examples of market failures. Governments have a variety of tools to correct these failures. Through antitrust action and regulation, governments can sometimes narrow the gap between price and marginal cost in imperfectly competitive markets. Externalities—unpriced by-products of economic transactions that affect outsiders—can be corrected through taxes or subsidies. And public goods—those that are nonrival and nonexcludable—can be provided by government itself.

K E Y T E R M S

tort
market failure
average cost pricing

Averch-Johnson effect
externality
public good

private good
rivalry
excludability

tragedy of the commons
Herfindahl-Hirschman Index

R E V I E W Q U E S T I O N S

1. Explain how each of the following enhances economic efficiency:
 - a. Criminal law
 - b. Property law
 - c. Contract law
 - d. Tort law
 - e. Antitrust law
2. What specific actions are forbidden by antitrust law?
3. How does regulation differ from court decisions in its effect on business?
4. Identify five main types of taxes in the United States.
5. What is a market failure? What are some main causes of market failure?
6. What is a natural monopoly? Give an example.
7. What is an externality? Give one example each of a positive and a negative externality not mentioned in the chapter. What are the effects of positive and negative externalities in the absence of government intervention?
8. What is a pure public good? How is a pure public good different from a private good?
9. What is the free rider problem?

P R O B L E M S A N D E X E R C I S E S

1. In class, one student frequently asks questions and engages the instructor in long discussions,
 - a. Does his behavior involve positive externalities? Negative externalities? Both? Neither? (Be clear about the assumptions you are making to arrive at your answers.)
 - b. Is the result efficient? If not, what kinds of “solutions” can you suggest?
2. When a negative externality creates an inefficiency, the government can sometimes correct the inefficiency by imposing a tax, as shown in Figure 5 (p. 446). An alternative approach is regulation. Suppose the gasoline market in Figure 5 is in equilibrium at point A. The government wishes to correct the externality by imposing an upper limit on the quantity of gasoline that can be produced and sold.

- a. What quantity should it choose as the upper limit?
 - b. Would regulating gasoline in this way correct the inefficiency? Explain.
 - c. Does this way of dealing with negative externality create any special problems compared to imposing a tax?
3. Classify each of the following goods, using your best judgment as to whether it is (a) rival or nonrival, and (b) excludable or nonexcludable.
 - a. Parks in a residential neighborhood
 - b. Military defense
 - c. Food
 - d. Clothing
 - e. Shelter
 - f. Health care
 4. Consider the following data on market shares in an industry.

Firm A	25%
Firm B	20%
Firm C	15%
Firm D	10%
Firm E	9%
Firm F	8%
Firm G	7%
Firm H	6%

 - a. What is the Herfindahl-Hirschman Index of the industry?
 5. Identify every merger of two firms in this industry that the Justice Department will not consider challenging. (*Hint:* There are not many.)
 6. Suppose all mergers you identified in part (b) have occurred. At that point, is there any merger among two firms that the Justice Department will not consider challenging?
 5. Give an example of a public good that is not provided by the government. Give an example of a good that is provided by government but is not a public good.
 6. Last year, Pat and Chris occupied separate apartments. Each consumed 400 gallons of hot water monthly. This year, they are sharing an apartment. To their surprise, they find that they are using a total of 1,000 gallons per month between them. Why?
 7. Many universities subsidize their football teams. That is, if ticket sales are insufficient to cover the cost of the football program, the university makes up the difference. Are there positive externalities that justify the use of a subsidy? If so, what are those externalities, and who are the third parties who benefit from them?
 8. In what sense is a public good like a positive externality?

C H A L L E N G E Q U E S T I O N S

1. Suppose Douglas and Ziffel have properties that adjoin the farm of Mr. Haney. The current zoning law permits Haney to use the farm for any purpose. Haney has decided to raise pigs (the best use of the land). A pig farm will earn \$50,000 per year, forever.
 - a. Assuming the interest rate is 10 percent per year, what is Haney's pig farm worth? (*Hint:* Use the special discounting formula of Chapter 13.)
 - b. Suppose the next-best use of Haney's property is residential, where it could earn \$20,000 per year. What is the minimum one-time payment Haney would accept to agree to restrict his land for residential use forever?
 - c. Suppose Douglas is willing to pay \$200,000 for an end to pig farming on Haney's land, while Ziffel is willing to pay no more than \$150,000. (For some reason, Ziffel does not mind pig farming as much as Douglas does.) If Douglas pays Haney \$200,000 and Ziffel pays Haney \$150,000, is this a Pareto improvement? Who benefits, who loses, and by how much?
 - d. If Douglas pays \$150,000 and Ziffel pays \$150,000, is this move a Pareto improvement? Who benefits, who loses, and by how much?
2. The purely competitive latte industry faces a demand curve given by

$$Q^D = 10 - P$$

or, equivalently,

$$P = 10 - Q^D$$

and the supply curve of lattes is given by

$$Q^S = P, \text{ or } P = Q^S$$

where P is the price per latte in dollars, and where Q^D and Q^S are quantities demanded and supplied, measured in millions of cups.

- a. Graph the original supply curve and demand curve. Determine the equilibrium price and equilibrium quantity.
- b. Now suppose an excise tax of \$2 per latte is imposed. This changes the supply curve to $P = 2 + Q^S$. Graph the new supply curve (after the tax) and the demand curve. Determine the equilibrium price and equilibrium quantity after the tax.
- c. How much tax revenue does the government earn?
- d. Show how a Pareto improvement is possible from the equilibrium you determined in part (b).

EXPERIENTIAL EXERCISES

1. Download Betty Joyce Nash's "Pollution Allowances Help Clear the Air" at <http://www.rich.frb.org/cross/cross134/2.html>. Based on what you've learned in this chapter, evaluate Nash's case for pollution allowances as a way of controlling negative externalities.



2. A good place for finding late-breaking information about antitrust activity is in the Legal Beat column of the *Wall Street Journal*, inside the Marketplace section. Try to find at least one example that mentions a merger and determine the basis for the government's response. Were the Justice Department's guidelines mentioned in the article?



CHAPTER

16

COMPARATIVE ADVANTAGE AND THE GAINS FROM TRADE

CHAPTER OUTLINE

The Logic of Free Trade

The Theory of Comparative Advantage

Opportunity Cost and Comparative Advantage

Specialization and World Production

Gains from International Trade
The Terms of Trade

Turning Potential Gains into Actual Gains

Some Important Provisos

The Sources of Comparative Advantage

Why Some People Object to Free Trade

The Impact of Trade in the Exporting Country

The Impact of Trade in the Importing Country

Attitudes Toward Free Trade:
A Summary

How Free Trade Is Restricted

Tariffs

Quotas

Protectionism

Myths About Free Trade

Sophisticated Arguments for Protection

Using the Theory: Trade Restrictions in the United States

Consumers love bargains. And the rest of the world offers U.S. consumers bargains galore—cars from Japan, computer memory chips from Korea, shoes from China, tomatoes from Mexico, lumber from Canada, and sugar from the Caribbean. But Americans' purchases of foreign-made goods have always been a controversial subject. Should we let these bargain goods into the country? Consumers certainly benefit when we do so. But don't cheap foreign goods threaten the jobs of American workers and the profits of American producers? How do we balance the interests of specific workers and producers on the one hand with the interests of consumers in general? These questions are important not just in the United States, but in every country of the world.

Over the post-World War II period, there has been a worldwide movement toward a policy of *free trade*—the unhindered movement of goods and services across national boundaries. An example of this movement was the creation—in 1995—of a new international body: the World Trade Organization (WTO). The WTO's goal is to help resolve trade disputes among its members, and to reduce obstacles to free trade around the world. And to some extent it has succeeded: Import taxes, import limitations, and all kinds of crafty regulations designed to keep out imports are gradually falling away. By the end of 1999, 135 countries had joined the WTO. And some 30 other countries, including China, Taiwan, Russia, and Vietnam, were eager to join the free-trade group.

But while many barriers have come down, others are being put up. Asian governments have been dragging their feet on allowing U.S. firms to sell telecommunications and financial services there. The United States has renewed its long-standing quota on sugar imports and—in the late 1990s—took serious steps to reduce imports of steel from Russia, Brazil, and Japan. Europeans have restricted the sale of American satellite communications services and American beef. Canada has interfered with the sale of American magazines and television programs within its borders. Poor countries have imposed tariffs on computers, semiconductors, and software exported by rich countries. Rich countries have announced their intention to maintain, at least through the year 2005, existing quotas on textiles and clothing sold by poor countries.

Looking at the contradictory mix of trade policies that exist in the world, we are left to wonder: Is free international trade a good thing that makes us better off, or is

it bad for us and something that should be kept in check? In this chapter, you'll learn to apply the tools of economics to issues surrounding international trade. Most importantly, you'll see how we can extend economic analysis to a global context, in which markets extend across international borders, and the decision makers are firms, households, and government agencies in different nations.

THE LOGIC OF FREE TRADE

Many of us like the idea of being self-reliant. A very few even prefer to live by themselves in a remote region of Alaska or the backcountry of Montana. But consider the defects of self-sufficiency: If you lived all by yourself, you would be poor. You could not *export* or sell to others any part of your own production, nor could you *import* or buy from others anything they have produced. You would be limited to consuming the goods and services that you produced. Undoubtedly, the food, clothing, and housing you would manage to produce by yourself would be small in quantity and poor in quality—nothing like the items you currently enjoy. And there would be many things you could not get at all—electricity, television, cars, airplane trips, or the penicillin that could save your life.

The defects of self-sufficiency explain why most people do not choose it. Rather, people prefer to specialize and trade with each other. In Chapter 2, you learned that specialization and exchange enable us to enjoy greater production and higher living standards than would otherwise be possible.

This principle applies not just to individuals, but also to *groups* of individuals, such as those living within the boundaries that define cities, counties, states, or nations. That is, just as we all benefit when *individuals* specialize and exchange with each other, so, too, we can benefit when *groups* of individuals specialize in producing different goods and services, and exchange them with other *groups*.

Imagine what would happen if the residents of your state switched from a policy of open trading with other states to one of self-sufficiency, refusing to import anything from “foreign states” or to export anything to them. Such an arrangement would be preferable to individual self-sufficiency—at least there would be specialization and trade *within* the state. But the elimination of trading between states would surely result in many sacrifices. Lacking the necessary inputs for their production, for instance, your state might have to do without bananas, cotton, or tires. And the goods that *were* made in your state would likely be produced inefficiently. For example, while residents of Vermont *could* drill for oil, and Texans *could* produce maple syrup, they could do so only at great cost of resources.

Thus, it would make no sense to insist on the economic self-sufficiency of each of the 50 states. And the founders of the United States knew this. They placed prohibitions against tariffs, quotas, and other barriers to interstate commerce right in the U.S. Constitution. The people of Vermont and Texas are vastly better off under free trade among the states than they would be if each state were self-sufficient.

What is true for states is also true for entire nations. The members of the WTO have carried the argument to its ultimate conclusion: National specialization and exchange can expand world living standards through free *international* trade. Such trade involves the movement of goods and services across national boundaries. Goods and services produced domestically, but sold abroad, are called **exports**; those produced abroad, but consumed domestically, are called **imports**. The long-term goal of the WTO is to remove all barriers to exports and imports in order to encourage among nations the specialization and trade that have been so successful within nations.

Exports Goods and services produced domestically, but sold abroad.

Imports Goods and services produced abroad, but consumed domestically.

THE THEORY OF COMPARATIVE ADVANTAGE

Economists who first considered the benefits of international trade focused on a country's *absolute advantage*.

Absolute advantage The ability to produce a good using fewer resources than another country.

A country has an absolute advantage in a good when it can produce it using fewer resources than another country.

As the early economists saw it, the citizens of every nation could improve their economic welfare by specializing in the production of goods in which they had an absolute advantage and exporting them to other countries. In turn, they would import goods from countries that had an absolute advantage in those goods.

Way back in 1817, however, the British economist David Ricardo disagreed. Absolute advantage, he argued, was not a necessary ingredient for mutually beneficial international trade. The key was *comparative advantage*:

Comparative advantage The ability to produce a good at a lower opportunity cost than another country.

A nation has a comparative advantage in producing a good if it can produce it at a lower opportunity cost than some other country.

Notice the difference between the definitions of absolute advantage and comparative advantage. While absolute advantage in a good is based on the resources used to produce it, comparative advantage is based on the *opportunity cost* of producing it. And we measure the opportunity cost of producing a good not by the resources used to produce it, but rather by the amount of *other goods* whose production must be sacrificed.

Ricardo argued that a potential trading partner could be absolutely inferior in the production of every single good—requiring more resources per unit of each good than any other country—and still have a comparative advantage in some good. The comparative advantage would arise because the country was *less* inferior at producing some goods than others. Likewise, a country that had an absolute advantage in producing everything could—contrary to common opinion—still benefit from trade. It would have a comparative advantage only in some—but not all—goods.

Mutually beneficial trade between any two countries is possible whenever one country is relatively better at producing a good than the other country is. Being relatively better means having the ability to produce a good at a lower opportunity cost—that is, at a lower sacrifice of other goods foregone.

OPPORTUNITY COST AND COMPARATIVE ADVANTAGE

To illustrate Ricardo's insight, let's consider a hypothetical world of two countries, China and the United States. Both are producing only two goods, men's suits and computers. Could they better themselves by trading with one another? Ricardo would have us look at opportunity costs. To find them, let's consider what it costs to produce these goods in each country. To keep our example simple, we assume that the costs per unit—for both suits and computers—remain the same no matter how many units are produced.

The relevant cost information is provided in Table 1. Since Chinese firms keep books in Chinese yuan (CNY) and American firms in dollars, our cost data are expressed accordingly. We can use the data in the table to calculate the opportunity cost of producing more of each good in each country.

TABLE 1

COSTS OF PRODUCTION

	Per Suit	Per Computer
China	2,000 CNY	10,000 CNY
United States	\$500	\$1,000

First, suppose China were to produce one additional computer. Then it would have to divert 10,000 yuan's worth of resources from the suit industry. This, in turn, would require China to produce fewer suits. How many fewer? Since each suit uses up 2,000 CNY in resources, using 10,000 CNY for one computer would require producing $10,000/2,000 = 5$ fewer suits. Thus, the opportunity cost of a computer in China is 5 suits. This opportunity cost is recorded in Table 2; check the table and make sure you can find this entry. In the United States, producing an additional computer requires diverting \$1,000 of resources from suit making. Since each suit costs \$500, this means a loss of 2 suits. Thus, in the United States, the opportunity cost of one computer is 2 suits—which can also be found in Table 2.

Summing up, we see that in China, the opportunity cost of a computer is 5 suits; in the United States, it is 2 suits. Therefore, the United States—with the lower opportunity cost of producing computers—*has a comparative advantage in making computers*.

Notice that in Table 2, we do similar calculations for the opportunity cost of making a suit, measuring the opportunity cost in terms of *computers foregone*. These computations are summarized in the first column of the table. Make sure you can use these numbers to verify that China has a comparative advantage in producing suits.

Now we can use our conclusions about comparative advantage to show how both countries can gain from trade. The explanation comes in two steps. First, we show that if China could be persuaded to produce more suits and the United States more computers, the world's total production of goods will increase. Second, we show how each country can come out ahead by trading with the other.

SPECIALIZATION AND WORLD PRODUCTION

Using the numbers in Table 2, if China produced, say, 10 more suits, it would have to sacrifice the production of 2 computers as resources were shifted between the two industries. If the United States, simultaneously, produced 4 extra computers, it would have to sacrifice 8 suits—again because fully employed resources would have to be moved. But note: As a result of even this small change, the world's production of suits would increase by 2, and its production of computers would also rise by 2—despite the fact that no more resources were used than before. Table 3 summarizes the changes.

TABLE 2

OPPORTUNITY COSTS

	Per Suit	Per Computer
China	$\frac{1}{5}$ computer	5 suits
United States	$\frac{1}{2}$ computer	2 suits

TABLE 3

A SMALL CHANGE IN PRODUCTION

	Suit Production	Computer Production
China	+10	-2
United States	-8	+4
World	+2	+2

The additional production of suits and computers in this example represents the gain from specializing according to comparative advantage—a gain, as the next section will show, that the two trading partners will share. It is also the kind of gain that multiplied a million times, lies behind the substantial benefits countries enjoy from free trade.

The particular example given here is not the only one that can be derived from our table of opportunity costs. For example, if China produced 20 more suits and, therefore, produced 4 fewer computers, while the United States changed as in Table 3, then world output of suits would increase by 12, while computer production would remain unchanged. And we could come up with other examples in which the world output of computers rises, but suits remain the same. (As an exercise, try to create such an example on your own.)

In all cases, however, the key insight remains the same:

If countries specialize according to comparative advantage, a more efficient use of given resources occurs. That is, with the same resources, the world can produce more of at least one good, without decreasing production of any other good.

GAINS FROM INTERNATIONAL TRADE

Now we proceed to the second step in Ricardo's case, showing that both sides can gain from trade. Let's first note that, if two countries change their production as in Table 3, but do *not* trade with each other, each country would have more of one good but *less* of another. However, because of the increase in *world* output, international trade flows could be arranged so that each country would share in the gain in total output. Many different arrangements are possible; here is one that would apportion the world output gain equally:

China exports (and the United States imports) 9 suits.

China imports (and the United States exports) 3 computers.

Table 4 summarizes the end result. The second column in the table shows the changes in *production* in each country based on the information in Table 3. The third column shows how much of each good is exported or imported. Finally, the last column shows how much more of each good the citizens of each country end up with. In our example, China and the United States each end up with 1 additional suit and 1 additional computer. Notice that if we add up these gains from trade (a total of 2 suits and 2 computers), they are precisely equal to the gains in world output noted earlier, in Table 3. This is no coincidence: With only two countries in our example, when world output of a good rises, one country or the other must end up consuming it.

TABLE 4

	Production	Loss from Exports (-) or Gain from Imports (+)	Net Gain
China			
Suits	+10	-9	+1
Computers	-2	+3	+1
United States			
Suits	-8	+9	+1
Computers	+4	-3	+1

**THE GAINS FROM
SPECIALIZATION
AND TRADE**

It is worth reiterating that the mutually beneficial changes summarized in Table 4 are based on *comparative* advantage, not *absolute* advantage. To make this point even clearer, let's look at the information in Table 1 from another perspective. Instead of thinking about the *cost* of producing a good, we'll look directly at the resources used up in making it. To keep things simple, we'll suppose that the only resource countries use in production is labor. Further, we'll suppose arbitrarily that an hour of labor costs 16 CNY in China and \$10 in the United States. Then the 2,000 CNY it costs to make a suit in China would mean that 125 hours of labor are needed to make a suit there, since 125 hours \times 16 CNY per hour = 2,000 CNY. Thus, in Table 5, we enter 125 hours for the labor needed per suit in China.

Making similar calculations, we find that it takes 625 hours to make a computer in China; and in the United States, it takes 50 hours to make a suit and 100 hours to make a computer.

Now it's easy to see that the United States has an absolute advantage—using less input per unit of output than China—in the production of *both* goods. Would specialization and mutually beneficial trade still be possible? Very much so. The opportunity cost data in Table 2 still apply (verify this on your own), and so do all the conclusions we derived in Tables 3 and 4, which were based on the information in Table 2. Thus,

as long as opportunity costs differ, specialization and trade can be beneficial to all involved. This remains true regardless of whether the parties involved are nations, states, counties, or individuals. It remains true even if one party has an all-round absolute advantage or disadvantage.

THE TERMS OF TRADE

In our ongoing example, China exports 9 suits in exchange for 3 computers. This exchange ratio (9 suits for 3 computers, or 3 suits per computer) is known as the

TABLE 5

	Per Suit	Per Computer
China	125 hours	625 hours
United States	50 hours	100 hours

LABOR INPUTS NEEDED

Terms of trade The ratio at which a country can trade domestically produced products for foreign-produced products.



If China has a comparative advantage in the production of men's suits, it can gain by exporting them to other countries.



The World Trade Organization's Web page (<http://www.wto.org/>) is a good source for all kinds of information on international trade.

Exchange rate The amount of one currency that is traded for one unit of another currency.

terms of trade. Our particular choice of 3 to 1 for the terms of trade happened to apportion the gain in world output equally between the two countries. (See Table 4.) With different terms of trade, however, the benefit would have been distributed unequally. We won't consider here precisely *how* the terms of trade are determined (it is a matter of supply and demand), but we can establish the limits within which they must fall.

Look again at Table 2 (p. 467). China would never give up *more* than 5 suits to import 1 computer. Why not? Because it could always get 1 computer for 5 suits *domestically*, by shifting resources into computer production.

Similarly, the United States would never export a computer for *fewer than 2* suits, since it can substitute 1 computer for 2 suits domestically (again, by switching resources between the industries). Therefore, the equilibrium terms of trade must lie *between* 5 suits for 1 computer and 2 suits for 1 computer. Outside of that range, one of the two countries would refuse to trade. Note that in our example, we assume terms of trade of 3 suits for 1 computer—well within the acceptable range.

TURNING POTENTIAL GAINS INTO ACTUAL GAINS

So far in this chapter, we have discussed the *potential* advantages of specialization and trade among nations, but one major question remains: How is that potential realized? Who or what causes a country to shift resources from some industries into others and then to trade in the world market?

Do foreign trade ministers at WTO meetings decide who should produce and trade each product? Does some group of omniscient and benevolent people in Washington and other world capitals make all the necessary arrangements? Not at all. Within the framework of the WTO, government officials are supposed to create the environment for free trade, but they do not decide who has a comparative advantage in what, or what should be produced in this or that country. In today's market economies around the world, it is individual consumers and firms who decide to buy things—at home or abroad. By their joint actions, they determine where things are produced and who trades with whom. That is, the promise of Ricardo's theory is achieved through markets. People only have to do what comes naturally: buy products at the lowest price. Without their knowing it, they are promoting Ricardo's dream!

Let's see how this works. In the absence of trade, the prices of goods within a country will generally reflect their domestic opportunity costs. That is, if producing one more computer in the United States requires the sacrifice of 2 suits, then the price of a computer should be about twice the price of a suit.

Let's imagine the situation before trade between two countries begins. We'll suppose that prices in each country are precisely equal to the costs of production in each country, as given earlier in Table 2. These prices are shown again in Table 6, in bold type. For the moment, ignore the prices in parentheses.

Now suppose we allow trade to open up between the two countries. Consider the decision of a U.S. consumer, who can choose to purchase suits and computers in either country. To buy goods from Chinese producers, Americans must pay in yuan. In order to obtain that currency, Americans must go to the *foreign exchange market* and trade their dollars for yuan at the going **exchange rate**—the rate at which one currency can be exchanged for another. Let's assume that the exchange rate is 8 yuan per dollar.

Now, at this exchange rate, an American can purchase a suit made in China priced at 2,000 CNY by exchanging \$250 for 2,000 CNY and then buying the suit. Thus, to the American, the *dollar price of a Chinese suit* is \$250, which appears in

TABLE 6

PRICES IN CHINA AND THE UNITED STATES WITH AN EXCHANGE RATE OF 8 CNY FOR \$1

	Per Suit	Per Computer
China	2,000 CNY (\$250)	10,000 CNY (\$1,250)
United States	\$500 (4,000 CNY)	\$1,000 (8,000 CNY)

parentheses below the price in yuan. Similarly, the dollar price of a 10,000 CNY Chinese computer is \$1,250—also in parentheses.

Looking at Table 6, you can see that, to an American, suits from China at \$250 are cheaper than U.S. suits at \$500, so *Americans will prefer to buy suits from China*. But when it comes to computers, we reach the opposite conclusion: A U.S. computer at \$1,000 is cheaper than a Chinese computer at \$1,250, so *Americans will prefer to buy computers in the United States*.

Now take the viewpoint of a Chinese consumer who can buy U.S. or Chinese goods. To buy U.S. goods, China's consumers will need dollars, which they can obtain at the going exchange rate: 8 CNY for \$1. The bottom row of the table (with figures in parentheses) shows the prices of U.S. goods in yuan. To a Chinese buyer, Chinese suits at 2,000 CNY are cheaper than U.S. suits at 4,000 CNY, while U.S. computers at 8,000 CNY are cheaper than Chinese computers at 10,000 CNY. Thus, *a Chinese, just like an American, will prefer to buy computers from the United States and suits from China*.

Now suppose that trade in suits and computers had previously been prohibited, but is now opened up. Everyone would buy suits in China and computers in the United States, and the process of specialization according to comparative advantage would begin. Chinese suit makers would expand their production, while Chinese computer makers would suffer losses, lay off workers, and even exit the industry. Unemployed computer workers in China would find jobs in the suit industry. Analogous changes would occur in the United States, as production of computers expanded there. These changes in production patterns would continue until China specialized in suit production and the United States specialized in computer production—that is, until each country produced according to its comparative advantage.

Our example illustrates a general conclusion:

When consumers are free to buy at the lowest prices, they will naturally buy a good from the country that has a comparative advantage in producing it. That country's industries respond by producing more of that good and less of other goods. In this way, countries naturally tend to specialize in those goods in which they have a comparative advantage.¹

¹ Something may be bothering you about the way we reached this conclusion: We merely *asserted* that the exchange rate was 8 yuan per dollar. What if we had chosen another exchange rate? With a little work, you can verify that at any exchange rate between 4 yuan per dollar and 10 yuan per dollar, our conclusion will still hold: Countries will automatically produce according to their comparative advantage. Further, you can verify that if the exchange rate went *beyond* those bounds, the residents of both countries would want to buy both goods from just one country. This would change the demand for yuan and force the exchange rate back between 8 yuan per dollar and 4 yuan per dollar.

SOME IMPORTANT PROVISOS

Look back at Tables 3 and 4 (pp. 468 and 469). There you saw how a small change in production—with China shifting toward suits and the United States shifting toward computers—caused world production of both goods to rise. But if this can happen once, why not again? And again? And again? In fact, our simple example seems to suggest that countries should specialize *completely*, producing *only* the goods in which they have a comparative advantage. In our example, it seems that China should get out of computer production *entirely*, and the United States should get out of suit production *entirely*.

The real world, however, is more complicated than our simplified examples might suggest. Despite divergent opportunity costs, sometimes it does *not* make sense for two countries to trade with each other, or it might make sense to trade, but *not* completely specialize. Following are some real-world considerations that can lead to reduced trade or incomplete specialization.

Costs of Trading. If there are high transportation costs or high costs of making deals across national boundaries, trade may be reduced and even become prohibitively expensive. High transportation costs are especially important for perishable goods, such as ice cream, which must be shipped frozen, and most personal services, such as haircuts, eye exams, and restaurant meals. None of these are typically traded internationally. (Imagine the travel cost for an American hair stylist who would like to sell a haircut to a resident of China.)

The costs of making deals are generally higher for international trade than for trade within domestic borders. For one thing, different laws must be dealt with. In addition, there are different business and marketing customs to be mastered. High transportation costs and high costs of making deals help explain why nations continue to produce some goods in which they do not have a comparative advantage and why there is less than complete specialization in the world.

One final cost of international trade arises from the need to exchange domestic for foreign currency. In international trade, either importers or exporters typically take some risk that the exchange rate might change. For example, suppose a U.S. importer of suits from China agrees in advance to pay 100,000 CNY for a shipment of suits. At the time the agreement is made, the exchange rate is 8 CNY per dollar, so the importer figures the shipment will cost him \$12,500. But suppose that, before he pays, the exchange rate changes to 5 CNY per dollar. Then the suit shipment—for which the importer must still pay 100,000 CNY—will cost him \$20,000. The rise in costs could wipe out the importer's profit, or even cause him to lose money on the shipment.

It is interesting to note that countries can work to reduce the cost of trading. Indeed, this was the primary reason behind the creation of a new, single currency—the *euro*—to be shared by 11 European countries, including France, Germany, Holland, and Italy. The euro was introduced into commerce in early 1999. By July 2002, the separate national currencies of the “Euroland” countries will be phased out of existence, and the French franc, the Italian lira, the German mark, and several other national currencies will become relics of the past. The move to a single currency will eliminate all the costs and risks of foreign exchange transactions from intra-European trade. This should enable these European countries to specialize more completely according to their comparative advantage, and increase the gains from trade even further.

Sizes of Countries. Our earlier example featured two large economies capable of fully satisfying each other's demands. But sometimes a very large country, such as the United States, trades with a very small one, such as the Pacific island nation of

Tonga. If the smaller country specialized completely, its output would be insufficient to fully meet the demand of the larger one. The larger country would continue to produce both goods and would specialize only in the sense of producing *more* of its comparative-advantage good rather than *nothing but* that good. The smaller country would specialize completely. This helps to explain why the United States continues to produce bananas, even though we do so at a much higher opportunity cost than many small Latin American nations.

Increasing Opportunity Cost. In all of our tables, we have assumed that opportunity cost remains constant as production changes. For example, in Table 2, the opportunity cost of a suit in China is $\frac{1}{3}$ of a computer—regardless of how many suits or computers China makes. But more typically, the opportunity cost of a good rises as more of it is produced. (Why? You may want to review the law of increasing opportunity cost in Chapter 2.) In that case, each step on the road to specialization would change the opportunity cost. A point might be reached—before complete specialization—in which opportunity costs became *equal* in the two countries, and there would be no further mutual gains from trading. (Remember: Opportunity costs must *differ* between the two countries in order for trade to be mutually beneficial.) In the end, while trading will occur, there will not be complete specialization. Instead, each country will produce both goods, just as China and the United States each produce suits *and* computers in the real world.

Government Barriers to Trade. Governments can enact barriers to trading. In some cases, these barriers increase trading costs; in other cases, they make trade impossible. Since this is such an important topic, we'll consider government-imposed barriers to trade in a separate section, later in the chapter.

THE SOURCES OF COMPARATIVE ADVANTAGE

We've just seen how nations can benefit from specialization and trade when they have comparative advantages. But what determines comparative advantage in the first place? In many cases, the answer is differences in natural resources. The top part of Table 7 contains some examples. Saudi Arabia has a comparative advantage in the production of oil because it has oil fields with billions of barrels of oil that can be extracted at low cost. Canada is a major exporter of timber because its climate and geography make its land more suitable for growing trees than other crops. Canada is a good example of comparative advantage without absolute advantage—it grows a lot of timber, not because it can do so using fewer resources than other countries, but because its land is even more poorly suited to growing other crops.

The bottom part of Table 7 shows examples of international specialization in which comparative advantage arises from some cause *other* than natural resources. Japan has a huge comparative advantage in making automobiles—over 40 percent of the world's automobiles are made there. And that number would be even larger, except for laws that limit the import of Japanese cars into Europe. Yet none of the natural resources needed to make cars are available in Japan; the iron ore, coal, and oil that provide the basic ingredients for cars are all imported.

Countries often specialize in products based on their own particular endowments of natural resources. But natural resources are not the only basis for comparative advantage.

TABLE 7

**EXAMPLES OF NATIONAL
SPECIALTIES IN
INTERNATIONAL TRADE**

Country	Specialty Resulting from Natural Resources or Climate
Saudi Arabia	Oil
Canada	Timber
United States	Grain
Spain	Olive oil
Mexico	Tomatoes
Jamaica	Aluminum ore
Italy	Wine
Israel	Citrus fruit
Specialty <i>Not</i> Based on Natural Resources or Climate	
Japan	Cars, consumer electronics
United States	Software, movies, music
Switzerland	Watches
Korea	Steel, ships
Hong Kong	Textiles
Great Britain	Financial services

Explaining the origins of the specialties in the bottom part of Table 7 is not easy. For example, if you think you know why Japan completely dominates the world market for VCRs and other consumer electronics—say, some unique capacity to mass-produce precision products—be sure you have an explanation for why Japan is a distant second in computer printers. The company that dominates the market for printers—Hewlett-Packard—is a U.S. firm. Moreover, the ability to mass-produce high-quality products is not unique to Japan, as Switzerland showed long ago in developing its international specialty in watches.

In even the most remote corner of the world, the cars, cameras, and VCRs will be Japanese, the movies and music American, the clothing from Hong Kong or China, and the bankers from Britain. Although we can't explain the reasons behind these countries' comparative advantages, we *can* explain why a country retains its comparative advantage once it gets started. Japan today enjoys a huge comparative advantage in cars and consumer electronics in large part because it has accumulated a capital stock—both physical capital and human capital—well suited to producing those goods. The physical capital stock includes the many manufacturing plants and design facilities that the Japanese have built over the years. But Japan's human capital is no less important. Japanese managers know how to anticipate the features that tomorrow's buyers of cars and electronic products will want around the world. And Japanese workers have developed skills adapted for producing these products. The stocks of physical and human capital in Japan sustain its comparative advantage in much the way as stocks of natural resources lead to comparative advantages in other countries. More likely than not, Japan will continue to have a comparative advantage in cars and electronics, just as the United States will continue to have a comparative advantage in making movies.



The International Trade Administration maintains a Web page that is full of information on U.S. international trade. Find it at <http://www.ita.doc.gov/>.

Countries often develop strong comparative advantages in the goods they have produced in the past, regardless of why they began producing those goods in the first place.

WHY SOME PEOPLE OBJECT TO FREE TRADE

Given the clear benefits that nations can derive by specializing and trading, why would anyone *ever* object to free international trade? Why do the same governments that join the WTO turn around and create roadblocks to unhindered trade? The answer is not too difficult to find: Despite the benefit to the nation as a whole, some groups within the country, in the short run, are likely to lose from free trade, even while others gain a great deal more. Unfortunately, instead of finding ways to compensate the losers—to make them better off as well—we often allow them to block free-trade policies. The simple model of supply and demand helps illustrate this story.

In our earlier example, after trade opens up, China exports suits and the United States imports them. Figure 1 illustrates the impact on the market for suits in the two countries. To keep things simple, we'll convert the price of suits in China into dollars, so that we can measure dollar prices on the vertical axis of both panels.

Before trade opens up, the Chinese suit market is in equilibrium at point *E*, with price equal to P_N (for “no trade”) and quantity equal to Q_N . The U.S. suit market is in equilibrium at point *F* with price P'_N and quantity Q'_N . Notice that before trade opens up, the price is lower in China—the country with a comparative advantage in suits.

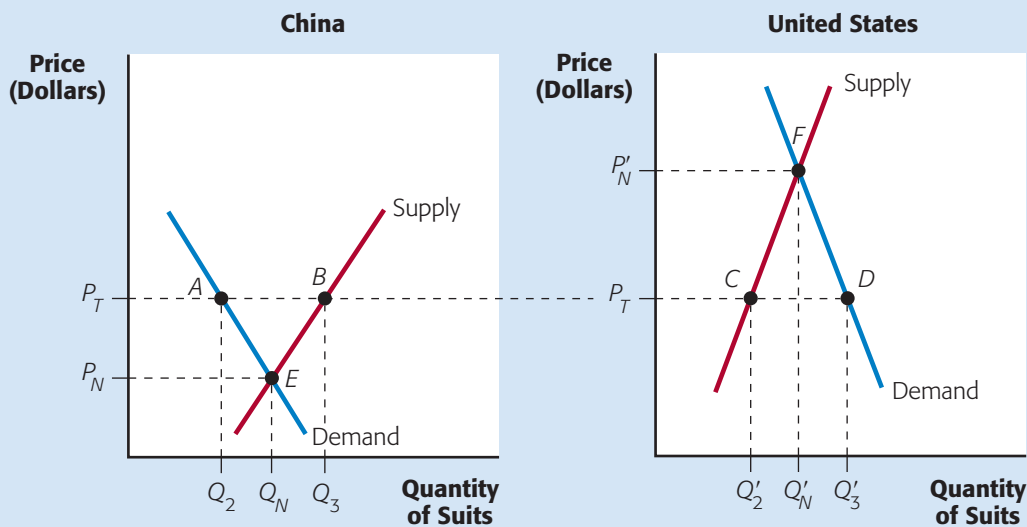
Now, when trade opens up, Americans will begin to buy Chinese suits, driving their price upward. As the price in China rises from P_N to P_T (for “trade”), Chinese producers increase their output, moving from *E* to *B* along the supply curve, and Chinese consumers decrease their purchases, moving from *E* to *A* along the demand curve

Find the Equilibrium

What Happens When Things Change?

THE IMPACT OF TRADE

FIGURE 1



Before trade, the Chinese suit market is in equilibrium at point *E*, and the U.S. market is in equilibrium at point *F*. When trade begins, Americans buy the cheaper Chinese suits, driving up their price. In response, Chinese manufacturers increase output, and Chinese consumers decrease their purchases. At the world equilibrium price P_T , the Chinese buy Q_2 suits, Americans buy $Q_3 - Q_2$ Chinese suits, and the total quantity of Chinese suits produced and sold is Q_3 . Distance *CD*, which shows U.S. imports of suits, equals distance *AB*, which shows Chinese exports.

As a result of trade, Chinese suit makers sell more units at a higher price, but Chinese consumers pay more for their suits. In the United States, suit makers are worse off, but suit buyers benefit from the lower price.

curve. This seems to create an “excess supply” of suits in China, equal to AB , but it is not *really* an excess supply, because AB is precisely the number of suits that are exported to the United States. That is, the entire output of suits— Q_3 —is purchased by either Chinese or Americans.

Now let’s consider the effects in the United States. There, consumers are switching from suits made in the United States to suits made in China. With less demand for United States suits, their price will fall. With free trade, the United States must be able to buy Chinese suits at the same price as the Chinese (ignoring transportation costs), so the price of suits in the United States must fall to P_T . As the price falls, U.S. suit producers will decrease their output, from F to C along the supply curve, and U.S. consumers will increase their purchases, from F to D along the demand curve. This seems to create a shortage of suits in the United States, equal to CD , but it is not a shortage: CD is precisely the number of suits imported from China.

Now let’s see how different groups are affected by the opening up of trade.

THE IMPACT OF TRADE IN THE EXPORTING COUNTRY

When trade opens up in suits, China is the exporting country. How are different groups affected there?

- *Chinese suit producers and workers are better off.* Before international trade, producers sold Q_N suits at price P_N , but with trade, they sell a larger quantity Q_3 at a higher price P_T . The industry’s workers are equally delighted because they undoubtedly share in the bonanza as the number of workers demanded rises along with the level of production. Both management and labor in the Chinese suit industry are likely to favor free trade.
- *Chinese suit buyers are worse off.* Why? Before trade, they bought Q_N suits at price P_N , and now they must pay the higher price, P_T , and consume the smaller quantity Q_2 . If the harm is great enough, consumers may band together and lobby the government to restrict free trade:

When the opening of trade results in increased exports of a good, the producers of the good are made better off and will support increased trade. Consumers of the good will be made worse off and will oppose increased trade.

The story told here is anything but hypothetical. A dramatic example is provided by American agriculture, which for decades exported a large percentage of various crops to the former Soviet Union. Growers of wheat, rye, and corn did everything they could to promote this trade. All kinds of people in grain-growing areas, ranging from car dealers to sellers of fertilizer, were equally behind the Russian trade deal; they benefited indirectly. American consumers, however, complained bitterly. Bread, cereals, and flour were more expensive. So were eggs and chicken, because chickens were fed with more expensive grain.

THE IMPACT OF TRADE IN THE IMPORTING COUNTRY

Now let’s consider the impact of free trade in suits on the United States, the importing country. Once again, it is easy to figure out who is happy and who is unhappy with the new arrangement.

- *U.S. suit producers and workers are worse off.* They formerly sold quantity Q'_N at price P'_N , but now they are furious because they sell the lower quantity Q'_2 at the lower price P_T . The industry’s workers suffer, too, because the number of

TABLE 8

ATTITUDES TOWARD FREE TRADE

	In Export Sectors That Enjoy Comparative Advantage	In Import Sectors That Suffer from Comparative Disadvantage
Pro Trade	Owners of firms, workers	Consumers
Anti Trade	Consumers	Owners of firms, workers

workers demanded falls with the level of production. Both management and labor are likely to oppose free trade.

- *U.S. suit buyers are better off.* They used to buy quantity Q'_N at price P'_N , but now they pay the lower price, P_T , and consume the larger quantity, Q'_3 . U.S. consumers will favor free trade:

When the opening of trade results in increased imports of a product, the domestic producers of the product are made worse off and will oppose the increased trade. Consumers are better off and will favor the increased trade.

This story, too, is anything but hypothetical, as an example from the mid-1990s illustrates. A Ukrainian clothing maker produced stylish, high-quality women's coats and sold them in the United States. With the coats priced between \$89 and \$139, over a million of them were sold. When American coat makers complained bitterly about the new competitor, the U.S. government stepped in. A tight import limitation killed off half of the Ukrainian imports in 1995. On top of that, the United States imposed a 21.5-percent tax on the offending coats. The interests of U.S. coat makers prevailed over the interests of U.S. coat consumers.

ATTITUDES TOWARD FREE TRADE: A SUMMARY

In our examples, we've been discussing the impact of free trade in suits. We could tell the same story about free trade in computers. In this case, the United States has the role of exporter, and China is the importer. But our conclusions about the impacts on different groups in exporting and importing countries would remain the same. And so would our conclusions about who favors, and who opposes, free trade. Table 8 summarizes the stance toward trade we can expect from these different groups.

HOW FREE TRADE IS RESTRICTED

So far in this chapter, you've learned that specialization and trade according to comparative advantage can dramatically improve the well-being of entire nations. This is why governments generally favor free trade. Yet international trade can, in the short run, hurt particular groups of people. These groups often lobby their government to restrict free trade.

When governments decide to accommodate the opponents of free trade, they are apt to use one of two devices to restrict trade: tariffs or quotas.

Tariff A tax on imports.

TARIFFS

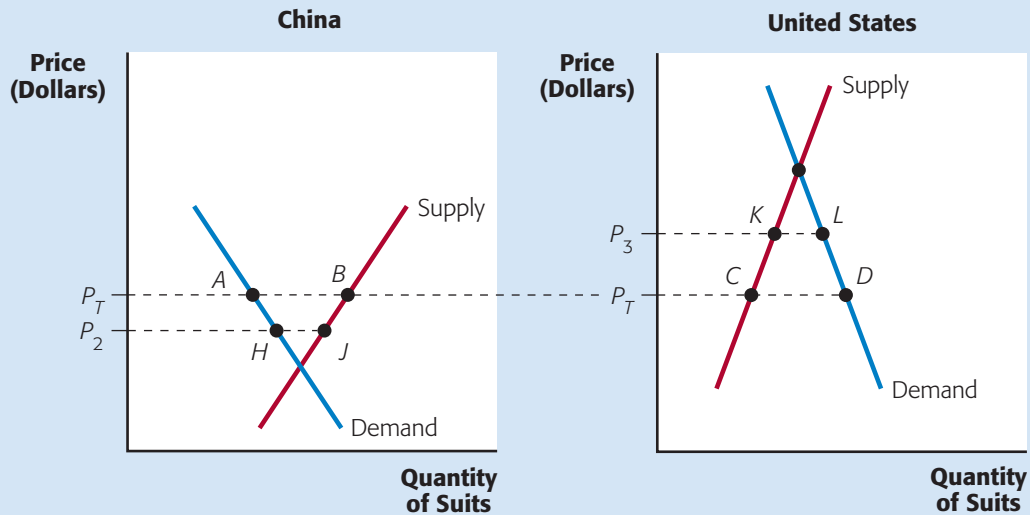
A **tariff** is a tax on imported goods. It can be a fixed dollar amount per physical unit, or it can be a percentage of the good's value. In either case, the effect in the tariff-imposing country is similar.



What Happens When Things Change?

FIGURE 2

THE EFFECTS OF A TARIFF ON SUITS



A U.S. tariff on imports of Chinese suits raises their price in the U.S. from P_T to P_3 . As a result of the price increase, U.S. imports fall to KL , which equals Chinese exports of HJ . With fewer suits produced, the price in China falls to P_2 .

Figure 2 illustrates the effects of a U.S. government tariff on Chinese suits. Initially, before the tariff is imposed, the price of suits in both countries is P_T , and China exports AB of them, while the U.S. imports the same number (represented by the distance CD in the U.S. market). Now, suppose the United States imposes a tariff on Chinese suits. Since it is more costly for Chinese suit makers to sell suits in the United States than before, they will shift some of their output back to the home market in China. In the United States, at the old price, P_T , this decrease in the supply of suits *would* create a shortage, but—as we know—shortages force the price up. In our diagram, the United States price rises to P_3 . As the price rises, the quantity of suits supplied domestically increases, and the quantity demanded domestically decreases. U.S. imports are accordingly cut back to KL . In China, the sale of suits formerly exported drives the price there down to P_2 . Notice that—in the final equilibrium with U.S. price equal to P_3 and the price in China equal to P_2 —U.S. imports (KL) and Chinese exports (HJ) are equal. That is, every suit Chinese suit that is *not* bought by a Chinese consumer is bought by an American consumer. As you can see, American consumers are worse off—they pay a higher price for fewer suits. U.S. producers, on the other hand, are much better off: They sell more suits at a higher price. In China, the impact is the opposite: The price of suits falls, so Chinese producers lose and Chinese consumers gain.

But we also know this: Since the volume of trade has decreased, the gains from trade according to comparative advantage have been reduced as well. Both countries, as a whole, are worse off as a result of the tariff:

Tariffs reduce the volume of trade and raise the domestic prices of imported goods. In the country that imposes the tariff, producers gain and consumers lose. But the world as a whole loses, because tariffs decrease the volume of trade and therefore decrease the gains from trade.

QUOTAS

A **quota** is a government decree that limits the imports of a good to a specified maximum physical quantity, such as 500,000 Ukrainian coats per year. Because the goal is to restrict imports, a quota is set below the level of imports that would occur under free trade. Its general effects are precisely the same as those of a tariff.

Figure 2, which we used to illustrate tariffs, can also be used to analyze the impact of a quota. In this case, we suppose that the U.S. government simply decrees that it will only allow KL suits into the country and that it is able to enforce this quota. Once again, the market price in the United States will rise to P_3 . (Why? Because at any price lower than P_3 , total imports of KL plus the domestic quantity supplied, given by the supply curve, would be smaller than quantity demanded. This would cause the price to rise.) And once again, the decrease in U.S. imports translates into a shrinkage in Chinese exports—down to HJ . Both countries' suit markets end up in exactly the same place as if the United States had imposed a tariff that raised the U.S. price to P_3 .

The previous discussion seems to suggest that tariffs and quotas are pretty much the same. But even though prices in the two countries may end up at the same level with a tariff or a quota, there is one important difference between these two trade-restricting policies. A tariff, after all, is a *tax* on imported goods. Therefore, when a government imposes a tariff, it collects some revenue every time a good is imported. (See if you can determine the amount of tariff revenue in Figure 2.) This revenue can be used to fund government programs or reduce other taxes, to the benefit of the country as a whole. When a government imposes a quota, however, it gains no revenue at all.

Quotas have effects similar to tariffs—they reduce the quantity of imports and raise domestic prices. While both measures help domestic producers, they reduce the benefits of trade to the nation as a whole. However, a tariff has one saving grace: increased government revenue.

Economists, who generally oppose measures such as quotas and tariffs to restrict trade, argue that, if one of these devices must be used, tariffs are the better choice. While both policies reduce the gains that countries can enjoy from specializing and trading with each other, the tariff provides some compensation in the form of additional government revenue.

PROTECTIONISM

This chapter has outlined the *gains* that arise from international trade, but it has also outlined some of the *pain* trade can cause to different groups within a country. While the country as a whole benefits, some citizens in both the exporting and importing countries are harmed. The groups who suffer from trade with other nations have developed a number of arguments against free trade. Together, these arguments form a position known as **protectionism**—the belief that a nation's industries should be *protected* from free trade with other nations. Some protectionist arguments are rather sophisticated and require careful consideration. We'll consider some of these a bit later. But anti-trade groups have also promulgated a number of myths to support their protectionist beliefs. Let's consider some of these myths.



What Happens When Things Change?

Quota A limit on the physical volume of imports.

Protectionism The belief that a nation's industries should be protected from foreign competition.

MYTHS ABOUT FREE TRADE

“A HIGH-WAGE COUNTRY CANNOT AFFORD FREE TRADE WITH A LOW-WAGE COUNTRY. THE HIGH-WAGE COUNTRY WILL EITHER BE UNDERSOLD IN EVERYTHING AND LOSE ALL OF ITS INDUSTRIES, OR ELSE ITS WORKERS WILL HAVE TO ACCEPT EQUALLY LOW WAGES AND EQUALLY LOW LIVING STANDARDS.”

It's true that some countries have much higher wages than others. Here are 1997 figures for average hourly wages, including benefits such as holiday pay and health insurance: Germany \$28.28; Japan \$19.37; United States \$18.24; Korea \$7.22, Mexico \$1.75; and less than a dollar in Russia, China, and India. As you can see, the wealthier, more-developed countries have wages far higher than poorer, less-developed countries. (The United States–China wage differential, for example, is in reality about 20 to 1—much higher than the 5 to 1 differential we used in our tables.) This leads to the fear that the poorer countries will be able to charge lower prices for their goods, putting American workers out of jobs unless they, too, agree to work for low wages.

But this argument is incorrect, for two reasons. First, it is true that American workers are paid more than Chinese workers, but this is because the average American worker is more *productive* than his Chinese counterpart. After all, the American workforce is more highly educated, and American firms provide their workers with more sophisticated machinery than do Chinese firms. If an American could produce 80 times as much output as a Chinese worker in an hour, then even though wages in the United States may be about 50 times greater, cost *per unit* produced would still be lower in the United States. This is reflected in our example in Tables 5 (p. 469) and 6 (p. 471). If you look closely, you'll see that even though American workers are paid more than their Chinese counterparts, they can produce a computer with so much less labor input that labor costs per computer are actually lower in the United States.

But suppose the cost per unit *were* lower in China. Then there is still another, more basic argument against the fear of a general job loss or falling wages in the United States: comparative advantage. Let's take an extreme case. Suppose that labor productivity were the same in the United States and China, so that China—with lower wages—could produce *everything* more cheaply than the United States could. Both countries would still gain if China specialized in products in which its cost advantage was relatively large and the United States specialized in goods in which China's cost advantage was relatively small. That is, even though China would have an absolute advantage in everything, the United States would still have a comparative advantage in some things. The mutual gains from trade arise not from absolute advantage, but from comparative advantage.

“A LOW-PRODUCTIVITY COUNTRY CANNOT AFFORD FREE TRADE WITH A HIGH-PRODUCTIVITY COUNTRY. THE FORMER WILL BE CLOBBERED BY THE LATTER AND LOSE ALL OF ITS INDUSTRIES.”

This argument is the flip side of the first myth. Here, it is the poorer, less-developed country that is supposedly harmed by trade with a richer country. But this myth, like the first one, confuses absolute advantage with comparative advantage. Suppose the high-productivity country (say, the United States) could produce *every* good with fewer resources than the low-productivity country (say, China). Once again, the low-productivity country would *still* have a comparative advantage in *some* goods. It would then gain by producing those goods and trading with the high-productivity country. This is the case in our example, where a glance at

Table 5 reminds us that that the United States has an absolute advantage in both goods, yet—as we’ve seen—trade still benefits both countries.

To make the point even clearer, let’s bring it closer to home. Suppose there is a small, poor town in the United States where workers are relatively uneducated and work with little capital equipment, so their productivity is very low. Would the residents of this town be better off sealing their borders and not trading with the rest of the United States, which has higher productivity? Before you answer, think what this would mean: The residents of the poor town would have to produce everything on their own: grow their own food, make their own cars and television sets, and even provide their own entertainment. Clearly, they would be worse off in isolation. And what is true *within* a country is also true *between* different countries: Closing off trade will make a nation, as a whole, worse off, regardless of its level of wages or productivity. Even a low-productivity country is made better off by trading with other nations.

“IN RECENT TIMES, AMERICA’S UNSKILLED WORKERS HAVE SUFFERED BECAUSE OF EVER-EXPANDING TRADE BETWEEN THE UNITED STATES AND OTHER COUNTRIES.”

True enough, unskilled workers lost ground from 1980 to the early 1990s, for *some* reason. College graduates have enjoyed growing purchasing power from their earnings, while those with only a grade school education have lost about 25 percent of their 1980 purchasing power. Rising trade with low-wage countries has been blamed for this adverse trend.

But before we jump to conclusions, let’s take a closer look. Our discussion earlier in this chapter tells us where to look for effects that come through trade. If the opening of trade has harmed low-skilled workers in the United States, it would have done so by lowering the prices of products that employ large numbers of those workers. For example, if the United States has been flooded recently with cheap clothes, then we should see a relative decline in U.S. clothing prices and reductions in earnings among clothing workers, who are mostly unskilled. A recent study taking this approach found almost no change in the relative prices of products in this country that employ large numbers of unskilled workers. Studies that take other approaches have found only modest effects. In general, economists who have looked at the relation between changes in trade patterns and the depressed earnings of unskilled American workers have concluded that foreign trade is a small contributor.²

SOPHISTICATED ARGUMENTS FOR PROTECTION

While most of the protectionist arguments we read in the media are based on a misunderstanding of comparative advantage, some more recent arguments for protecting domestic industries are based on a more sophisticated understanding of how markets work. These arguments have become collectively known as *strategic trade policy*. According to its proponents, a nation can gain in some circumstances by assisting certain “strategic” industries that benefit society as a whole, but that may not thrive in an environment of free trade.

Strategic trade policy is most effective in situations where a market is dominated by a few large firms.³ With few firms, the forces of competition—which

² The studies include Robert Z. Lawrence and Matthew J. Slaughter, “Trade and U.S. Wages: Giant Sucking Sound or Small Hiccup?” *Brookings Papers on Economic Activity: Microeconomics*, 2:1993, pp. 161–210; and Jeffrey D. Sachs and Howard J. Shatz, “Trade and Jobs in U.S. Manufacturing,” *Brookings Papers on Economic Activity*, 1:1994, pp. 1–84.

ordinarily reduce profits in an industry to very low levels—will not operate. Therefore, each firm in the industry may earn high profits. These profits benefit not only the owners of the firm, but also the nation more generally, since the government will be able to capture some of the profit with the corporate profits tax. When a government helps an industry compete internationally, it increases the likelihood that high profits—and the resulting general benefits—will be shifted from a foreign country to its own country. Thus, interfering with free trade—through quotas, tariffs, or even a direct subsidy to domestic firms—might actually benefit the country as a whole.

An argument related to strategic trade policy is the *infant industry argument*. This argument begins with a simple observation: In order to enjoy the full benefits of trade, markets must allocate resources toward those goods in which a nation has a comparative advantage. This includes not only markets for resources such as labor and land, but also *financial markets*, where firms obtain funds for new products. But in some countries—especially developing countries—financial markets do not work very well. Poor legal systems or incomplete information about firms and products may prevent a new industry from obtaining financing, even though the country would have a comparative advantage in that industry once it was formed. In this case, government assistance to the “infant industry” may be warranted until the industry can “stand on its own feet.”

Strategic trade policy and support for infant industries are controversial. Opponents of these ideas stress three problems:

1. Once the principle of government assistance to an industry is accepted, special-interest groups of all kinds will lobby to get the assistance, whether it benefits the general public or not.
2. When one country provides assistance to an industry by keeping out foreign goods, other nations may respond in kind. If they respond with tariffs and quotas of their own, the result is a shrinking volume of world trade and falling living standards. If subsidies are used to support a strategic industry, and another country responds with its own subsidies, then both governments lose revenue, and neither gains the sought-after profits.
3. Strategic trade policy assumes that the government has the information to determine which industries, infant or otherwise, are truly strategic and which are not.

Still, the arguments related to strategic trade policy suggest that government protection or assistance *may* be warranted in some circumstances, even if putting this support into practice proves difficult. Moreover, the arguments help to remind us of the conditions under which free trade is most beneficial to a nation:

Production is most likely to reflect the principle of comparative advantage when firms can obtain funds for investment projects and when they can freely enter industries that are profitable. Thus, free trade, without government intervention, works best when markets are working well.

This may explain, in part, why the United States, where markets function relatively well, has for decades been among the strongest supporters of the free-trade ideal.

³ Why might there be only a few firms in a market? In Chapter 8, you learned some of the reasons. These include economies of scale, legal barriers like patent protection, and strategic behavior on the part of existing firms to keep out competitors.

TRADE RESTRICTIONS IN THE UNITED STATES

No country has completely free trade with the rest of the world; every government limits trade in one way or another. And in spite of its strong pro-trade stance, the United States has restricted imports in many cases. Among the trade restrictions currently imposed by the U.S. government are the following:

- Foreign airlines may not carry domestic passengers from one point to another inside the United States.
- Canadian lumber can enter the United States only in limited quantities.
- Imports of fibers and textiles are tightly limited.
- Importers of many products have to pay tariffs.
- The amount of sugar that can be imported is tightly limited and is far less than would occur with free trade.

In addition, the government often takes temporary steps to limit certain kinds of imports or to raise their prices. For example, the United States has required Japan to limit exports of automobiles during certain periods, and the government required Asian manufacturers of computer memory chips to double the U.S. prices of their products for a time. Again, these practices, though restrictive, are not nearly as severe as those of many other governments: Japanese carmakers sell millions of cars in the United States, but almost none in Europe, where there is a flat ban on imports of their cars.

As we learned earlier in the chapter, protection benefits those who make the protected product, but it is bad for consumers. As a result, there is a tug-of-war between consumer interests and producer interests. Generally, in the United States, consumers have won the tug-of-war. Because so many imports are allowed into the country free of tariffs, the average U.S. tariff rate for all imports (which once approached 50 percent) was down to 3 percent by the mid-1990s. Thus, U.S. consumers enjoy the benefits of importing many of the products listed in Table 7—olive oil from Spain, tomatoes from Mexico, and cars and VCRs from Japan. Consumers also benefit indirectly when domestic producers buy inputs abroad, such as oil, aluminum, timber, and steel.

On the other side of the ledger, U.S. consumers suffer, and U.S. producers gain, from some persistent quotas, such as the sugar import quota. As you saw in Figure 2, a quota on imports raises the price to domestic residents. It is no surprise that the price of sugar in the United States is about 10 times higher than the world market price.

But quotas—like the U.S. sugar quota—create further problems of their own. First, because a quota raises the domestic price above prices elsewhere in the world, importers have an incentive to buy the good on the international market, violating the quota. The U.S. sugar quota, for example, has to be enforced by “sugar police.” Their job is to be sure that much of the sugar that is grown in the United States is exported, rather than sold domestically. Otherwise that sugar would eliminate the price differential and reduce the price of sugar in the United States to the free-trade price, like P_T in Figure 2 (p. 478). In this way, valuable resources—such as the labor of the sugar police—are used up to enforce the quota.

Another problem with a quota is how to decide who gets to import the restricted good. Importers have a lot to gain, since they can buy at the lower world price and sell at the artificially high domestic price. One logical approach would be for the government to auction off tickets that entitle the holder to import a given amount of the restricted good. Then the government would collect some revenue

Using the
THEORY



from the auction, making the quota similar to a tariff in its total impact. But this approach is never used in practice. Instead, the right to import is typically *given* away by the government, as in the case of sugar.

The impact of quotas in general can be understood by looking closely at the harm caused by the U.S. sugar quota:

1. It denies U.S. consumers the benefits of free trade—the ability to buy sugar cheaply from countries that have comparative advantages in sugar production.
2. It lowers the incomes of sugar producers in the generally poor, tropical countries that have comparative advantages in sugar production.
3. The gap between the U.S. and world market prices creates an incentive for illegal and wasteful activities, such as smuggling sugar, bribing the “sugar police,” or importing candy and refining it back into sugar. (Some people are actually in jail for defying the sugar import quota.)
4. The government’s power to grant sugar-importing rights causes people to waste resources lobbying for those rights, and it may cause corruption of the government officials in charge.
5. The government does not collect revenue that it could.

Who benefits from the sugar quota? A look back at Table 8 (p. 477) provides the answer: U.S. sugar producers and foreign sugar consumers. But as the principle of comparative advantage shows, the world as a whole is the loser.

S U M M A R Y

International specialization and trade enable people throughout the world to enjoy greater production and higher living standards than would otherwise be possible. The benefits of unrestrained international trade can be traced back to the idea of comparative advantage. Mutually beneficial trade is possible whenever one country can produce a good at a lower opportunity cost than its trading partner can. Whenever opportunity costs differ, countries can specialize according to their comparative advantage, trade with each other, and end up consuming more.

Despite the net benefits to each nation as a whole, some groups within each country lose, while others gain. When trade leads to increased exports, domestic producers gain and domestic consumers are harmed. When imports increase as a

result of trade, domestic producers suffer and domestic consumers gain. The losers often encourage government to block or reduce trade through the use of tariffs—taxes on imported goods—and quotas—limits on the volume of imports.

A variety of arguments have been proposed in support of protectionism. Some are clearly invalid, and fail to recognize the principle that both sides gain when countries trade according to their comparative advantage. More sophisticated arguments for restricting trade may have merit in certain circumstances. These include strategic trade policy—the notion that governments should assist certain strategic industries—and the idea of protecting “infant” industries when financial markets are imperfect.

K E Y T E R M S

exports
imports
absolute advantage

comparative advantage
terms of trade

exchange rate
tariff

quota
protectionism

R E V I E W Q U E S T I O N S

1. Describe the theory of comparative advantage.
2. What is the difference between absolute advantage and comparative advantage?
3. What are the terms of trade and why are they important?
4. What are the sources of comparative advantage?

5. What is a tariff? What are its main economic effects? How does a quota differ from a tariff?
6. What arguments have been made in support of protectionism? Which of them may be valid, and under what circumstances?
7. List the ways in which a quota on imported coffee would harm the nation that imposes it.

P R O B L E M S A N D E X E R C I S E S

1. Suppose that the costs of production of winter hats and wheat in two countries are as follows:

	Per Winter Hat	Per Bushel of Wheat
United States	\$10	\$1
Russia	5,000 rubles	2,500 rubles

- a. What is the opportunity cost of producing one more winter hat in the United States? In Russia?
 - b. What is the opportunity cost of producing one more bushel of wheat in the United States? In Russia?
 - c. Which country has a comparative advantage in winter hats? In wheat?
 - d. Construct a table similar to Table 3 that illustrates how a change in production in each country would increase world production.
 - e. If the exchange rate were 1,000 rubles per dollar, would mutually beneficial trade occur? If yes, explain what mechanism would induce producers to export according to their country's comparative advantage. If no, explain why not, and explain in which direction the exchange rate would change. (*Hint*: Construct a table similar to Table 6.)
 - f. Answer the same questions for an exchange rate of 100 rubles per dollar.
2. The following table gives information about the supply and demand for beef in Paraguay and Uruguay. (You may wish to draw the supply and demand curves for each country to help you visualize what is happening.)

Paraguay			Uruguay		
Price	Quantity Supplied	Quantity Demanded	Price	Quantity Supplied	Quantity Demanded
0	0	1,200	0	0	1,800
5	200	1,000	5	0	1,600
10	400	800	10	0	1,400
15	600	600	15	0	1,200
20	800	400	20	200	1,000
25	1,000	200	25	400	800
30	1,200	0	30	600	600
35	1,400	0	35	800	400
40	1,600	0	40	1,000	200
45	1,800	0	45	1,200	0

- a. In the absence of trade, what is the equilibrium price and quantity in Paraguay? In Uruguay?
 - b. If the two countries begin to trade, what will happen to the price of beef? How many sides of beef will be purchased in Paraguay and how many in Uruguay at that price?
 - c. How many sides of beef will be produced in Paraguay and how many in Uruguay? Why is there a difference between quantity purchased and quantity produced in each country?
 - d. Who benefits and who loses from the opening of trade between these two countries?
3. Use the data on supply and demand given in Question 2 to answer the following questions:
 - a. Suppose that Uruguay imposed a tariff that raised the price of beef imported from Paraguay to \$25 per side. What would happen to beef consumption in Uruguay? To beef production there? How much beef would be imported from Paraguay?
 - b. How would the tariff affect Paraguay? What would happen to the price of beef there after Uruguay imposed its tariff? How would domestic production and consumption be affected?
 - c. Suppose, instead, that Uruguay imposed a quota on the import of beef from Paraguay—only 200 sides of beef can be imported each year. What would happen to the price of beef in Uruguay? What would happen to beef consumption in Uruguay? To beef production there?
 - d. How would the quota affect Paraguay? What would happen to the price of beef there after Uruguay imposed its quota? How would domestic production and consumption be affected?

C H A L L E N G E Q U E S T I O N

Suppose that the Marshall Islands does not trade with the outside world. It has a competitive domestic market for VCRs. The market supply and demand curves are reflected in this table:

Price (\$/VCR)	Quantity Demanded	Quantity Supplied
500	0	500
400	100	400
300	200	300
200	300	200
100	400	100
0	500	0

- a. Plot the supply and demand curves and determine the domestic equilibrium price and quantity.
- b. Suddenly, the islanders discover the virtues of free exchange and begin trading with the outside world. The Marshall Islands is a very small country, and so its trading has no effect on the price established in the world market. It can import as many VCRs as it wishes at the world price of \$100 per VCR. In this situation, how many VCRs will be purchased in the Marshall Islands? How many will be produced there? How many will be imported?
- c. After protests from domestic producers, the government decides to impose a tariff of \$100 per imported VCR. Now how many VCRs will be purchased in the Marshall Islands? How many will be produced there? How many will be imported?
- d. What is the government's revenue from the tariff described in part (c)?

E X P E R I E N T I A L E X E R C I S E S

1. Visit the Office of the U.S. Trade Representative at <http://www.ustr.gov>. The Trade Representative—a Cabinet-level appointee—acts as the president's principal trade advisor, negotiator, and spokesperson on trade and related matters. Look at some of the most recent press releases. What are some trade-related issues facing the United States today? Be as specific as you can be about the countries, the products, and the problems.



2. The *Wall Street Journal* is a good source of information regarding international trade. A good place to look is the International page toward the back of the First Section of each day's *Journal*. Look at a recent issue and find an article dealing with trade barriers—tariffs, quotas, and so on. Model the trade barrier, using a graph, and try to determine who are the beneficiaries and who are the losers. If you are lucky, the article will provide sufficient information for you to determine the impact of the trade barrier on the price and quantity of the good in question.

THE MICROECONOMICS OF ONLINE RETAILING

Using All the THEORY

During the 1990s, something big happened to the United States and world economies. It can be summarized in the word *Internet*.

The technical foundations for the Internet were established much earlier—during the 1960s—by researchers working for the U.S. Department of Defense. But it wasn't until the early 1990s, when the point-and-click graphical interface of the World Wide Web was developed, that the Internet began to get the public's attention. From 1995 on, the number of people connected to the Internet grew rapidly, doubling every year. And by the end of 1999, 200 million people—half of them Americans—had Internet access.

The development of the Internet has provided a major shock to the entire economy, and the economy is still adjusting to that shock. It is changing the way that business firms produce goods. It has led to the creation and rapid growth of entirely new industries, including online retailing, online auctions, on-demand entertainment, Web consulting, and more. The Internet is creating unprecedented opportunities for entrepreneurs who understand the new technology's potential. And it has rocked financial markets around the world, and especially in the United States.

Economic changes of this magnitude—which cause a reconfiguration of the entire economy—don't happen often. When they do, it may seem as if the old rules no longer apply. Media pundits have argued that we are in a new economy, in which basic economic principles must be entirely revamped. Even the staid *Wall Street Journal*, to dramatize this viewpoint, declared in a headline: “So Long, Supply and Demand.”¹

But supply and demand have not gone. On the contrary, the Internet is a proving ground for the usefulness of supply and demand and the other microeconomic tools you've learned. As you will see in this chapter, these tools are needed more than ever if we are to understand how the economy is responding to the Internet, and how it will continue to respond in the years to come.



CHAPTER OUTLINE

Online Retailing: The Basics

The Big Picture: Online Retailing and Living Standards

How the Four-step Process Helps Us Analyze the Online Retail Industry

Resource Allocation: From Bricks and Mortar to the Internet

Online Retailing and Labor Markets

Online Retailing and the Stock Market

Time for You to Use the Theory

¹ Special section on “Industry and Economics,” *Wall Street Journal*, January 1, 2000.

But this “Using the Theory” is different from the others you’ve seen in this text in two ways. First, instead of looking at just *one* piece of the economy at a time—such as an individual firm, or the market for a particular product—we’ll be looking at several parts of the economy, viewing them as an integrated whole. Second, instead of analyzing small changes in mature markets—like the markets for wheat, gold, airline travel, or an exterminator’s services—we’ll be looking at *big* changes, at industries in upheaval, and at an entirely new industry: *online retailing*.

ONLINE RETAILING: THE BASICS

What is an online retailer? At the most basic level, it is a firm that provides goods and services directly to consumers who order on line. You will no doubt recognize the names of many of these firms: Amazon.com, barnesandnoble.com, Etoys, pets.com, CameraWorld.com, and so on. But to really understand this industry, we have to recognize that online retailers produce and sell *retail services*.

“But wait,” you might be thinking—“I buy *goods* from online companies. What’s this about *services*?” This is a subtle, but important, question. An online retailer does, indeed, ship goods to you. But these goods are not what it *produces*. Books are produced by book publishers, and CDs by music companies. The online retailer—like any retailer—just makes these goods readily available, so you can buy them. Thus, it is more accurate to say that what an online retailer sells, and what you buy from it, are the services of *making goods available*. The price you pay for such services is the retailer’s markup over the cost of the wholesale goods. For example, when barnesandnoble.com buys a book from the publisher for \$20 and sells it to you for \$29.95, you have paid a price of \$9.95 for the service of making the book conveniently available so you could buy it. In a sense, you paid \$20 for the book, and \$9.95 for the retail services associated with buying it.

The online retail industry is growing fast. In 1999, households bought only \$30 billion of goods over the Internet—a tiny fraction of total retail purchases. But, some observers forecast that within a decade, more than half of our retail purchases will be made on line. While others think such forecasts are exaggerated, no one doubts that online retailing will experience phenomenal growth over the next several years. To get an idea of how rapid this growth can be, consider that by 2003, *total* retail sales are forecast to exceed \$3 trillion. If even 10 percent of those purchases are made on line—\$300 billion—it would mean an increase of 900 percent in online retail purchases in just four years.

This dramatic increase in sales will be accompanied by an equally dramatic increase in the number of *firms* competing to offer online retail services. In fact, in the few months that you have been studying microeconomics, dozens of new online retail firms have been born.

What sorts of changes can we expect in the economy as a result of this explosive growth in online retail services?

THE BIG PICTURE: ONLINE RETAILING AND LIVING STANDARDS

One big change we can expect from the Internet in general—and from online retailing in particular—is an improvement in living standards. To understand why, remember that providing any kind of retail services—that is, making goods available

for people to buy—uses up some of society's *resources*. But online retailing will enable us to produce retail services using fewer resources than ever before.

To see why, let's compare traditional *bricks and mortar* (*B&M*) retailers—supermarkets, bookstores, appliance stores, furniture stores, and so on—with their online counterparts. Some resources are used in similar quantities by *both* types of retailers. For example, both must acquire wholesale goods to sell to the public, both must have warehousing facilities, both must maintain inventories of goods for sale, both must hire professional labor to advertise their services, design effective business strategies, file tax returns, and comply with government regulations.

But there are other resources that online retailers can either do without, or use more sparingly than B&M retailers. For example, online retailers do not use storefronts at all. This is a huge savings in physical capital, and especially in land. For example, Amazon.com ships many of its books from its huge distribution center in the desert of Nevada—some of the most inexpensive land available. By contrast, Borders—in order to sell books to its customers in the traditional way—must have stores on prime real estate in towns and cities across the country.

Not having stores saves online retailers more than just the cost of buildings and the land underneath them. It also saves on the labor of sales staff, cashiers, custodians, and security guards, and all of the resources embodied in display cases, automated security systems, air conditioning systems, and more. While online retailers *do* use more technically skilled labor (such as Web page designers and computer technicians), the *total* labor they need to provide any given amount of retail service is substantially lower on line than through stores.

How much of a saving in resources can online retailers enjoy? That is hard to say, because few online retailers have reached their long-run anticipated level of sales—a level that would take advantage of huge economies of scale. As these firms grow, the long-run average cost of providing their services should fall dramatically. We can, however, get some idea of the potential savings by looking at another industry that provides services over the Internet—the banking industry. In early 1998, Wells Fargo Bank calculated that the marginal cost of processing an in-person transaction—mostly labor time—was about \$1.07. By contrast, the marginal cost of processing an *online* transaction was just \$0.01.²

Another resource that is saved when retail services are provided on line is customers' time. Simply put, it takes time to shop the traditional way. For example, to shop for a new CD, you have to drive to the store, park, deal with crowds, find your CD, take it to the cash register, wait in line, go back to your car, and drive home. In many communities, this would take half an hour or longer. But on line, the entire transaction may take only two minutes—even less if you are shopping at a site you've visited before. Moreover, you may be able to hear a sample from your CD on-line, and decide you don't want it after all—saving you a trip to return unwanted merchandise. When this example is extended to other errands—shopping for food, clothing, toys, books, videos, tools, perhaps even groceries—the time savings can be substantial. The time freed up can be devoted to leisure activities, to work for pay, or even to more shopping.

Online retailing enables the same retail service to be provided using fewer resources. The resource savings occur on both the selling side (where less labor, land, and capital are needed to make goods available on line) and on the buying side (where consumers save time by shopping on line).

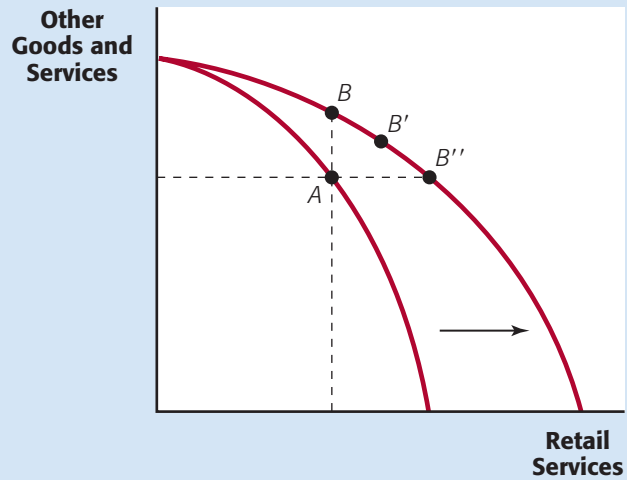
² *ComputerWorld*, January 5, 1998.

FIGURE 1

ONLINE RETAILING AND LIVING STANDARDS

The inner production possibilities frontier shows productively efficient combinations of retail services and other goods and services that the economy is capable of producing with its existing collection of resources. Through market interactions, society chooses to produce at point A.

The emergence of online retailing enables society to choose a point along the new, outer PPF. At point B, society would have the same amount of retail services as before, but *more* of other goods. At point B'', it would have more retail services, and the same amount of other goods. And at point B', it would have more of both. The economy's standard of living is higher because of the advent of online retailing.



To see how this affects society as a whole, look at Figure 1, where we use a familiar tool: the production possibilities frontier (PPF), first introduced in Chapter 2. To keep things simple, we'll assume that amount of leisure time that families enjoy remains constant, so that any resources freed up from the development of online retailing are used to produce more goods and services. The figure divides our total yearly production into two categories: retail services (measured along the horizontal axis) and all other goods and services (measured along the vertical axis). Initially, before the development of online retailing, we are restricted to a point on the inner PPF, such as point A.

Now we introduce online retailing, which uses fewer resources to produce a given amount of retail services. As a result, the maximum amount of retail services society can provide with its available resources increases—the horizontal intercept of the PPF moves rightward. The vertical intercept, however, remains unaffected by the development of online retailing.³

The end result is that the PPF pivots outward from the vertical axis. Society can now choose any point on the *outer* PPF. Point B represents one such choice: where society consumes the same amount of retail services as before, and uses all the freed-up resources to produce *other* things that we value. Point B'' represents another choice: the same amount of other goods and services, and more retail services. Finally, point B' represents an intermediate case: a higher level of retail services *and* more of all other goods and services.

Online retailing will shift the economy's production possibilities frontier outward, and enable us to enjoy a higher standard of living. That is, we can have more and better retailing services, or more of other things we value, or both.

³ At the same time that online retailing is shifting out the horizontal intercept of the PPF, other Internet developments are helping to shift out the *vertical* intercept (not shown in the figure). See end-of-chapter Problem 1 to explore this further.

A higher average living standard is certainly a good thing. But our analysis leaves many questions unanswered. For example, *how* will resources be shifted from brick and mortar stores to Internet retailers? And how will the *gains* to our society be distributed among different economic players: the stockholders who own the online retailers, the consumers who buy from them, and the people who work in them? Will there be losers as well as gainers? And why should the stock prices of these companies fluctuate as wildly as they have?

To answer these questions, we must look at the individual parts of the economy and understand how these parts fit together. It won't surprise you that we'll be using our four-step process to do this.

HOW THE FOUR-STEP PROCESS HELPS US ANALYZE THE ONLINE RETAIL INDUSTRY

KEY STEP #1: CHARACTERIZE THE MARKET

The first step in answering almost any question about the economy is to *characterize the market*. But which market should we look at? Online retailers are involved in a number of *different* markets, and the ones we choose for our analysis—and how we characterize them—will depend on the specific question we are trying to answer.

For example, if we want to determine whether online retailers can earn economic profit in the long run, we'll need to look at *product markets*, in which retailers sell their services to consumers. To analyze the effects on wages and salaries of skilled and unskilled workers, we'll need to focus on *labor markets*, in which online retailers hire their employees. To analyze what has been happening to the value of Internet stocks, we'll be looking at a specific *financial market*—the *stock market*. In each of these cases, we'll be characterizing a market or group of markets. That means we'll have to identify the buyers and sellers who have the potential to trade, and decide which type of market model to use—perfect competition, monopolistic competition, oligopoly, or monopoly.



Characterize the Market

KEY STEP #2 IDENTIFY GOALS AND CONSTRAINTS

In every market, the buyers and sellers who come together to trade have goals and face constraints. In most cases, the goals and constraints are similar to those we've discussed elsewhere in this text. For example, in the markets for online and traditional retail services, a consumer—in deciding whether to buy products over the Internet or at a local store—strives to achieve the highest possible level of utility, and faces the constraints of having to shop for his purchases with limited income and limited time. The goal of maximum utility and the constraints of limited income and limited time are built into the demand curves we see in these markets.

Similarly, much of what we'll assume about the goals and constraints of online retailers is familiar. For example, firms face the familiar constraints of a given production technology, of having to pay for their inputs, and—when they sell in competitive markets—of having to sell their product at the going market price

But what about the *goal* of retail firms? Is their goal—like all other firms we've discussed in this text—the familiar one of maximum profit?

Yes . . . and no. Because of the newness and the nature of this industry, we'll have to extend our theory of profit maximization beyond the material in Chapter 7. In that



Identify Goals and Constraints

chapter—and throughout the text—we’ve been able to get away with a simplification. We’ve assumed that the firm’s goal is to maximize profit in the *current* period, such as this year. But more realistically, the firm’s goal is not to maximize profit just *this* year, or just *next* year, but rather to maximize profits over a long period of time—as long as the firm will exist.

What does that mean in practice? Should a firm take an action that will result in a loss of \$10 million this year if, by doing so, it can earn an additional \$2 million in annual profit for the next 10 years? Or, more generally, how can a firm make decisions involving trade-offs of profits in one year for profits in another?

Chapter 13 gave the answer: present value.⁴ And it also suggests a more complete way of stating the firm’s goal—especially when it faces important trade-offs over time:

The firm’s goal—as it makes decisions in its product market, factor markets, and in financial markets—is to maximize the total present value of its future profits.

In a mature industry, a firm that maximizes each year’s profits will ordinarily be maximizing the total present value of profits as well. Thus, our assumption throughout this text that a firm’s goal is to maximize profit during the current period is a useful and realistic assumption in most cases.

But in a new Internet industry like online retailing, maximizing the total present value of future profits may require suffering huge losses in the early years. After all, new dot.com firms must pay the initial setup costs to establish a presence on the Web, and the initial high advertising costs needed to attract first-time customers. For this reason, the simplified view of the firm’s goal—to maximize current-year profit—will not work at all if we want to understand the motives and actions of Internet firms.

Find the Equilibrium



KEY STEP #3: FIND THE EQUILIBRIUM

Many of the important questions we will ask about the online retail industry center on Key Step #3. Can online retail firms become—and remain—profitable? Will the compensation of employees there remain high? Are the stocks of individual firms—and the sector as a whole—a good investment? To answer these questions, we must find the long-run equilibrium in online retail markets. And, as you’ll see, there is considerable disagreement among observers over the nature of that equilibrium.

What Happens When Things Change?



KEY STEP #4: WHAT HAPPENS WHEN THINGS CHANGE?

In most economic analysis, this is the most interesting and important step. But it’s especially important when analyzing Internet industries, for two reasons. First, the Internet itself is an important change—one that is causing a fundamental re-

⁴ A dollar received in the future is worth less than a dollar received today. To compare dollars received at different points in time, we need to convert them all to their present-day equivalents. We do this by choosing an appropriate discount rate and then using the formula $PV = Y_t/(1 + i)^t$, where Y_t is the amount of money to be received t years in the future, i is the discount rate, and PV is the value today of the sum to be received t years from now.

configuration of the economy. Second, because the Internet is so new, the changes that *affect* it—changes in technology, in government policy, and in consumer tastes—are much bigger and have much more impact.

In the remainder of this chapter, we'll use the four-step process—along with many aspects of the microeconomic theory you've learned in this text—in order to answer a number of questions about the impact of online retailing. That is, we'll be looking at several different types of markets where buyers and sellers come together, each trying to achieve their goals and each facing constraints. We'll be examining the equilibrium in each of these markets, and we'll observe what happens when that equilibrium changes.

RESOURCE ALLOCATION: FROM BRICKS AND MORTAR TO THE INTERNET

In Figure 1, we saw that society has much to gain by shifting production of retail services from traditional stores to the Internet. But this, in turn, requires our economy to reallocate resources from one sector to another. Resources that would otherwise be used up building and maintaining stores or display cases, or providing assistance to traditional shoppers, must now be shifted toward manufacturing fiberoptic cable, creating transmission networks, and designing and maintaining Web pages. How does this shift in resources come about?

We can answer this question by using the model of perfect competition. True, perfect competition is not an *exact* fit for online and traditional retail markets. But it comes close enough to be useful. For example, in most retail markets, there are many close competitors. While each firm is not *strictly* a price taker, it is pretty close to being one. If it raises the price of its retail services outside of a narrow range (that is, begins marking up its goods by, say, 10 percent or 20 percent more than its competitors), it will soon lose all or most of its sales to competitors. While retail services are not a completely standardized product, they are somewhat standardized: Most retailers give the same guarantees, have roughly the same level of service, and offer a very similar product selection. Finally, barriers to entry and exit, if they exist at all, are not too significant. Traditional retailing has always been an easy market to enter and exit, and so far, the same has been true for dot.com retailing, as shown by the many new competitors that have entered virtually every online retail market in the past few years.

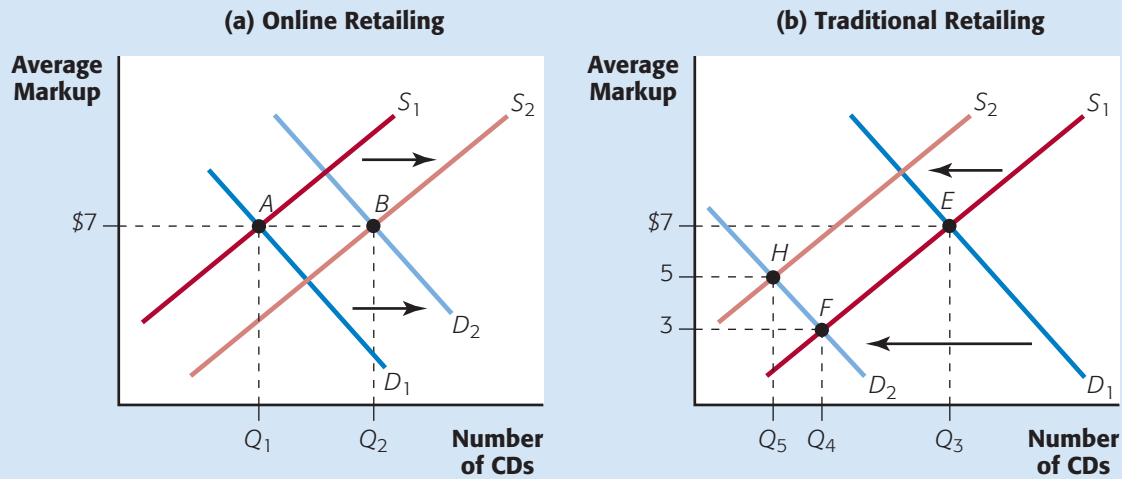
Let's narrow our focus and look at two very specific retail markets: online and traditional retail services for compact discs. The *traditional* CD retailers include Wal-Mart, MediaPlay, Kmart, and Tower Records, as well as all the smaller, independently owned shops. The online sellers include Amazon.com, barnesandnoble.com (partially owned by Barnes & Noble, a separate firm), buy.com, cduniverse.com, and CDNow.

CHANGES IN THE ONLINE RETAIL MARKET

Panel (a) of Figure 2 shows what is happening in the online market. In this diagram, we are using the quantity of CDs as our measure of retail services on the horizontal axis, and the average markup as our measure of the price of retail services on the vertical axis. The demand curve D_1 shows the demand for online retail services. The

FIGURE 2

PRODUCT MARKET EFFECTS OF GROWTH IN INTERNET RETAILING



Panel (a) depicts the market for CDs sold on line. Initially, the supply and demand curves intersect at point A to determine a retail markup of \$7 per CD; Q_1 CDs are sold at that markup. As online purchasing become more popular, the demand curve shifts rightward to D_2 . The supply curve shifts rightward as well because new firms enter this industry in search of profits. Point B shows one possible long-run equilibrium: The average markup is the same as before, but the quantity of CDs sold online is higher at Q_2 .

Panel (b) shows what happens to traditional bricks and mortar CD retailers. As customers shift their purchases on line, the market demand curve for traditional firms shifts leftward to D_2 . In the short-run equilibrium at point F , existing firms suffer economic losses. Some of them exit the industry, shifting the supply curve leftward to S_2 . In the new long-run equilibrium at point H , fewer CDs are sold in bricks and mortar stores.

short-run supply curve S_1 shows the amount of online retail services provided by *firms already in the industry*. The current short-run equilibrium in this market is at point A . (We don't assume this is a long-run equilibrium; indeed, no Internet market has yet achieved a long-run equilibrium.) At point A , the amount of retail services offered (measured by the number of CD's sold) is Q_1 , and the price of those services (the markup over cost) is \$7 per CD.

The second set of supply and demand curves in panel (a) illustrates what will happen in the online retail market over the long run. The demand curve will shift rightward for two reasons. First, more people will become connected to the Internet and have the *ability* to order goods on line. And second, even among those who have Internet access, tastes for buying goods on line should increase over time. (See end-of-chapter Problem 2 to explore this further.)

But the demand shift is not the only long-run change; the supply curve will shift rightward as well. That's because entry will continue to occur over the next several years. Some new entrants may be entirely new firms—firms with new ideas about customer service or firms serving specialized markets in particular types of music. And traditional retailers, such as Tower Music or Kmart, may enter and try to grab some of the online market for themselves.

What is the result of these changes in demand and supply? We'll discuss that soon. But before we do, it's time to address an issue that may be troubling you.

A DETOUR: ENTRY AND ECONOMIC PROFIT

In Figure 2, the supply curve shifts rightward because of entry of new firms. Does that make sense? So far in this text, we've assumed that entry occurs when firms that are already in the industry are enjoying economic profit. We've even identified profit as one of the major forces that help to allocate resources in a market economy: Profit attracts new firms, while losses cause exit.

But that is certainly *not* what is happening in online retailing. After all, almost *all* online retailers have suffered significant losses since their founding, and they expect these losses to continue for several years into the future. Why, then, are so many firms scrambling to enter this industry when they *should* be rushing to the exit doors?

The answer centers on our more complete statement of the firm's goal, which we introduced earlier in this chapter: Firms are concerned not with just current-year profit, but rather with the *total present value of future profit*.

The forces driving entry and exit in the long run are not current profit, but rather the total present value of future profit that firms anticipate. When a potential entrant anticipates positive total present value of future profit, it will enter the industry, even if it anticipates short-run losses.

The online retailers, themselves, always accompany admissions of current losses—in documents they are required by law to file with the Securities and Exchange Commission—with assertions that their losses are due to *temporarily* high expenses. The implication is that, while losses may continue in the short run, they will eventually—as the firm matures—turn into profit. (See Table 1.) Such statements help to reassure the firm's shareholders that they should hang onto their stock in spite of current losses, because future profit will more than make up for it.

SOME CONCLUSIONS ABOUT THE MARKET FOR ONLINE RETAIL SERVICES

In panel (a) of Figure 2, we end up at point *B*, where the price of online retail services—the average markup on a CD—remains the same as at point *A*. This is because—in our diagram—we happened to assume that both the supply and demand curves shifted rightward by the same amount. But this need not be the case. If the supply curve shifts out by more than the demand curve, then the average markup on CDs purchased over the Internet will fall from its current level. If the demand curve shifts out by more than the supply curve, the average markup will rise.

Finally, there is another possibility: that the market for buying CDs on line begins to deviate from perfect competition so much that our supply and demand model is no longer useful. For example, if online retailing in CDs becomes an oligopoly with just two or three large firms, we would need to use a game theory model to analyze and predict the average markup in the industry. No doubt, average markups in such an oligopoly would be higher than under perfect competition, for reasons that you learned about it in Chapter 10.

Regardless of what happens to price, however, we can be reasonably certain that the *quantity* of online retail services in the CD market will increase. Or, more simply, people will buy more of their CDs on-line five years from now than they do today.

TABLE 1

RECENT LOSSES FOR ONLINE RETAILERS, AND EXPLANATIONS

Internet Retailer	Core Offering	Accounting Loss in 1999	Reason for Loss
Amazon.com	Books/media	\$720 million	"We have incurred significant losses since we began doing business. . . . To succeed we must invest heavily in marketing and promotion and in developing our product, technology and operating infrastructure. Our aggressive pricing programs have resulted in relatively low product gross margins. . . . For these reasons we believe that we will continue to incur substantial operating losses for the foreseeable future, and these losses may be significantly higher than our current losses."
Egghead.com	Computers/software	\$154.9 million	"We expect gross margins to continue to be low due to our aggressive efforts to gain market share . . . by expanding and enhancing our customer service operations, our promotional offerings . . . extending credit to certain business customers . . . waiving all or part of the shipping and handling fee for limited periods of time and offering promotional pricing on specific products."
Beyond.com	Software	\$124.8 million	"For the foreseeable future, the Company intends to expend significant financial and management resources on brand development, marketing and promotion, site content development, strategic relationships . . . , and technology and operating infrastructure, including ESD capabilities. As a result, the Company expects to incur additional losses and continued negative cash flow from operations for the foreseeable future . . . [T]he Company . . . believes that period-to-period comparisons of its operating results are not necessarily meaningful and should not be relied upon as an indication of future performance."
Garden.com	Gardening	\$19.1 million	"The Company expects to experience operating losses and negative cash flow for the foreseeable future . . . [due to] expenses related to brand development, marketing and other promotional activities, content development and technology and infrastructure development. . . . To date, the Company has funded its operations from the sale of equity securities and has not generated sufficient cash from operations."

Sources: Susan Reda, "1999 Top 100 Internet Retailers," *STORES* (September 1999) (available at <http://www.stores.org/eng/archives/sept99cover.htm>); Charles Schwab Web site (www.schwab.com), accessed on March 16, 2000; and various filings by these firms with the Securities and Exchange Commission.

And what is true of CDs is true of other retail markets as well: markets for books, videos, software, computers, electronic equipment, automobiles, and more. In general:

firms expect future profits from selling goods online, and they expect the total present value of future profits to be positive, in spite of early losses. These expected future profits serve as a market signal, encouraging firms to enter online retail markets, and provide more online retail services to society.

THE IMPACT ON TRADITIONAL RETAILERS

Markets are interconnected. Changes that take place in one market can have important effects in other markets. The changes we've been discussing in markets for online retail services have an important impact on markets for *traditional* retail services.

Look at panel (b) of Figure 2 (on p. 494), which shows the market for *traditional* retail services provided by bricks and mortar sellers of CDs. Once again, the quantity of retail services is measured by the number of CDs sold, while the price of retail services is the markup over cost. Initially, this market is in long-run equilibrium at point *E*.

But the increased demand in the *online* market is associated with a *decreased* demand in the *traditional* market. Thus, the demand curve for traditional retail services is shifting leftward.

In the short run—a period too short for traditional retailers to exit the industry—equilibrium moves to point *F*. Traditional markups fall, creating losses (not shown) for the typical traditional retailer. These losses work as *market signals*, telling retail firms that society would be better off if their resources were freed up for other uses. In the long run, the prospect of future losses will cause some firms to exit the industry, and the supply curve will shift leftward, eventually reaching S_2 . In the end, the market settles at point *H*, with fewer traditional CD retailers, supplying fewer retail services, and charging a lower markup on each CD than initially.

The changes seen in panel (b) of Figure 2 can be predicted for many traditional retail markets, not just CDs. In general,

over the next several years, the demand for traditional retail services will decrease. All else equal, this will lead to lower markups and short-run losses at traditional retailers, causing some of them to exit the industry. In long-run equilibrium, there will be fewer traditional retailers than initially.

Note that our analysis assumes no other changes in traditional retail markets, except for the leftward shift in the demand curve and the long-run changes in supply that follow. In the real world, however, other things may change at the same time. For example, over the next 5 or 10 years, we can expect consumer incomes to grow. This will shift demand curves rightward in all markets for normal goods—including markets for traditional retail services. It is possible that the growth in income could have such a strong effect that the demand curve in panel (b) would shift rightward, not leftward, and our conclusions would have to be rephrased. In this case, there would be *two* changes, each having an impact on traditional retailers. The change in income, alone, would tend to *increase* the number of traditional retailers and *increase* the markup they could charge. The Internet

would then work *against* these effects, tending to push the demand curve leftward, decreasing the number of retailers and the markups they could charge. In the end, the impact on the market will depend on which effect dominates. But we can certainly conclude the following:

Because of online retailing, there will be fewer traditional retailers than there would otherwise be, and each will charge a lower markup than it otherwise would.

ONLINE RETAILING AND LABOR MARKETS

The shift from bricks and mortar to online retail services requires that *resources* be reallocated from one sector to the other. And one of the most important resources being reallocated in this way is *labor*. New online retailers need to hire Web page designers, Internet consultants, marketers, strategic planners, business attorneys, product packers, and more. At the same time, sales staff, security guards, cashiers, and shelf-stockers in traditional retail outlets must move to other jobs that society values more highly.

But how, exactly, does this reallocation of labor take place? What, for example, causes workers needed by online retailers to move there from other jobs? And what causes labor to leave the traditional retail sector and move to jobs that society finds more valuable? Finally, what are the implications of all these changes for different types of workers?

THE IMPACT ON INTERNET PROFESSIONALS

Let's start by exploring the market for a particular type of worker increasingly hired by Internet firms: business attorneys. Online retailers and other Internet firms hire these attorneys to advise on patent, tax, and labor policies; to write and negotiate contracts with business partners, suppliers, and investors; and to prepare official filings for the SEC and other government agencies.

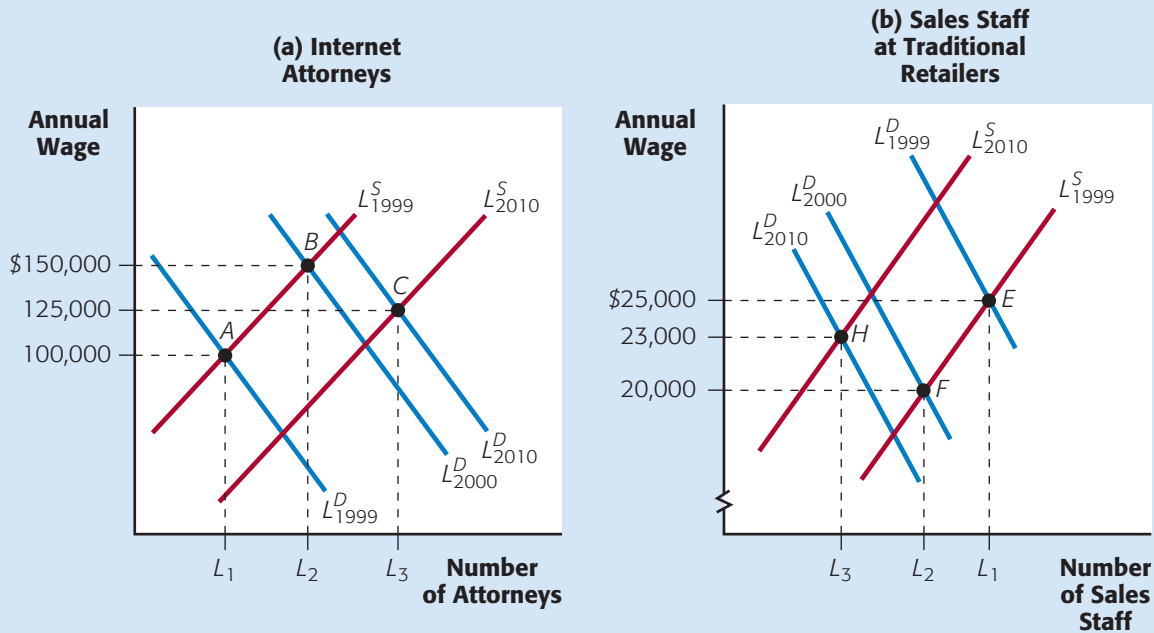
Panel (a) of Figure 3 shows us what has been happening in this labor market in recent years. In the figure, we assume that the labor market for Internet attorneys is perfectly competitive.

We begin our analysis in 1999. The labor demand curve for that year, L_{1999}^D , sloped downward: the lower the salary, the more attorneys dot.com firms would want to hire. The labor supply curve L_{1999}^S is the *short-run* labor supply curve for that year. It tells us the amount of labor supplied by *those who were already attorneys qualified to work in Internet firms*. The curve slopes upward: At higher salaries, more qualified business attorneys offer their services to Internet firms.

In 1999, the labor market was in equilibrium at point A, with the typical entry-level Internet lawyer earning about \$100,000 per year. But between 1999 and 2000, things changed. As more dot.com firms were established, they entered this labor market to hire attorneys, shifting the labor demand curve rightward, to L_{2000}^D . The labor supply curve, however, did *not shift* during this time period, since one year is too short a time for people to acquire law degrees or change their specialty from, say, family law to business law. In the short run, then, the shift in labor demand

LABOR MARKET EFFECTS OF GROWTH IN ONLINE RETAILING

FIGURE 3



Panel (a) shows the market for Internet attorneys. In 1999, the labor supply and demand curves intersected at point A to determine an annual wage of \$100,000. However, as more Internet firms set up business, the demand for attorneys increased, shifting the demand curve rightward to L_{2000}^D . In the short-run equilibrium at point B, the annual wage is higher—\$150,000 per year. Eventually, though, this high wage will attract additional lawyers, and the market supply curve will shift to L_{2010}^S . Simultaneously, the growth of online retailing will continue to shift the demand curve to the right. In the figure, long-run equilibrium is re-established at point C, with a wage of \$125,000.

Panel (b) depicts the labor market for less-skilled workers in traditional retailing, where labor demand is decreasing. In the new short-run equilibrium at point F, the annual wage drops to \$20,000. In the long run, some workers leave this industry, shifting the supply curve to L_{2010}^S . At the same time, the demand curve will continue to shift leftward as traditional retailers exit the industry. In the new long-run equilibrium at point H, fewer unskilled workers are employed in the traditional retail sector.

caused the equilibrium to move along the short-run labor supply curve to point B, with a new salary of \$150,000. This change in salary might seem unrealistically large for one year, but it is an accurate representation of actual events: Salary and other compensation for new Internet attorneys from good law schools actually *did* rise by about 50 percent between 1999 and 2000.

But point B is *not* the long-run equilibrium in this market. Over the long run, the high salaries of Internet attorneys will cause entry into this profession. More college graduates will choose law school over, say, medical school, and more law students will choose specialties in business law. Within a few years—as these new business lawyers hit the market—the labor supply curve for Internet lawyers will begin shifting rightward. If this were the only change occurring, the salary would fall back toward its initial level. But, as we discussed earlier in this chapter, there is another change we can expect over this period: continued entry by

new dot.com firms. This will shift the labor demand curve farther and farther rightward over time.

When the number of Internet firms and business lawyers stabilizes—the figure assumes this occurs in the year 2010—the labor supply and labor demand curves will stop shifting. At that point, with L_{2010}^S and L_{2010}^D , the new, long-run equilibrium is at point C. Notice that, in our diagram, the long-run compensation of Internet lawyers is higher than in 1999, but lower than in 2000. But that *need not* be our result. If the dot.com sector grows large enough (and the labor demand curve shifts far enough rightward), the salary could end up *higher* than its initial value. (You'll be asked to diagram this case in end-of-chapter Problem 6.)

Our analysis applies not just to Internet attorneys, but also to *many* types of professional labor needed by online retailers and other Internet firms. More generally,

the demand for highly skilled workers needed by Internet firms is increasing, causing salaries for these workers to soar. The rise in salaries acts as a market signal—telling individuals that society would be better off if they took jobs in Internet firms. Entry of new workers would ordinarily bring salaries back down somewhat. But continued entry by dot.com firms will work against the drop in salaries. In the new long-run equilibrium, there will be more highly skilled professionals working at Internet firms, earning higher salaries than initially, but not necessarily as high as in the short run.

THE IMPACT ON TRADITIONAL RETAIL WORKERS

Now shift your attention to panel (b) of Figure 3, which shows the *other* side of the labor market story—the unpleasant side. Here, we look at the market for sales staff at traditional retail outlets. We've already seen that, in the short run, these outlets will have to lower their markups in order to compete with online retailers. In our diagram, this lower market price for retail services causes the labor demand curve for sales staff to shift leftward, from L_{1999}^D to L_{2000}^D . This one-year period is too short for sales staff to acquire the skills or training for other jobs, so there is no shift in the labor supply curve. The equilibrium moves from point E to point F, and the salary of a typical sales person, which was low to begin with, drops further—in our example, from \$25,000 in 1999 to \$20,000 in 2000.

But this is not the end of the story. In the long run, the drop in salaries will encourage some workers to leave the traditional retail sector entirely—shifting the labor supply curve leftward. At the same time, some traditional retailers—continuing to suffer losses—will exit the industry, shifting the labor demand curve *further* leftward. When these adjustments stop—in the year 2010 in the figure—the market reaches its new, long-run equilibrium at point H. In our example, the salaries of sales staff rebound somewhat, but they remain lower than they were initially. However, if the traditional retail sector shrinks enough in the long run, salaries could actually drop *below* their initial level. (End-of-chapter Problem 7 asks you to diagram this case.)

Our analysis applies not just to sales help, but also to security guards, cashiers, and other less-skilled workers who have been working in the traditional retail sector. More generally,

because of online retailing, the demand for traditional retail workers is decreasing, or rising more slowly than it otherwise would. As a result, salaries for these workers are stagnating. This acts as a market signal—telling individuals that society would be better off if they took jobs elsewhere. Exit of traditional retail workers would ordinarily bring salaries back up somewhat. But continued exit by bricks and mortar retailers may work against this effect. In the new long-run equilibrium, there will be fewer people working at traditional retail outlets, earning lower salaries than initially, but not necessarily as low as in the short run.

EFFECTS IN OTHER LABOR MARKETS

So far, our analysis has centered on two types of workers directly affected by online retailing: professionals who work for dot.com firms, such as Internet lawyers, and less-skilled people who work at traditional retailers, such as retail sales clerks.

But the effects extend to *other* labor markets as well. Indeed, the shift to online retailing—and the development of the Internet more generally—is working to exacerbate a trend we first observed in Chapter 11: the rising earnings differential between highly skilled workers (college and professional school graduates) and less-skilled workers (high school graduates or less). In 1998, the average college graduate earned almost twice as much as the average high school graduate. And economists project that the difference will grow over the next decade.

How does the shift from traditional to online retailing contribute to this trend across the economy? Figure 4 tells the story. Panel (a) shows the market for business lawyers who work *outside* the Internet sector, such as in traditional law firms. As lawyers are attracted to the Internet sector, as described earlier, the labor supply curve for *non*-Internet lawyers shifts leftward, and the equilibrium moves from point *J* to point *K*. Salaries rise—in our diagram, from \$100,000 in 1999 to \$150,000 in 2010. And this is more than a hypothetical possibility. Traditional law firms—even those that have nothing to do with the Internet—have had to increase compensation of new attorneys by 50 percent or more in order to counter offers from Internet firms.

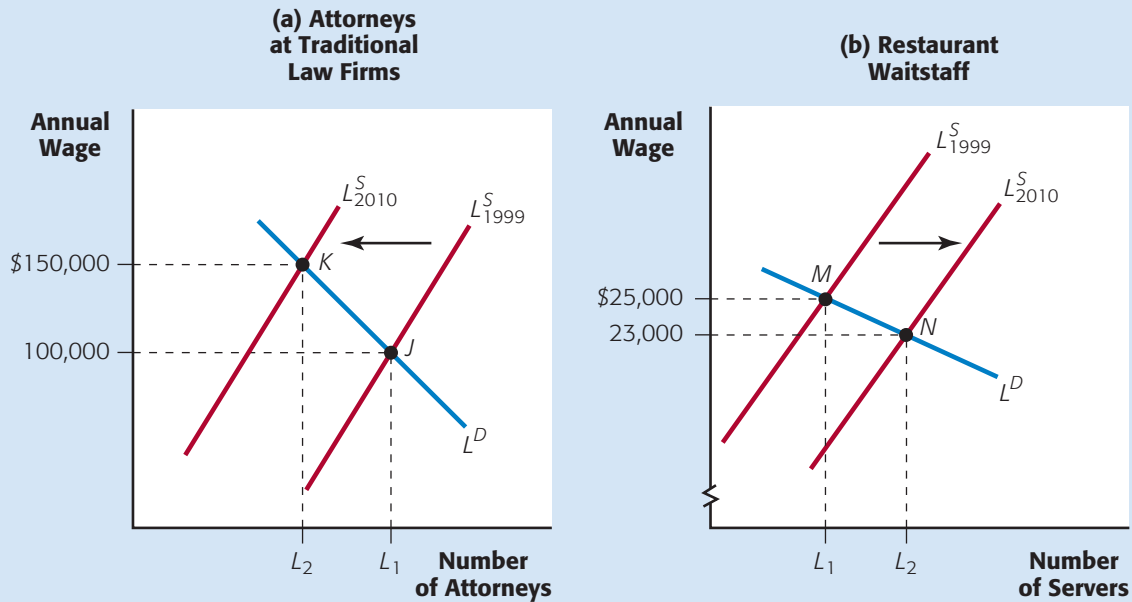
Similar changes will be observed in professional markets across the economy. In the long run, salaries of doctors, engineers, college professors, and commercial artists will rise, because labor supply curves in all of these markets will be shifting leftward as some people leave for high-paying careers in Internet firms.

By contrast, panel (b) of Figure 4 shows what happens in a *less*-skilled labor market: the market for restaurant waiters and waitresses. As the traditional retail sector declines, retail sales staff move to other labor markets, such as the market for restaurant help. The labor supply curve in this market shifts rightward, and the equilibrium wage falls. And the same will occur in markets for less-skilled labor throughout the economy: Labor supply curves in these markets will shift rightward, and wages there will drop.

The impact of the shift from traditional to online retailing extends beyond the markets that are directly affected. Because Internet firms tend to hire more highly skilled workers, the wages of these workers will increase, whether they work for Internet firms or not. Because traditional retailers tend to hire mostly less-skilled workers, their wages will decrease or rise more slowly, whether they work in traditional retail stores or not.

FIGURE 4

EFFECTS OF ONLINE RETAILING IN OTHER LABOR MARKETS



Panel (a) shows the market for attorneys in traditional law firms. Starting from the initial equilibrium at point *J*, the labor supply curve shifts leftward as attorneys are lured away by the higher wages at Internet firms. As a result, the wage rate rises to \$150,000 at point *K*.

In panel (b), the initial equilibrium in the market for restaurant waitstaff is at point *M*. As workers leave the traditional bricks and mortar retail sector, they seek work in this and other markets for less-skilled labor. The increased supply in the restaurant labor market drives the wage down to \$23,000 per year at point *N*.

These effects, if not addressed, will lead to a further widening of the wage gap between more educated and less-educated workers. Of course, over the next 10 years, other changes might offset or reverse this trend. For example, subsidies that encourage young people to attend college would work to counteract the labor supply shifts in *both* panels of Figure 4. (End-of-chapter Problem 8 asks you to illustrate and explain.)

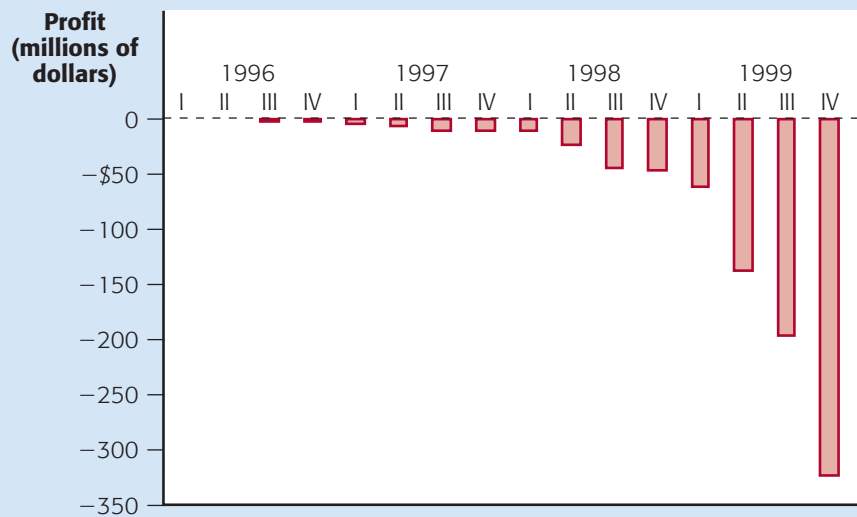
ONLINE RETAILING AND THE STOCK MARKET

As this is being written, Internet stocks—especially the stocks of online retailers—seem to be violating all of the rules of stock pricing.

All else equal, we would expect firms with high earnings per share to have high stock prices, and firms with low earnings per share to have low stock prices. But consider the stock of Amazon.com—the most widely visited Internet retailer. It has never earned a profit, and has no prospect of earning a profit during the next several years. In fact, its losses have been astounding. Figure 5 shows Amazon's quarterly losses—as negative profit—from 1997 through 1999. Over that period, Amazon lost a total of \$881 million.

QUARTERLY ACCOUNTING PROFIT AT AMAZON.COM, 1996–1999

FIGURE 5



Note: Negative numbers represent losses

Yet Amazon's market value on March 20, 2000—what it would cost to buy up all the shares of Amazon in circulation on that day—was \$21.8 billion. That is eight times the combined market value of Barnes and Noble and Borders, which together earned \$182 million in profit in 1999.

Or consider the amount of physical capital a firm owns. If worse came to worst, a company could always liquidate itself, and sell all of its capital—the office buildings it owns, manufacturing plants, computers, office furniture, vehicles, and so on. In theory, we might think that a firm's market value should bear some relation to the value of its physical capital. But consider TWA. In early 2000, the airline owned 185 aircraft, valuable landing rights in two dozen countries, and office space in prime real estate across the country. Yet the company's market value was \$146 million—substantially less than the value of its capital. Meanwhile, the market value of Yahoo—which owned only \$20 million worth of physical capital—was \$90 billion!

Finally, consider *information* about Internet firms. Everyone knows the business they are in. Everyone knows how much profit or loss they've made each quarter, since by law, a publicly owned corporation must publish this information using standard accounting practices. And everyone reads the same forecasts of these firms' long-term prospects. So it seems there should be some agreement—on any day, in any week, or in any month—on just how much a share of stock in an Internet firm is worth.

But in fact, the stocks of new Internet firms—especially online retailers—have had a wild ride in recent years. For example, on June 1, 1999, a share of Amazon's stock could be bought for \$52.91. By December 10, 1999, the price had risen to \$106.69. Then, in just over three months, it fell again to \$62.50 on March 3, 2000. These overall swings masked even more volatile day-to-day movements. For example, on December 9, 1999, the stock rose 17 percent over its value the day before. On January 5, 2000, the value plunged by almost 15 percent—again, in just a single day!

What has been going on here?

In fact, what has been happening to Internet stocks—their high valuations in spite of losses and little physical capital, and their wild price fluctuations—makes perfect sense when you use the tools of microeconomic theory.

Let's start with a fundamental concept from Chapter 13: the principle of asset valuation. It tells us that the value of a firm is the total present value of its future earnings. According to this principle, a firm that has no earnings—or has suffered year after year of losses—could still have value if its anticipated future earnings are high enough to compensate for the losses.

Of course, future earnings are *discounted*, so a dollar in the future is worth considerably less than a present dollar. Thus, the longer a firm's earnings are postponed into the future, the higher those earnings have to be to justify a high stock price now.

Let's calculate how much profit Amazon would have to earn in future years to make its market value, which was \$21.8 billion in March 2000, an attractive price at which to buy the company. We'll assume a discount rate of 12 percent (a few points above the going interest rate in early 2000, to adjust for risk), and also assume that Amazon stops losing money and turns profitable immediately as we start our calculations. We'll also assume, for simplicity, that once Amazon turns profitable, it earns a *constant* profit forever. This will allow us to use our special discounting formula (from Chapter 13) to calculate the total present value of a constant stream of future payments. Letting X represent that constant profit, our formula tells us that the total present value would be $X/0.12$. We want the value for X that makes this total present value equal to \$21.8 billion. So we must solve the following equation:

$$\frac{X}{0.12} = \$21.8 \text{ billion,}$$

giving us

$$X = \$2.61 \text{ billion.}$$

To recap, in order to make Amazon an attractive buy at its \$21.8 billion market value in March 2000, it would have to make future profits equivalent to a constant \$2.6 billion per year, beginning immediately and continuing forever. That is a huge amount of profit. By current standards, it would make Amazon the thirty-eighth most profitable corporation in the United States—ahead of American Express, Coca-Cola, Boeing, and Chevron, and just behind AT&T. Remember, too, that our calculations assume that Amazon will begin earning profit right away. But in actuality, the firm itself forecasts growing losses for several years, so it would have to earn a profit greater than \$2.6 billion to justify its market value.⁵

Stock market investors in early 2000 were clearly very optimistic about Amazon's future. But how *confident* were they? Apparently, not very. As noted earlier, Amazon's market value has several times been halved in a matter of months, only to double again a few months later. It seems that the market's view of Amazon's future is unstable.

⁵ Of course, Amazon stockholders hope that Amazon's profit will *grow* forever, not remain constant. Our estimate, based on a constant annual profit, is meant to illustrate orders of magnitude only.

As it *should* be. Because Amazon's future—and the future of *all* online retailers—depends crucially on the long-run market structure of the industry. But online retailing is so new—and the questions surrounding it are so profound—that long-run predictions at this stage can only be speculative.

Why should the future market structure of online retailing matter so much? Recall that in two of the four market structures you've learned about—perfect competition and monopolistic competition—new firms can easily enter the market, so competition reduces economic profit to zero in the long run. Thus, if the online retail industry is heading toward either perfect or monopolistic competition, the stocks of Amazon and other Web retailers are hopelessly overvalued.

On the other hand, if the industry is heading toward a monopoly or oligopoly structure, significant *barriers* to entry will keep out potential entrants. In these market structures, long-run profit is possible (although not guaranteed).

Thus, part of the instability observed in online retail stocks arises from two competing views of online retailing's future. On the one hand, there is the *long-run profits view*: that online retailing will end up as profitable oligopoly or monopoly markets. On the other hand, there is *the zero-profits view*: that online retailing will end up perfectly or monopolistically competitive. And the wind keeps changing direction. An investor leaning heavily toward one view can, rationally, begin tilting toward the other in a matter of weeks, days or even hours, based on some new information. And when people move *en masse* from one camp to the other, stock prices can swing dramatically.

This is shown in Figure 6. The vertical supply curve shows the number of shares outstanding for a Web retailer. The downward sloping demand curve labeled $D_{\text{pessimistic}}$ shows how many shares people would like to hold at each price if—based on recent information—investors tilt toward the zero-profits view. The demand

STOCK MARKET VALUATION IN ONLINE RETAILING

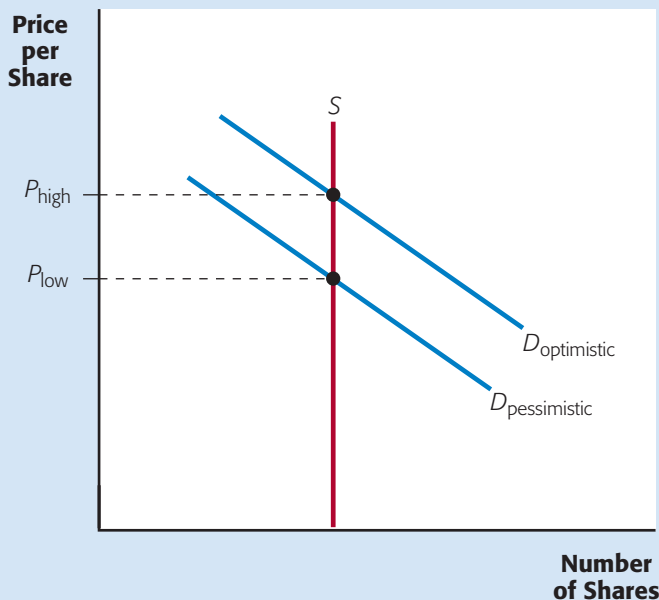


FIGURE 6

At a particular moment in time, the number of shares of any online retail firm is fixed, as illustrated by the vertical supply curve. The demand curve $D_{\text{pessimistic}}$ shows how many shares people would like to hold at each price if they tilt toward the zero-profits view. $D_{\text{optimistic}}$ shows demand if investors favor the long-run profits view. As new information becomes available, the demand curve may shift back and forth, so that the price per share oscillates between P_{low} and P_{high} .

curve labeled $D_{\text{optimistic}}$ shows the quantity of shares demanded when investors believe that the company will make profits in the long run. As new information becomes available, and the demand curve shifts back and forth, so does the equilibrium price of the stock.

What's behind each of these views about the future of the online retail industry? As you are about to see, both views are based on the microeconomic theory you've already learned.

THE LONG-RUN PROFITS VIEW

Adherents of the long-run profits view must believe there are barriers to entry in the online retail industry—barriers significant enough to create hugely profitable monopoly or oligopoly markets. You learned about barriers to entry in Chapters 9 and 10. Which of these barriers might be relevant to the future of Web retailers?

1. Economies of scale. Suppose that online retailers enjoy economies of scale until they are serving a large fraction of the market. In that case, there would be room for only a few firms in each retail market, since if there were many firms, each would have higher costs per unit.

Economies of scale may be particularly relevant in online retailing because of the costs of *lumpy inputs*. For example, most of the *information* costs of online retailing—the costs of developing and maintaining sophisticated Web sites, inventory tracking systems, and systems for organizing and tracking delivery of goods—are all lumpy costs. It costs just as much to maintain these information networks whether a firm sells 10 CDs per day or 10,000.

2. Reputation. Reputations certainly matter in traditional retailing. People are more likely to shop at stores that they know carry good merchandise, offer good service, and permit returns with minimum hassle. In addition, it takes time to find out about new retailers and to test them out, so people may shop at the same retailers again and again just because they know and feel comfortable with them.

In the long-run profits view, reputation may be even *more* important on the Internet, because customers must pay *before* the goods are received. When you buy on line, you need to have faith that the goods will arrive, that they will arrive in good condition, and that they will arrive when promised.

In this view, early entrants in an online retail market have a big *first-mover* advantage: They have already established their reputations. Newcomers will shy away because they'd have to pay for costly advertising campaigns and product giveaways in order to wrest customers from the early entrants. This is a cost that these early entrants have not had to bear.

If reputation is important in online retailing, then we'd expect those with the best reputations to charge higher prices for their retail services than their lesser-known competitors. And some early evidence suggests that this is indeed the case. In 1999, several studies found that the most recognized brands in online retailing—such as Amazon.com and CDNow—tended to sell identical goods for prices 7 to 12 percent higher than lesser-known retailers.⁶

⁶ See, for example, the studies cited in Michael Smith, Joseph Bailey, and Erik Brynjolfsson, "Understanding Digital Markets: Review and Assessment," July 1999. Draft available at <http://ecommerce.mit.edu/papers/ude>.

3. Protection of Intellectual Property. You might think that the protection of intellectual property through copyrights and patents has nothing to do with online retailers. After all, they don't *write* the books or *invent* any of the consumer products they sell. But in 1998, a U.S. Court of Appeals decision, upheld by the Supreme Court in 1999, permitted firms to patent *ways of doing business* over the Internet. For example, Amazon.com holds a patent on the process of ordering products with a single click of the mouse. It holds another patent on a basic method of rewarding affiliated sites for referring business to Amazon. In 1999, 700 such patents were awarded to online firms for ways of doing business on the Web. And under current law, each of these patents lasts for 20 years.

If a patent either lowers costs, or makes a retailer more attractive to customers, the patent's owner can make a profit without fearing that others will enter and eliminate that profit.

4. Other Barriers. In the long-run profits view, online retailing could benefit from some additional barriers to entry. One of these barriers is *network effects*—the benefit that accrues to one customer when *other* people become customers. For example, whenever you click on a book at Amazon's Web site, you are immediately informed that "other people who bought this book, also bought. . ." followed by a list of other books you might like. The more people who buy from Amazon, the larger the database that Amazon can use to come up with recommendations, and the more valuable the recommendations are. Thus, the first retailer to grow large will attract more and more customers, and will grow even larger. A new entrant would never have a chance in such a market.

Some online retailers go even further in recommending products—homing in on the tastes of individual customers based on their *past purchases*. A CD seller, for example, can offer a different home page—with different featured products—to each person who logs on, based on the type of music they have ordered in the past. In this way, the seller hopes to take advantage of *lock-in*—the special benefits that accrue to customers who keep coming back. These benefits are sacrificed when a customer switches to another firm. If the early online retailers can lock-in their customers, new firms will be hesitant to enter the industry.

If the long-run profits view is correct, and barriers to entry will keep out new entrants—and cause exit in markets that are already overcrowded—then the online retail industry could end up as a collection of oligopolies, one for each product.

But remember that oligopoly firms do not *necessarily* make economic profit. In order to be profitable, they must either have a monopoly on some aspect of their service that the other firms cannot copy, or else cooperate with their competitors in order to prevent costly price wars. (The airlines, for example, are oligopolies, but have never been able to cooperate long enough to give them significant profits.)

In the long-run profits view, the online retail industry satisfies both of these requirements for profitability. First, as we discussed, the leading Web sellers will have special patents—such as Amazon's one-click patent—that will prevent others from competing their profits away. Second, online retailing has features—such as easily observed prices—that may facilitate cooperation among firms and help prevent cheating.

In sum,

in the long-run profits view, online retail markets can be profitable—hugely profitable—due to barriers to entry. Only a few firms will survive in each retail market, and each survivor will have a monopoly on some valuable aspect of retail service, or be able to cooperate with its few competitors to boost markups and profits.

THE ZERO-PROFITS VIEW

In the zero-profits view, online retail markets are heading toward perfect or monopolistic competition in the long run. These are market structures in which free entry drives economic profit to zero. In this view, barriers to entry will either be nonexistent or insignificant. Let's look at each of the potential barriers to entry and the counterargument of the zero profits view.

Economies of Scale. While economies of scale are no doubt present in online retailing, no one yet knows at what point the minimum efficient scale (MES)—the share of the market at which cost per unit hits bottom—occurs. Moreover, online retail markets are huge, national markets in which every seller can sell to any buyer anywhere in the nation. With larger markets, the MES should occur at a much smaller *percentage* of the market than would occur in a traditional, local retail market.

For example, let's consider the market for books—one of the oldest and biggest online retail markets. Suppose that the MES in this market occurs when 1 million books are sold annually, and that cost per unit remains about the same for sales beyond this number. Suppose, too, that the online book market grows from 10 percent of total book sales in early 2000 to about 25 percent of total book sales by the time the market matures. Based on current totals, that would mean online purchases of about 100 million books. Under these assumptions, a seller that had only a 1 percent share of the market—1 million books—would have no cost disadvantage compared to one who had a 50 percent share—50 million books. So, based on economies of scale alone, there would be room for 100 different book-selling firms. In other words, in the zero-profits view, economies of scale are a weak argument for monopolies or oligopolies in the huge national retail markets created by the Internet.

Reputation. A good reputation can certainly give a firm an advantage. But newcomers can develop good reputations, too. And it may actually be *easier* for a new online retailer to develop a reputation than a new traditional retailer. For example, rating agencies—such as the Better Business Bureau—will soon go online, so information about a Web retailer will be just a click away. In addition, Web sites have ways of fostering good reputations rather quickly. They can set up communities (bulletin boards and chat rooms), or rely on good word of mouth on existing bulletin boards and chat rooms. They can buy links from other trusted Web sites. Moreover, in virtually every retail market, there are many *conventional* firms that *already* have good reputations. They may be able to go on line themselves, or “lend” their reputation to another firm through a partnership. For example, barnesandnoble.com—a spinoff of traditional retailer Barnes and Noble—entered the

bookselling market after Amazon had established its lead, but enjoyed the benefits of an instant reputation. Similarly Wal-Mart—a recent entrant in online retailing—enjoyed an instant reputation based on years of serving millions of customers across the country. Thus, in the zero-profits view, reputation is an unlikely barrier to the entry of new competitors.

Protection of Intellectual Property. The role of intellectual property on the Internet remains one of the big unknowns. In early 2000, the legal system was supporting long-term patents on ways of doing business, and this certainly furthered the long-run profits view. But these patents are becoming increasingly unpopular in the Internet community. So unpopular, in fact, that Jeff Bezos—the CEO and largest stockholder in Amazon.com—flipped his position on Web patents. In March 2000, he publicly argued for reducing the life of Internet patents to 3 or 4 years from the current 20 years. In the zero-profits view, the U.S. legal system—and legal systems around the world—will evolve to help keep e-commerce markets open and highly competitive. Patents will not be a significant barrier to entry.

Other Barriers. What about networks and lock-in? There is certainly value to buying from a seller who monitors your purchases, knows your tastes, and has a large enough information base to make recommendations on other products. But there is a countervailing force: the desire for privacy. Many people are uncomfortable when their purchases are monitored—a prerequisite for lock-in. And network effects may be countered by an individualist spirit that resents being told to buy products just because others did. Remember, too, that the ability of a retailer to know our tastes and make personal recommendations did not save the small general store from being replaced by huge mass merchants. (When was the last time any one of the sales staff recognized you, and recommended something, at your local Target, Tower, or Borders? Yet these stores thrived because they lured people from smaller stores with higher prices.) In the zero-profits view, network effects and lock-in are dubious foundations on which to build long-run economic profit.

Finally, there is one special feature of online retailing that strengthens the zero-profit view: the ease and speed of getting information on the Net. To see why, think about *traditional* shopping in the physical world where comparing prices is time consuming and troublesome. It involves transportation, waiting in line, and remembering information about prices and quality as you travel from one store to the other. Thus, in *traditional* retail markets, a store might be able to charge higher prices than other similar stores, without losing all of its customers.

Now think about buying on the Web. Comparison shopping there involves no travel time, just pointing and clicking. Moreover, powerful *shopbots* can make the comparisons for you. You just click on the good you are trying to buy, and the shopbot will report on a wide selection of sellers and prices. Two shopbots go even further: Clickthebutton.com keeps a button on your computer screen that you can push just before making a purchase, to check prices at competitors' Web sites. And R-U-Sure.com puts a program on your computer that *automatically* reports on competitors' prices every time you make a purchase. This suggests that in cyberspace, retailers will have a harder time charging more than their competitors, making economic profit even more unlikely.

In the zero-profits view, online retail markets will be unable to earn economic profit in the long run. Barriers to entry will not be high enough to keep out new entrants, and comparison shopping will be easy. As a result, online retail markets will most closely resemble monopolistic or perfect competition, with zero economic profit in the long run.

Interestingly, Amazon.com—a major proponent of the long-run profits view, recently seemed to support the zero-profits view. In an official filing with the Securities and Exchange Commission in August 1999, Amazon stated,

[C]ompetition in the Internet and online commerce markets probably will intensify. As various Internet market segments obtain large, loyal customer bases, participants in those segments may use their market power to expand into the markets in which we operate. In addition, new and expanded Web technologies may increase the competitive pressures on online retailers. For example, “shopping agent” technologies permit customers to quickly compare our prices with those of our competitors. This increased competition may reduce our operating margins, diminish our market share or impair the value of our brand.

Of course, Amazon has also released statements taking the opposite position. But if a leading firm seems to be on both sides of the fence, it is no surprise that stock market investors, trying to understand the future of an entirely new industry, are uncertain. Every new court decision, every release of sales figures, every entry of a new competitor or failure of an existing firm, causes radical shifts of opinion, and shifts the demand curve for online retail stocks, as in Figure 6 (p. 505). No doubt, the prices of these stocks will continue to fluctuate in value for years to come.

TIME FOR YOU TO USE THE THEORY

In this chapter, we’ve applied microeconomic theory—and especially the four-step process—to the online retail industry. We’ve explained some of the ways in which online retailing affects product markets, labor markets, and financial markets. But there are many more questions raised by this new industry—the impact on labor market discrimination, on economic efficiency, on international trade, on the distribution of income, and more.

Now it’s time for you to use the theory yourself. Take a look at the questions at the end of this chapter. See how many you can answer by applying the tools of microeconomics—and especially the four-step process.

PROBLEMS AND EXERCISES

1. In Figure 1, the development of online retailing shifts out the horizontal intercept of the PPF, but leaves the vertical intercept unaffected. Give examples of *other* Internet developments that are shifting out the vertical intercept of the PPF, and explain why they do so.
2. In Figure 2, one of the reasons for the rightward shift of the demand curve is an increase in tastes for ordering goods on line among those *already* connected to the Internet. Can you give some reasons for such a change in tastes? (*Hint*: Did you or your family begin ordering goods on the Internet the first month you were connected? The first year you were connected? What were the reasons for the delay?)
3. Figure 2 shows the market for traditional retail services in an initial long-run equilibrium, a new short-run equilibrium, and finally, a new long-run equilibrium. Show

- all three equilibria with a diagram for the typical online retail firm. You will need to draw ATC and MC curves, as well as the demand curve facing the firm. Assume that traditional retailing is an *increasing cost industry*.
4. Figure 2 shows that, as a result of online retailing, there will be fewer traditional retailers of CDs, each charging a lower markup in the future. Come up with a realistic story (a set of assumptions that differs from those behind Figure 2) that would lead to each of the following conclusions, and illustrate each story with a graph.
 - a. In the long run, there are fewer traditional retailers of CDs, each charging a *higher* markup than before.
 - b. In the long run, there are more traditional retailers of CDs, each charging a *higher* markup than before.
 5. Are there any current technological developments that might decrease the cost of providing *traditional* retail services? If so, what are they? How would they affect the analysis of traditional retailing in Figure 2?
 6. In panel (a) of Figure 3, the salary of Internet business attorneys in 2010 is higher than in 1999, but lower than in 2000. Show, using a similar diagram, that the salary cannot logically end up lower than in 1999, but that it could logically end up higher than in 2000. Assume there are no other changes affecting the salary, apart from those considered in the figure. (*Hint*: Imagine that online retailing undergoes a *huge* expansion between 2000 and 2010.)
 7. In panel (b) of Figure 3, the salary of traditional retail sales staff in 2010 is lower than in 1999, but higher than in 2000. Show, using a similar diagram, that the salary cannot logically end up higher than in 1999, but that it could logically end up lower than in 2000. Assume there are no other changes affecting the salary, apart from those considered in the figure. (*Hint*: Imagine that the traditional retail sector shrinks significantly between 2000 and 2010.)
 8. Figure 4 shows how the Internet works to widen the gap between highly skilled (college-educated) and less-skilled (high school educated) workers. How would increased government subsidies for college students affect both panels of Figure 4? Explain why college subsidies have an impact not only on the wage of those who *attend* college, but also on the wage of those who *don't* attend college.
 9. Some policy makers have called for an easing of immigration restrictions to help deal with labor market effects of rapidly changing technologies. For each of the following cases, explain how the graphical analysis in Figure 4, and the conclusion we reached from that figure, must be modified.
 - a. The government allows increased immigration among the highly skilled, but not among the less skilled.
 - b. The government allows increased immigration among the less skilled, but not among the highly skilled.

Would your answer in part (b) be affected if less-skilled immigrants also tend to patronize establishments (e.g., traditional retailers) that tend to employ less-skilled workers?
 10. Explain whether network effects, lock-in, or both are present in each of the following cases, and how they might work as a barrier to the entry or growth of new firms.
 - a. Frequent flier mileage awards by airlines
 - b. The time and trouble it takes to learn a new word processing program
 - c. AOL's efforts in 1999 to prevent outsiders from entering its chat rooms to communicate with AOL subscribers
 - d. Your decision to buy either an Apple or a Windows-based computer
 - e. A small, local video store, whose owner is getting to know the kind of movies you like, and the fact that you usually rent them on Thursday nights. She is starting to put certain videos aside for you on Thursdays in case you come in.
 11. Some economists have argued that online retailers will be able to *price discriminate* more easily than traditional retailers, while others have argued that price discrimination will eventually prove more difficult on the Web. Can you think of arguments to support each side of this debate? Does price discrimination imply anything about long-run profits?
 12. In Chapter 12, you learned about a variety of ways in which unfavored groups are discriminated against in labor markets. Which types of labor market discrimination are weakened, and which types are strengthened, by the rise of online retailing?
 13. Which type of retailing—on line or traditional—creates more negative externalities? Give examples to support your position.
 14. “Online retailers should have a much easier time selling abroad than traditional retailers do.” Suppose this is true, and online retail markets become *international* markets. Does this strengthen the long-run profits view or the zero-profits view?

C H A L L E N G E Q U E S T I O N S

1. Suppose a potential entrant in an online retail market can anticipate three years of economic losses equal to \$300 million, \$500 million, and \$200 million, respectively, and then a constant annual economic profit beginning in the fourth year and continuing forever. Assuming the appropriate discount rate is 10 percent, what is the minimum constant annual economic profit that would induce the firm to enter the industry? (*Hint:* Use the formula, given in Chapter 13, for determining the total present value of a constant stream of payments beginning this year. But be sure to deduct the total present value of the first three years of missing profit, and then subtract the total present value of the first three years' losses.)
2. Traditional retailers have been lobbying to apply the sales tax—which they have to pay—to online retail sales (which by early 2000 remained mostly free from taxation). If the traditional retailers are successful, how would this change the analysis in Figure 2? Be sure to state any assumptions you are making in your analysis.
3. Some economists believe that the United States has a comparative advantage in providing online retail services. What could account for this view?
4. Is the growth of online retailing likely to increase or decrease the returns to “superstars” discussed in Chapter 12? Explain

WHAT MACROECONOMICS TRIES TO EXPLAIN

You have no doubt seen photographs of the earth taken from satellites thousands of miles away. Viewed from that great distance, the world's vast oceans look like puddles, its continents like mounds of dirt, and its mountain ranges like wrinkles on a bedspread. In contrast to our customary view from the earth's surface—of a car, a tree, a building—this is a view of the big picture.

What, you may be wondering, could this possibly have to do with economics? Actually, quite a bit: These two different ways of viewing the earth—from up close or from thousands of miles away—are analogous to two different ways of viewing the economy. When we look through the *microeconomic* lens—from up close—we see the behavior of *individual decision makers* and *individual markets*. When we look through the *macroeconomic* lens—from a distance—these smaller features fade away, and we see only the broad outlines of the economy.

Which view is better? That depends on what we're trying to do. If we want to know why rents are so high in big cities, why computers are getting better and cheaper each year, or why the earnings of anesthesiologists are falling, we need the close-up view of microeconomics. But to answer questions about the *overall* economy—what determines the amount of unemployment, how fast the average standard of living will rise over the next decade, or how fast prices will rise—we need the more comprehensive view of *macroeconomics*.

MACROECONOMIC GOALS

While there is some disagreement among economists about *how* to make the macroeconomy perform well, there is widespread agreement about the goals we are trying to achieve:

Economists—and society at large—agree on three important macroeconomic goals: rapid economic growth, full employment, and stable prices.

Why is there such universal agreement on these three goals? Because achieving them gives us the opportunity to make *all* of our citizens better off. Let's take a closer look at each of these goals and see why they are so important.

CHAPTER OUTLINE

Macroeconomic Goals

- Rapid Economic Growth
- High Employment
- Stable Prices

The Macroeconomic Approach

- Aggregation in Macroeconomics

Macroeconomic Controversies

As You Study

- Macroeconomics . . .

RAPID ECONOMIC GROWTH

Imagine that you were a typical American worker living at the beginning of the twentieth century. You would work about 60 hours every week, and your yearly salary—about \$450—would buy a bit less than \$8,000 would buy today. You could expect to die at the age of 47. If you fell seriously ill before then, your doctor wouldn't be able to help much: There were no X-ray machines or blood tests, and little effective medicine for the few diseases that could be diagnosed. You would probably never hear the sounds produced by the best musicians of the day, or see the performances of the best actors, dancers, or singers. And the most exotic travel you'd enjoy would likely be a trip to a nearby state.

Today, the typical worker has it considerably better. He or she works about 35 hours per week, and is paid about \$31,000 per year, not to mention fringe benefits such as health insurance, retirement benefits, and paid vacation. Thanks to advances in medicine, nutrition, and hygiene, the average man can expect to live to age 73, and the average woman to age 80. And more of a worker's free time today is really free: There are machines to do laundry and dishes, cars to get to and from work, telephones for quick communication, and—increasingly—personal computers to keep track of finances, appointments, and correspondence. Finally, during their lifetimes, most Americans will have traveled—for enjoyment—to many locations in the United States and abroad.

What is responsible for these dramatic changes in economic well-being? The answer is three words: *rapid economic growth*. In the United States—as in most developed economies—our output of goods and services has risen faster than the population. As a result, the average person can consume much more today—more food, clothing, housing, medical care, entertainment, and travel—than in the year 1900.

Economists monitor economic growth by keeping track of *real gross domestic product (real GDP)*—the total quantity of goods and services produced in a country over a year. When real GDP rises faster than the population, output per person rises, and so does the average standard of living.

Figure 1 shows real GDP in the United States from 1920 to 1999, measured in dollars of output at 1996 prices. As you can see, real GDP has increased dramatically over the greater part of the century. Part of the reason for the rise is an increase in population: More workers can produce more output. But real GDP has actually increased *faster* than the population: During this period, while the U.S. population did not quite triple, the quantity of goods and services produced each year has increased more than tenfold. Hence, the remarkable rise in the average American's living standard.

But when we look more closely at the data, we discover something important: Although output has grown, the *rate* of growth has varied over long periods of time. From 1959 to 1973, output per person grew, on average, by 4.1 percent per year. But from 1973 to 1991, average annual growth slowed to 2.7 percent. Then, from 1991 to 1999, growth picked up again, averaging 3.6 percent per year. These may seem like slight differences. But over long periods of time, such small differences in growth rates can cause huge differences in living standards. For example, suppose that for the entire period from 1973 to 1999, output per person had grown at its previous pace of 4.1 percent per year, instead of its actual rate. Then we'd have produced more output in *each* of those 26 years. By 1999, our real annual GDP would have been \$11,578 billion instead of \$8,861 billion. That increase in GDP would have been enough to give every man, woman, and child in the country an additional \$10,000 in goods and services from that year's production alone.

U.S. REAL GROSS DOMESTIC PRODUCT, 1920–1999

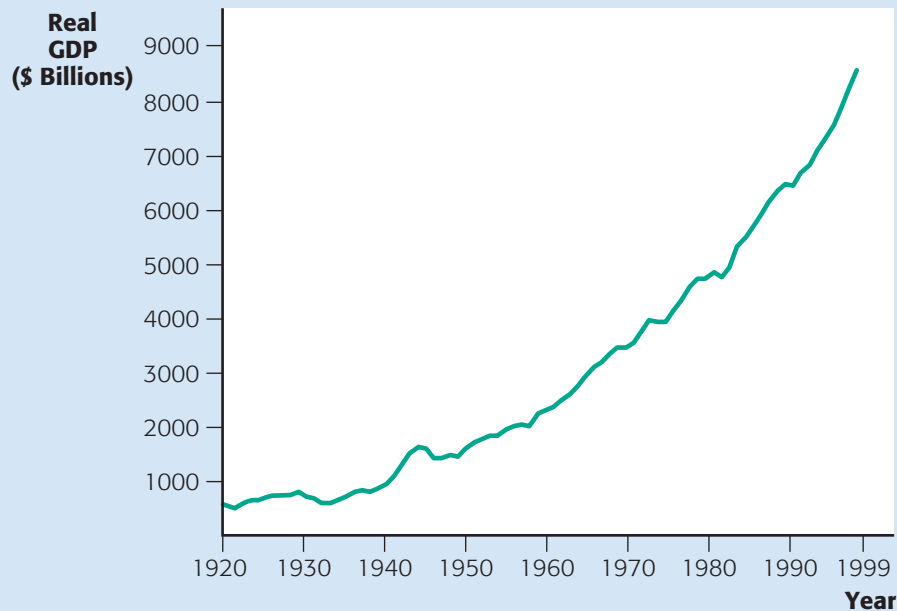


FIGURE 1

Real GDP has increased dramatically over the past 80 years. In the figure, real GDP is measured in dollars of output valued at 1996 prices. (The measurement of real GDP will be discussed in more detail in the next two chapters.)

Economists and government officials are very concerned when economic growth slows down. Growth increases the size of the economic pie, so it becomes possible—at least in principle—for every citizen to have a larger slice. This is why economists agree that growth is a good thing.

But in practice, growth does *not* benefit everyone. Living standards will always rise more rapidly for some groups than for others, and some may even find their slice of the pie shrinking. For example, since the late 1980s, economic growth has improved the living standards of the highly skilled, while less-skilled workers have actually become worse off. Partly, this is due to improvements in technology that have lowered the earnings of workers whose roles can be taken by computers and machines. But very few economists would advocate a halt to growth as a solution to the problems of unskilled workers. Some believe that, in the long run, everyone will indeed benefit from growth. Others see a role for the government in taxing successful people and providing benefits to those left behind by growth. But in either case, economic growth—by increasing the size of the overall pie—is seen as an important part of the solution.

HIGH EMPLOYMENT

Economic growth is one of our most important goals, but not the only one. Suppose our real GDP were growing at, say, a 3 percent annual rate, but 10 percent of the workforce was unable to find work. Although the economy would be growing at a healthy pace, we would not be achieving our full economic potential—our average standard of living would not be as high as it *could be*. There would be millions of people who wanted jobs, who *could* be producing output we could all use, but who would not be producing anything. This is one reason why consistently

high employment—or consistently *low unemployment*—is an important macroeconomic goal.

But there is another reason, too. In addition to its impact on our average standard of living, unemployment also affects the distribution of economic well-being among our citizens. People who cannot find jobs suffer. Their incomes, and their ability to buy goods and services, decrease. And even though many of the jobless receive unemployment benefits and other assistance from the government, the unemployed typically have lower living standards than the employed.

One measure economists use to keep track of employment is the *unemployment rate*, which is the percentage of the workforce that would like to work, but cannot find jobs. Figure 2 shows the average unemployment rate during each of the past 80 years. Notice that the unemployment rate is never zero—there are always *some* people looking for work, even when the economy is doing well. But in some years, unemployment is unusually high. The worst example occurred during the Great Depression of the 1930s, when millions of workers lost their jobs and the unemployment rate reached 25 percent. One in four potential workers could not find a job. More recently, in 1982 and 1983, the unemployment rate averaged almost 10 percent.

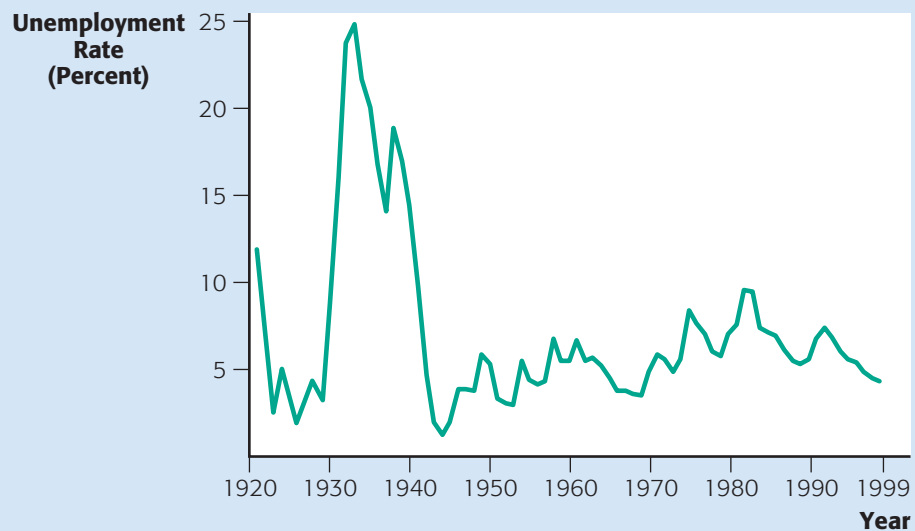
The nation's commitment to high employment has twice been written into law. With the memory of the Great Depression still fresh, Congress passed the *Employment Act of 1946*, which required the federal government to “promote maximum employment, production, and purchasing power.” It did not, however, dictate a target rate of unemployment the government should aim for. A numerical target was added in 1978, when Congress passed the *Full Employment and Balanced Growth Act*, which called for an unemployment rate of 4 percent.

A glance at Figure 2 shows how seldom we have hit this target over the last few decades. In fact, we did not hit it at all through the 1970s and 1980s. But in the 1990s, we came closer and closer and finally—in January 2000—we reached the target again for the first time since the 1960s. In future chapters, you will learn why

FIGURE 2

The unemployment rate fluctuates over time. During the Great Depression of the 1930s, unemployment was extremely high, reaching 25 percent in 1933. In the early 1980s, the rate averaged 10 percent.

U.S. UNEMPLOYMENT RATE, 1920–1999



THE BUSINESS CYCLE

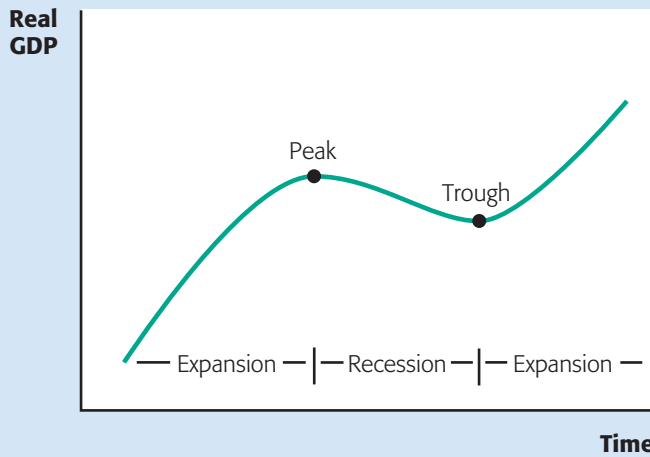


FIGURE 3

Over time, real GDP fluctuates around an overall upward trend. Such fluctuations are called *business cycles*. When output rises, we are in the expansion phase of the cycle; when output falls, we are in a *recession*.

the unemployment rate has often been higher than its target, why we were able to hit the target in January 2000, and—more generally—what the government can and cannot do to achieve its goal of low unemployment.

Employment and the Business Cycle. When firms produce more output, they hire more workers; when they produce less output, they tend to lay off workers. We would thus expect real GDP and employment to be closely related, and indeed they are. In recent years, each 1 percent drop in output has been associated with the loss of about half a million jobs. Consistently high employment, then, requires a high, stable level of output. Unfortunately, output has *not* been very stable. If you look back at Figure 1, you will see that while real GDP has climbed upward over time, it has been a bumpy ride. The periodic fluctuations in GDP—the bumps in the figure—are called **business cycles**.

Figure 3 shows a close-up view of a hypothetical business cycle. When output rises, we are in the **expansion** phase, which continues until we reach a **peak**. Then, as output falls, we enter a **recession**—a period of declining output. When output hits bottom, we are in the **trough** of the recession.

Of course, real-world business cycles never look quite like the smooth, symmetrical cycle in Figure 3, but rather like the jagged, irregular cycles of Figure 1. Recessions can be severe or mild, and they can last several years or less than a single year. When a recession is particularly severe and long lasting, it is called a **depression**. In the twentieth century, the United States experienced just one decline in output serious enough to be considered a depression—the worldwide *Great Depression* of the 1930s. From 1929 to 1933, the first four years of the Great Depression, U.S. output dropped by more than 25 percent.

But even during more normal times, the economy has gone through many recessions. Since 1959, we have suffered through two severe recessions (in 1974–75 and 1981–82) and several less severe ones, such as the recession of 1990 to 1991. Later in this book, you will learn about some of the causes of recessions, why we have not been able to eliminate them entirely, and what we *may* be able to do to make them milder in the future.

Business cycles Fluctuations in real GDP around its long-term growth trend.

Expansion A period of increasing real GDP.

Peak The point at which real GDP reaches its highest level during an expansion.

Recession A period of declining or abnormally low real GDP.

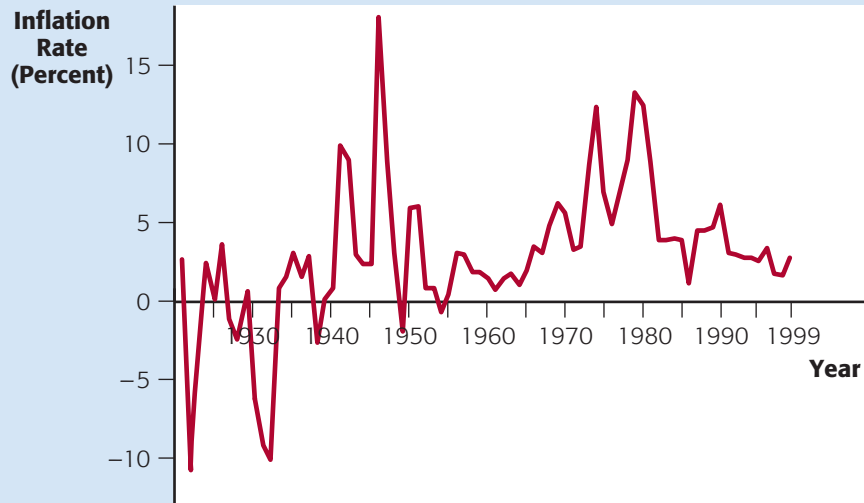
Trough The point at which real GDP reaches its lowest level during a recession.

Depression An unusually severe recession.

FIGURE 4

In most years, the inflation rate has been positive. The overall price level increased during those years.

U.S. ANNUAL INFLATION RATE, 1922–1999



STABLE PRICES

Figure 4 shows the annual inflation rate—the percentage increase in the average level of prices—from 1922 to 1999.¹ With very few exceptions, the inflation rate has been positive—on average, prices have risen in each of those years. But notice the wide variations in inflation. In 1979 and 1980, we had double-digit inflation—prices rose by more than 12 percent in both years. During that time, polls showed that people were more concerned about inflation than any other national problem—more than unemployment, crime, poverty, pollution, or anything else. During the 1990s, the inflation rate averaged less than 3 percent per year, and it has stayed low through early 2000. As a result, we hardly seem to notice it at all. Pollsters no longer include “rising prices” as a category when asking about the most important problems facing the country.

Other countries have not been so lucky. In the 1980s, several Latin American nations experienced inflation rates of thousands of percent per year. In the early 1990s, some of the newly emerging nations of Central Europe and the former Soviet Union suffered annual inflation rates in the triple digits. An extreme case was the new nation of Serbia, where prices rose by 1,880 percent in the single month of August 1993. If prices had continued to rise at that rate all year, the annual inflation rate would have been 363,000,000,000,000,000 percent.

Why are stable prices—a low inflation rate—an important macroeconomic goal? Because inflation is *costly* to society. With annual inflation rates in the thousands of percent, the costs are easy to see: The value of the currency—its purchasing power—declines so rapidly that people are no longer willing to hold it. This breakdown of the monetary system forces people to waste valuable time and resources bartering with each other—for example, trading plumbing services for dentistry services. With so much time spent trying to find trading partners, there is little time left for producing goods and services. As a result, the average standard of living falls.

¹ The figure is based on the Consumer Price Index, the most popular measure of the price level, as well as historical estimates of what this index *would* have been in the early part of the twentieth century, before the index existed. We’ll discuss the Consumer Price Index and other measures of inflation in more detail in later chapters.

With more modest inflation, like the double-digit rates the United States experienced in the late 1970s, the costs to society are less obvious and less severe. But they are still significant. And when it comes time to bring down even a modest inflation rate, painful corrective actions by government are required. These actions cause output to decline and unemployment to rise. For example, in order to bring the inflation rate down from the high levels of the early 1980s (see Figure 4), government policy purposely caused a severe recession in 1981–82, reducing output (Figure 1) and increasing unemployment (Figure 2).

The previous paragraph raises a number of questions. How, precisely, does a modest inflation harm society? Why would a recession reduce inflation? And how does the government create a recession? If you're a bit confused, don't worry. You are just beginning your study of macroeconomics, and we have a lot of ground to cover.

THE MACROECONOMIC APPROACH

If you have already studied *microeconomics*, you will notice much that is familiar in *macroeconomics*. The *four-step procedure* plays an important role in both branches of the field. But the macroeconomic approach is different from the microeconomic approach in significant ways. Most importantly, in *microeconomics*, we typically apply our 4 Key Steps to *one market at a time*—the market for soybeans, for neurosurgeons, or for car washes. In *macroeconomics*, by contrast, we want to understand how the entire economy behaves. Thus, we will be applying the key steps to *all markets simultaneously*. This includes not only markets for goods and services, but also markets for labor and for financial assets like bonds and foreign currency.

How can we possibly hope to deal with all of these markets at the same time? One way would be to build a gigantic model that included every individual market in the economy. The model would have tens of thousands of supply and demand curves, which could be used to determine tens of thousands of prices and quantities. With today's fast, powerful computers, we could, in principle, build this kind of model.

But it would not be easy. We would need to gather data on every good and service in the economy, every type of labor, every type of financial asset, and so on. As you might guess, this would be a formidable task, requiring thousands of workers just to gather the data alone. And in the end, the model would not prove very useful. We would not learn much about the economy from it: With so many individual trees, we could not see the forest. Moreover, the model's predictions would be highly suspect: With so much information and so many moving parts, high standards of accuracy are difficult to maintain. Even the government of the former Soviet Union, which directed production throughout the economy until the 1990s, was unable to keep track of all the markets under its control. In a market economy, where production decisions are made by individual firms, the task would be even harder.

What, then, is a macroeconomist to do? The answer is a word that you will become very familiar with in the chapters to come: **aggregation**—the process of combining different things into a single category and treating them as a whole. Let's take a closer look at how aggregation is used in macroeconomics.

Aggregation The process of combining different things into a single category.

AGGREGATION IN MACROECONOMICS

Aggregation is a basic tool of reasoning, one that you often use without being aware of it. If you say, "I applied for five jobs last month," you are aggregating five very different workplaces into the single category, *jobs*. Whenever you say, "I'm going out with my friends," you are combining several different people into



In many English words, the prefix *macro* means “large” and *micro* means “small.” As a result, you might think that in microeconomics, we study economic units in which small sums of money are involved, while in macroeconomics we study units involving greater sums. But this is not correct: The annual output of General Motors is considerably greater than the total annual output of many small countries, such as Estonia or Guatemala. Yet when we study the behavior of General Motors, we are practicing *microeconomics*, and when we study the causes of unemployment in Estonia, we are practicing *macroeconomics*. Why? Microeconomics is concerned with the behavior and interaction of *individual* firms and markets, even if they are very large; macroeconomics is concerned with the behavior of *entire economies*, even if they are very small.

a single category: people you consider *friends*.

Aggregation plays a key role in both micro- and macroeconomics. Microeconomists will speak of the market for automobiles, lumping Toyotas, Fords, BMWs, and other types of cars into a single category. But in macroeconomics, we take aggregation to the extreme. Because we want to consider the entire economy at once, and yet keep our model

as simple as possible, we must aggregate all markets into the broadest possible categories. For example, we lump together all the millions of different goods and services—computers, coffee tables, egg rolls, newspapers—into the single category, *output*. Similarly, we combine the thousands of different types of workers in the economy—doctors, construction workers, plumbers, college professors—into the category, *labor*. By aggregating in this way, we can create workable and reasonably accurate models that teach us a great deal about how the overall economy operates.

MACROECONOMIC CONTROVERSIES

Macroeconomics is full of disputes and disagreements. Indeed, modern macroeconomics—which began with the publication of *The General Theory of Employment, Interest, and Money*, by British economist John Maynard Keynes in 1936—originated in controversy. Keynes was taking on the conventional wisdom of his time—*classical economics*—which held that the macroeconomy worked very well on its own, and the best policy for the government to follow was *laissez faire*—“leave it alone.” As he was working on *The General Theory*, Keynes wrote to his friend, the playwright George Bernard Shaw, “I believe myself to be writing a book on economic theory which will largely revolutionize—not, I suppose, at once but in the course of the next ten years—the way the world thinks about economic problems.” Keynes’s prediction was on the money. After the publication of his book, economists argued about its merits, but 10 years later, the majority of the profession had been won over; they had become Keynesians. This new school of thought held that the economy does *not* do well on its own (one needed only to look at the Great Depression for evidence) and requires continual guidance from an activist and well-intentioned government.

From the late 1940s until the early 1960s, events seemed to prove the Keynesians correct. Then, beginning in the 1960s, several distinguished economists began to challenge Keynesian ideas. Their counterrevolutionary views—which in many ways mirrored those of the classical economists—were strengthened by events in the 1970s, when the economy’s behavior began to contradict the most important Keynesian ideas. While some of the early disagreements have been resolved, others have arisen to take their place.

Some of today’s controversies are purely *positive* in nature. For example, in a later chapter you will learn about the Federal Reserve System—the central bank in the United States—which can influence many important macroeconomic aggregates, such as output, employment, and the inflation rate. As this is being written (early 2000), most economists believe that the Federal Reserve (or “Fed”) is doing an excellent job

of managing these aggregates. They point out that the Fed has successfully engineered high employment and rapid economic growth for almost a decade, without overheating the economy and risking future inflation. A few economists, however, think that the Fed has made a mistake. They believe that it has indeed been overheating the economy, which will lead to higher inflation in the future. (Are you confused about the connections between rapid economic growth, overheating the economy, and future inflation? Don't worry: All of this will be explained in later chapters.)

To some extent, this is a *positive* disagreement: The two sides have different views about how the economy is performing now, and what that performance implies about the future. That is, it's in part a disagreement about *how the economy works*.

But for some, the controversy is also *normative*. We might find two economists who agree about the extent to which the Fed's current policies are risking future inflation. But they might disagree strongly about the wisdom of the gamble because of differences in *values*. One economist may place more weight on high employment and rapid growth, and may be willing to risk future inflation to achieve them. The other might put more weight on avoiding future inflation, even if it means lower employment and slower growth now.

Economists, like all other human beings, hold different values, and often hold them strongly. Not surprisingly, disagreements among economists are often emotionally charged. But there is also more agreement than meets the eye. Macroeconomists agree on many basic principles, and we will stress these as we go. And even when there are strong disagreements, there is surprising consensus on the approach that should be taken to resolve them.

AS YOU STUDY MACROECONOMICS . . .

Macroeconomics is a fascinating and wide-ranging subject. You will find that each piece of the macroeconomic puzzle connects to all of the other pieces in many different ways. Each time one of your questions is answered, 10 more will spring up in your mind, each demanding immediate attention. This presents a problem for a textbook writer, and for your instructor as well: What is the best order to present the principles of macroeconomics? We could follow the line of questions that occur to the curious reader, but this would be an organizational disaster. For example, learning about unemployment raises questions about international trade, but it also raises questions about government spending, government regulations, economic growth, wages, banking, and much, much more. And each of these topics raises questions about still others. Organizing the material in this way would make you feel like a ball in a pinball machine, bouncing from bumper to bumper. Still, the pinball approach—bouncing from topic to topic—is the one taken by the media when reporting on the economy. If you have ever tried to learn economics from a newspaper, you know how frustrating this approach can be.

In our study of macroeconomics, we will follow a different approach: presenting material as it is *needed* for what follows. In this way, what you learn in one chapter will form the foundation for the material in the next, and your understanding of macroeconomics will deepen as you go.

But be forewarned: This approach requires considerable patience on your part. Many of the questions that will pop into your head will have to be postponed until the proper foundations for answering them have been established. It might help, though, to give you a *brief* indication of what is to come.

In the next two chapters, we will discuss three of the most important aggregates in macroeconomics: output, employment, and the price level. You will see why each



Two excellent print sources for news on the U.S. and world economies are *The Wall Street Journal* and *The Economist*, a British Magazine.

of these is important to our economic well-being, how we keep track of them with government statistics, and how to interpret these statistics with a critical eye.

Then, in the remainder of the book, we study how the macroeconomy operates, starting with its behavior in the long run. Here, you will learn what makes an economy grow over long periods of time, and which government policies are likely to help or hinder that growth.

Then, we turn our attention to the short run. You will learn why the economy behaves differently in the short run than in the long run, why we have business cycles, and how these cycles may be affected by government policies. Then we'll expand our analysis to include the banking system and the money supply, and the special challenges they pose for government policy makers.

Finally, we'll turn our attention to the special problems of a global economy. You'll learn how trade with other nations constrains and expands our macro policy options at home and how economic events abroad influence our own economy. You will also learn why the United States has run persistent trade deficits with the rest of the world and what that means for our citizens.

This sounds like quite a lot of ground to cover, and indeed, it is. But it's not as daunting as it might sound. Remember that the study of macroeconomics—like the macroeconomy itself—is not a series of separate units, but an integrated whole. As you go from chapter to chapter, each principle you learn is a stepping-stone to the next one. Little by little, your knowledge and understanding will accumulate and deepen. Most students are genuinely surprised at how well they understand the macroeconomy after a single introductory course, and find the reward well worth the effort.

S U M M A R Y

Macroeconomics is the study of the economy as a whole. It deals with issues such as economic growth, unemployment, inflation, and government policies that might influence the overall level of economic activity.

Economists generally agree about the importance of three main macroeconomic goals. The first of these is rapid economic growth. If output—real gross domestic product—grows faster than population, the average person can enjoy an improved standard of living.

High employment is another important goal. In the United States and other market economies, the main source of

households' incomes is labor earnings. When unemployment is high, many people are without jobs and must cut back their purchases of goods and services.

The third macroeconomic goal is stable prices. This goal is important because inflation imposes costs on society. Keeping the rate of inflation low helps to reduce these costs.

Because an economy like that of the United States is so large and complex, the models we use to analyze the economy must be highly aggregated. For example, we will lump together millions of different goods to create an aggregate called “output” and combine all their prices into a single “price index.”

K E Y T E R M S

business cycles
expansion

peak
recession

trough
depression

aggregation

R E V I E W Q U E S T I O N S

1. Discuss the similarities and differences between macroeconomics and microeconomics.
2. What is the basic tool macroeconomists use to deal with the complexity and variety of economic markets and institutions? Give some examples of how they use this tool.
3. List the nation's macroeconomic goals and explain why each is important.
4. Consider an economy whose real GDP is growing at 4 percent per year. What else would you need to know in order to say whether the average standard of living is improving or deteriorating?

CHALLENGE QUESTION

Speculate about some factors that might help explain the post-1973 growth slowdown. What changes in the economy or in society as a whole may have contributed to

this phenomenon? Why might growth have speeded up again in the late 1990s?

EXPERIENTIAL EXERCISES

1. Which of the three macroeconomic goals mentioned in this chapter do you think is the most important today? Use *The Wall Street Journal* or Infotrac to support your conclusions. Do this by finding several recent articles that mention rapid growth, high employment, and stable prices. Then point to the emphasis each goal receives in these articles.
2. The Index of Leading Economic Indicators is an economic statistic that is sometimes used to predict how the economy will behave over the next several months. You can find it at <http://www.conference-board.org/products/frames.cfm?main=lei1.cfm>. Read up on the Index and see if you can explain how some of the individual components are related to overall economic performance.





CHAPTER

18

PRODUCTION, INCOME, AND EMPLOYMENT

CHAPTER OUTLINE

Production and Gross Domestic Product

- GDP: A Definition
- The Expenditure Approach to GDP
- Other Approaches to GDP
- Measuring GDP: A Summary
- Real Versus Nominal GDP
- How GDP Is Used
- Problems with GDP

Employment and Unemployment

- Types of Unemployment
- The Costs of Unemployment
- How Unemployment Is Measured
- Problems in Measuring Unemployment

Using the Theory: Society's Choice of GDP

On the first Friday of every month, at 8:00 A.M., dozens of journalists mill about in a room in the Department of Labor. They are waiting for the arrival of the press officer from the government's Bureau of Labor Statistics. When she enters the room, carrying a stack of papers, the buzz of conversation stops. The papers—which she passes out to the waiting journalists—contain the monthly report on the experience of the American workforce. They summarize everything the government knows about hiring and firing at businesses across the country; about the number of people working, the hours they worked, and the incomes they earned; and about the number of people *not* working and what they did instead. All of this information is broken down by industry, state, city, race, sex, and age. But one number looms large in the journalists' minds as they scan the report and compose their stories: the percentage of the labor force that could not find jobs, or the nation's *unemployment rate*.

Once every three months, a similar scene takes place at the Department of Commerce, as reporters wait for the release of the quarterly report on the nation's output of goods and services and the incomes we have earned from producing it. Once again, the report includes tremendous detail. Output is broken down by industry and by the sector that purchased it (ordinary households, businesses, government agencies, and foreigners), and income is broken down into the different types of earners—wage earners, property owners, and owners of small businesses. And once again, the reporters' eyes will focus on a single number, a number that will dominate their stories and create headlines in newspapers across the country: the nation's *gross domestic product*.

The government knows that its reports on employment and production will have a major impact on the American political scene, and on financial markets in the United States and around the world. So it takes great pains to ensure fair and equal access to the information. For example, the Bureau of Labor Statistics allows journalists to look at the employment report at 8:00 A.M. on the day of the release (the first Friday of every month). But all who see the report must stay inside a room—appropriately called the lockup room—and cannot contact the outside world until the official release time of 8:30 A.M. At precisely 8:29 A.M., the reporters are permitted to hook up their laptop modems, and then a countdown begins, ending at precisely 8:30 A.M. At that moment—and not a second before—the reporters

are permitted to transmit their stories. At the same instant, the Bureau posts its report on an Internet Web site. (The URL is <http://stats.bls.gov/blshome.html>.)

The reactions to the government's reports come almost immediately. Within seconds, wire-service headlines appear on computer screens across the country—"Unemployment Rate Up Two-Tenths of a Percent" or "Nation's Production Steady." Within minutes, financial traders, regarding these news flashes as clues about the economy's future, make snap decisions to buy or sell, and prices move in the stock and bond markets. This creates further headlines—"Stock Market Plunges on Unemployment Data" or "Bonds Rally on Output Report." Within the hour, politicians and pundits will respond with sound bites, attacking or defending the administration's economic policies.

Why is so much attention given to the government's reports on production and employment, and—in particular—to those two numbers: gross domestic product and the unemployment rate? Because they describe aspects of the economy that dramatically affect each of us individually and our society as a whole. In this chapter, we will take our first look at production and employment in the economy. The purpose here is not to explain what causes these variables to rise or fall—that will come a few chapters later, when we begin to study macroeconomic models. Here, we will focus on the reality behind the numbers: what the statistics tell us about the economy, how the government obtains them, and how they are sometimes misused.

PRODUCTION AND GROSS DOMESTIC PRODUCT

You have probably heard the phrase *gross domestic product*—or its more familiar abbreviation, GDP—many times. It is one of those economic terms that is frequently used by the media and by politicians. In the first part of this chapter, we take a close look at GDP, starting with a careful definition.

GDP: A DEFINITION

The U.S. government has been measuring the nation's total production since the 1930s. You might think that this is an easy number to calculate, at least in theory: Simply add up the output of every firm in the country during the year. Unfortunately, measuring total production is not so straightforward, and there are many conceptual traps and pitfalls. This is why economists have come up with a very precise definition of GDP.

The nation's gross domestic product (GDP) is the total value of all final goods and services produced for the marketplace during a given year, within the nation's borders.

Gross Domestic Product (GDP) The total value of all final goods and services produced for the marketplace during a given year, within the nation's borders.

Quite a mouthful. Is everything in this definition really necessary? Absolutely. To see why, let's break the definition down into pieces and look more closely at each one.

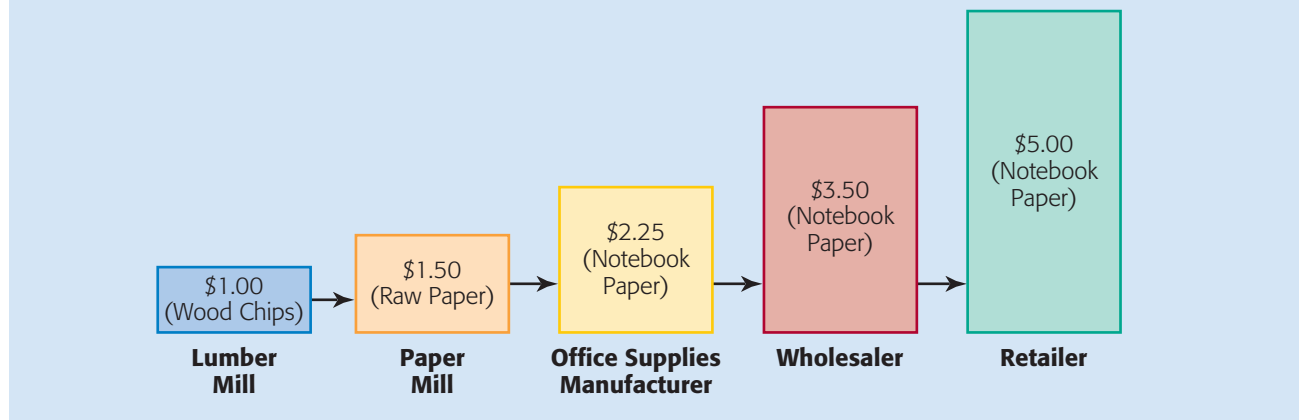
The total value . . .

An old expression tells us that "you can't add apples and oranges." But that is just what government statisticians must do when they measure our total output. In a typical day, American firms produce millions of *loaves* of bread, thousands of *pounds* of peanut butter, hundreds of *hours* of television programming, and so on. These are *different* products, and each is measured in its own type of units. Yet, somehow, we must combine all of them into a single number. But how?

The approach of GDP is to add up the *dollar value* of every good or service—the number of dollars each product is *sold* for. As a result, GDP is measured in

FIGURE 1

STAGES OF PRODUCTION



dollar units. For example, in 1999, the GDP of the United States was about \$9,248,000,000,000—give or take a few billion dollars. (That’s about \$9.2 trillion.)

Using dollar values has two important advantages. First, it gives us a common unit of measurement for very different things, thus allowing us to add up “apples and oranges.” Second, it ensures that more valuable goods (like a hundred computer chips) will count more in our GDP than less valuable ones (a hundred tortilla chips).

... of all final ...

When measuring production, we do not count *every* good or service produced in the economy, but only those that are sold to their *final users*. An example will illustrate why.

Figure 1 shows a simplified version of the stages of production needed to produce a ream (500 sheets) of notebook paper: A lumber company cuts down trees and produces wood chips, which it sells to a paper mill for \$1.00. The mill cooks, bleaches, and refines the wood chips, turning them into paper rolls, which it sells to an office supplies manufacturer for \$1.50. This manufacturer cuts the paper, prints lines and margins on it, and sells it to a wholesaler for \$2.25. The wholesaler sells it to a retail store for \$3.50, and then, finally, it is sold to a consumer—perhaps you—for \$5.00.

Should we add the value of *all* this production, and include $\$1.00 + \$1.50 + \$2.25 + \$3.50 + \$5.00 = \13.25 in GDP each time a ream of notebook paper is produced? No, this would clearly be a mistake, since all of this production ends up creating a good worth only \$5 in the end. In fact, the \$5 you pay for this good already *includes* the value of all the other production in the process.

In our example, the goods sold by the lumber company, paper mill, office supplies manufacturer, and wholesaler are all **intermediate goods**—goods used up in the process of producing something else. But the retailer (say, your local stationery store) sells a **final good**—a product sold to its *final user* (you). If we separately added in the production of intermediate goods when calculating GDP, we would be counting them more than once, since they are already included in the value of the final good.

Intermediate goods Goods used up in producing final goods.

Final good A good sold to its final user.

To avoid overcounting intermediate products when measuring GDP, we add up the value of final goods and services only. The value of all intermediate products is automatically included in the value of the final products they are used to create.

... *goods and services* ...

We all know a good when we see one: We can look at it, feel it, weigh it, and, in some cases, eat it, strum it, or swing a bat at it. Not so with a service: When you get a medical checkup, a haircut, or a car wash, the *effects* of the service may linger, but the service itself is used up the moment it is produced. Nonetheless, final services are as much a part of our GDP as are final goods. The services we are talking about include Internet access and the U.S. Navy, which you many not think of as services when you first hear the word.

Services have become an increasingly important part of our total output in recent decades. The service sector has grown from 31 percent of total output in 1950 to more than half of total output in 1999.

... *produced* ...

In order to contribute to GDP, something must be *produced*. This may sound obvious, but it is easy to forget. Every day, Americans buy billions of dollars worth of things that are *not* produced, or at least not produced this year, and so are not counted in this year's GDP. For example, people may buy land, or they may buy financial assets such as stocks or bonds. While these things cost money, they are not counted in GDP because they are not "goods and services *produced*." Land (and the natural resources on it or under it) is not produced at all. Stocks and bonds represent a claim to ownership or to receive future payments, but they are not themselves goods or services.

In addition, people and businesses buy billions of dollars in *used* goods during the year, such as secondhand cars, previously occupied homes, used furniture, or an old photo of Elvis talking to an extraterrestrial. These goods were all produced, but not in the current period. We include only *currently produced* output when figuring this year's GDP.

... *for the marketplace* ...

GDP does not include *all* final goods and services produced in the economy. Rather, it includes only the ones produced for the marketplace—that is, with the intention of being *sold*. Because of this restriction, we exclude many important goods and services from our measure. For example, when you clean your own home, you have produced a final service—housecleaning—but it is *not* counted in GDP because you are doing it for yourself, not for the marketplace. If you *hire* a housecleaner to clean your home, however, this final service *is* included in GDP; it has become a market transaction.

The same is true for many services produced in the economy. Taking care of your children, washing your car, mowing your lawn, walking your dog—none of these services are included in GDP if you do them for yourself, but all *are* included if you pay someone else to do them for you.

... *during a given year* ...

This part of the definition of GDP tells us that GDP is an example of a **flow variable**—a measure of a *process* that takes place over a *period* of time:



You've learned that GDP excludes the value of many things that are bought and sold—such as land, financial assets, and used goods—because they are not *currently produced goods and services*. But all of this buying and selling *can* contribute to GDP indirectly. How? If a dealer or broker is involved in the transaction, then that dealer or broker is producing a current service: bringing buyer and seller together. The value of this service is part of current GDP.

For example, suppose you bought a secondhand book at your college bookstore for \$25. Suppose, too, that the store had bought the book from another student for \$15. Then the purchase of the used book will contribute \$10 to this year's GDP. Why? Because \$10 is the value of the bookstore's services; it's the premium you pay to buy the book in the store, rather than going through the trouble to find the original seller yourself. The remainder of your purchase—\$15—represents the value of the used book itself, and is *not* counted in GDP. The book was already counted when it was newly produced—in this or a previous year.

Flow variable A measure of a process that takes place over a period of time.

Gross domestic product is a flow variable: It measures a process—production—over a period of time.

The value of a flow depends on the length of the period over which we choose to measure it. For example, if you are asked, “What is your *income*?” (another flow variable), your answer will be different depending on whether the question refers to your hourly, weekly, monthly, or yearly income. The same is true of GDP: We can measure it per day, per month, or per year. (For example, in 1999, the United States produced \$25 billion worth of output on a typical day, \$786 billion in a typical month, and \$9,248 billion for the year as a whole. By tradition, the basic period for reporting GDP is a year.)

Stock Variable A measure of an amount that exists at a moment in time.

Not all macroeconomic variables are flow variables; some are **stock variables**—measures of things that *exist* at a *moment* in time. The U.S. population, the number of homes in the nation, the current value of your wealth—all these are stock variables because they are values measured at a particular instant. In this case, we never need to add the phrase *per week* or *per month*, since there is no *period* attached to the variable. (For example, it makes no sense to ask, “What is your wealth per month?” Instead, we would ask, “What is your wealth *right now*?”)

... *within the nation's borders.*

GDP measures output produced *within U.S. borders*—regardless of whether it was produced by Americans. This means we *include* output produced by foreign-owned resources and foreign citizens located in the United States, and we *exclude* output produced by Americans located in other countries. For example, when the rock star Sting, a resident of Britain, gives a concert tour in the United States, the value of his services is counted in U.S. GDP, but not in British GDP. Similarly, the services of an American nurse working in an Ethiopian hospital are part of Ethiopian GDP and not U.S. GDP.

THE EXPENDITURE APPROACH TO GDP

The Commerce Department’s Bureau of Economic Analysis (BEA)—the agency responsible for gathering, reporting, and analyzing movements in the nation’s output—calculates GDP in several different ways. The most important of these is the *expenditure approach*. Because this method of measuring GDP tells us so much about the structure of our economy, we’ll spend the next several pages on it.

In the expenditure approach, we divide output into four categories according to which group in the economy purchases it. The four categories are:

1. *Consumption goods and services (C)*, purchased by households
2. *Private investment goods and services (I)*, purchased by businesses
3. *Government goods and services (G)*, purchased by government agencies
4. *Net exports (NX)*, purchased by foreigners.¹

This is an exhaustive list: Every purchaser of U.S. output belongs to one of these four sectors. Thus, when we add up the purchases of the four sectors, we must get GDP:

Expenditure approach Measuring GDP by adding the value of goods and services purchased by each type of final user.

In the expenditure approach to measuring GDP, we add up the value of the goods and services purchased by each type of final user:

$$\text{GDP} = C + I + G + \text{NX}.$$

¹ The meaning and measurement of the term *net exports* will become clear in a few pages.

As you can see in Table 1, applying the expenditure approach to GDP in 1999 gives us $GDP = C + I + G + NX = \$6,255 + \$1,621 + \$1,629 + (-\$257) = \$9,248$ billion.

Now let's take a closer look at each of the four components of GDP.



GDP is measured and reported each *quarter*. But be careful: Quarterly GDP is almost always reported at an *annual rate*. For example, in the first quarter of 1999, we produced \$1,983 billion in final goods and services; but the GDP was reported at the *annual rate* of $4 \times \$1,983$ billion = \$7,933 billion. This is what we *would have* produced in 1999 if production had continued at the first quarter's rate for the entire year.

Consumption Spending. Consumption (C) is both the largest component of GDP—making up about three-quarters of total production in recent years—and the easiest to understand:

Consumption is the part of GDP purchased by households as final users.

Consumption (C) The part of GDP purchased by households as final users.

Almost everything that households buy during the year—restaurant meals, gasoline, new clothes, doctors' visits, movies, electricity, and more—is included as part of consumption spending when we calculate GDP.

But notice the word *almost*. Two categories of things that households buy during the year are *not* part of consumption because they are not part of GDP at all. The two categories, referred to in an earlier Dangerous Curves warning, are *used* goods (such as secondhand textbooks or cars) and assets such as stocks, bonds, or real estate.

There are also some quirky exceptions to the definition of consumption. For example, two things are included even though households do not actually buy them: (1) the total value of all food products that farm families produce and consume themselves (meat, dairy products, fruit, and vegetables) and (2) the total value of the shelter provided by homes that are owned by the families living in them. The government estimates (and adds to GDP) what farm families *would* pay if they had to buy all of their farm products in the marketplace like everyone else. It also estimates the rent that homeowners *would* pay for their homes if they were renting from someone

GDP IN 1999: THE EXPENDITURE APPROACH (BILLIONS OF DOLLARS)

TABLE 1

Consumption Purchases		Private-Investment Purchases		Government Purchases		Net Exports	
Services	\$3,656	Plant and Equipment	\$1,166	Government Consumption	\$1,333	Exports	\$ 996
Nondurable Goods	\$1,841	New-Home Construction	\$ 411	Government Investment	\$ 296	Imports	\$1,253
Durable Goods	\$ 758	Changes in Business Inventories	\$ 44				
Consumption =	\$6,255	Private Investment =	\$1,621	Government Purchases =	\$1,629	Net Exports =	-\$257

$$\begin{aligned}
 GDP &= C + I + G + NX \\
 &= \$6,255 + \$1,621 + \$1,629 + (-\$257) \\
 &= \$9,248
 \end{aligned}$$

Source: Economic Report of the President, 2000 (average of 1999 second and third quarter annual rates).

else. Another exception is that the construction of new homes—even when households buy them—is not counted as consumption, but rather as private investment.

Private Investment. What do oil-drilling rigs, cash registers, office telephones, and the house you grew up in all have in common? They are all examples of *capital goods*—goods that will provide useful services in future years. When we sum the value of all of the capital goods in the country, we get our **capital stock**. As the name suggests, this is a *stock* variable—a value that exists at a moment in time.

Understanding the concept of capital stock helps us understand and define the concept of investment. A rough definition of **private investment** is *capital formation*—the *increase* in the nation’s capital stock during the year. Investment, like the other components of GDP, is a *flow* variable—a process (capital formation) that takes place over a period of time.

More specifically,

Private investment has three components: (1) business purchases of plant and equipment; (2) new home construction; and (3) changes in business firms’ inventory stocks (stocks of unsold goods).

Each of these components requires some explanation.

Business Purchases of Plant and Equipment. This category might seem confusing at first glance. Why aren’t plant and equipment considered intermediate goods? After all, business firms buy these things in order to produce other things. Doesn’t the value of their final goods include the value of their plant and equipment as well?

Actually, no, and if you go back to the definition of intermediate goods, you will see why. Intermediate goods are *used up* in producing the current year’s GDP. But a firm’s plant and equipment are intended to last for many years; only a small part of them is used up to make the current year’s output. Thus, we regard newly produced plant and equipment as final goods, and the firms that buy them as the final users of those goods.

For example, suppose our paper mill—the firm that turns wood chips into raw paper—buys a new factory building that is expected to last for 50 years. Then only a small fraction of that factory building—one-fiftieth—is used up in any one year’s production of raw paper, and only a small part of the factory building’s value will be reflected in the value of the firm’s current output. But since the factory is produced during the year, we must include its value *somewhere* in our measure of total production. In calculating GDP, we therefore count the factory building as an investment good.

Plant and equipment purchases are always the largest component of private investment. In 1999, businesses purchased and installed \$1,166 billion worth of plant and equipment, which was about 70 percent of total private investment spending that year. (See Table 1.)

New Home Construction. As you can see in Table 1, new home construction made up a significant part of total private investment in 1999. But it may strike you as odd that this category is part of investment spending at all, since most new homes are purchased by households and could reasonably be considered consumption spending instead. Why do we treat new home construction as investment spending in GDP?

Largely because residential housing is an important part of the nation’s *capital stock*. Just as an oil-drilling rig will continue to provide oil-drilling services for many years, so, too, a home will continue to provide shelter services into the future. If we want our measure of private investment spending to roughly correspond to

Capital stock The total value of all goods that will provide useful services in future years.

Private investment (I) The sum of business plant and equipment purchases, new home construction, and inventory changes.

the increase in the nation's capital stock, we must include this important category of capital formation in investment spending.

Changes in Inventories. Inventories are goods that have been produced, but not yet sold. They include goods on store shelves, goods making their way through the production process in factories, and raw materials waiting to be used. We count the *change* in firms' inventories as part of investment in measuring GDP. Why? When goods are produced but not sold during the year, they end up in some firm's inventory stocks. If we did *not* count changes in inventories, we would be missing this important part of current production. Remember that GDP is designed to measure total *production*, not just the part of production that is sold during the year.

To understand this more clearly, suppose that in some year, the automobile industry produced \$100 billion worth of automobiles, and that \$80 billion worth was sold to consumers. Then the other \$20 billion remained unsold and was added to the auto company's inventories. If we counted consumption spending alone (\$80 billion), we would underestimate automobile production in GDP. To ensure a proper measure, we must include not only the \$80 billion in cars sold (consumption), but also the \$20 billion *change* in inventories (private investment). In the end, the contribution to GDP is \$80 billion (consumption) + \$20 billion (private investment) = \$100 billion, which is, indeed, the total value of automobile production during the year.

What if inventory stocks *decline* during the year, so that the change in inventories is negative? Our rule still holds: We include the change in inventories in our measure of GDP—but in this case, we must add a *negative* number. For example, if the automobile industry produced \$100 billion worth of cars, but consumers bought \$120 billion, then \$20 billion worth of cars must have come from inventory stocks—cars that were produced (and counted) in previous years, but that remained unsold until this year. In this case, the consumption spending of \$120 billion will *overestimate* automobile production during the year, and subtracting \$20 billion corrects for this overcount. In the end, GDP would rise by \$120 billion (consumption) – \$20 billion (private investment) = \$100 billion.

Inventory changes are included in investment spending, rather than some other component of GDP, because unsold goods are part of the nation's capital stock. They will provide services in the future, when they are finally sold and used. An increase in inventories represents capital formation: a decrease in inventories—negative investment—is a decrease in the nation's capital.

Inventory changes are generally the smallest component of private investment, but the most highly volatile in percentage terms. In 1999, for example, inventories increased by about \$44 billion; one year earlier, they increased by \$71.2 billion—almost twice as much, and in some years, inventories *decrease*. Part of the reason for this volatility is that, while some inventory investment is intended, much of it is *unintended*. During recessions, for example, businesses are often unable to sell all of the goods they have produced and had planned to sell. The unsold output will be added to inventory stocks—an unintended increase in inventories. During rapid expansions, the opposite may happen: Businesses find themselves selling more than they produced—an unintended decrease in inventories.

Private Investment and the Capital Stock: Some Important Provisos. A few pages ago, it was pointed out that private investment corresponds only *roughly* to the increase in the nation's capital stock. Why this cautious language? Because changes in the nation's capital stock are somewhat more complicated than we are able to capture with private investment alone.



Unsold goods, like those pictured in this warehouse, are considered inventories. The *change* in these inventories—positive or negative—is included as investment when calculating GDP.

First, an important part of the nation's capital stock is owned and operated not by businesses, but by government—federal, state, and local. Courthouses, police cars, fire stations, weather satellites, military aircraft, highways, and bridges are all examples of government capital. In any given year, some of the nation's capital formation consists of an increase in government capital, which is not included in our measure of private investment. Thus, private investment spending alone tends to *underestimate* the increase in the nation's capital stock. A better measure of capital formation would include both private and government investment:

Total investment during the year is the sum of private investment and government investment.

In 1999, for example, the BEA estimated that \$296 billion of government spending was devoted to capital formation, so that total investment in that year was

$$\begin{aligned}\text{Total Investment} &= \text{Private Investment} + \text{Government Investment} \\ &= \$1,621 \text{ billion} + \$296 \text{ billion} = \$1,917 \text{ billion.}\end{aligned}$$

Second, in any given year, some of the nation's existing capital stock will wear out, or *depreciate*. Total investment spending, because it ignores depreciation, *overestimates* the increase in the nation's capital stock. We can fix this, however, by subtracting depreciation from total investment, to obtain *net investment spending*. This is the amount by which private and government investment actually causes the capital stock to increase:

$$\text{Net Investment} = \text{Total Investment} - \text{Depreciation.}$$

For example, the government estimates that in 1999, \$1,141 billion of private and government capital depreciated, so that net investment for the year was

$$\text{Net Investment} = \$1,917 \text{ billion} - \$1,141 \text{ billion} = \$776 \text{ billion.}$$

Net investment comes close to being a true measure of the increase in the capital stock during the year. But in the minds of many economists, we are still not completely there, because we are still ignoring two kinds of capital formation. One is the purchase of *consumer durables*—goods such as furniture, automobiles, washing machines, and personal computers for home use. All of these goods can be considered capital goods, since they will continue to provide services for many years. In 1999, households purchased \$758 billion in durables. If we deduct from this an estimate of depreciation on the existing stock of durables (say, \$100 billion), we would get the increase in the stock of durables: \$758 billion – \$100 billion = \$658 billion. Some economists would argue that *if* we included this \$658 billion or so as part of investment, we would have an even better measure of the increase in the capital stock

Finally, our typical measures of capital formation ignore *human capital*—the skills and training of the labor force. Think about a surgeon's skills in performing a heart bypass operation, or a police detective's ability to find clues and solve a murder, or a Web-page designer's mastery of HTML and Java. These types of knowledge will continue to provide valuable services well into the future, just like plant and equipment or new housing. To measure the increase in the capital stock most broadly, then, we *should* include the additional skills and training acquired by the workforce during the year. But human capital growth—like growth in con-

Net investment Total investment minus depreciation.

sumer durables—is *not* included in the official measure of investment by the BEA.

Government Purchases. In 1999, the government bought \$1,629 billion worth of goods and services that were part of GDP—about a sixth of the total. This component of GDP is called **government purchases**, although in recent years the Department of Commerce has begun to use the phrase *government consumption and investment purchases*. Government *investment*, as discussed earlier, refers to capital goods purchased by government agencies. The rest of government purchases is considered government *consumption*—spending on goods and services that are used up during the year. This includes the salaries of government workers and military personnel, and raw materials such as computer paper for government offices, gasoline for government vehicles, and the electricity used in government buildings.

There are a few things to keep in mind about government purchases in GDP. First, we include purchases by state and local governments as well as the federal government. In macroeconomics, it makes little difference whether the purchases are made by a local government agency like the parks department of Kalamazoo, Michigan, or a huge federal agency such as the U.S. Department of Defense.

Second, government purchases include *goods*—like fighter jets, police cars, school buildings, and spy satellites—and *services*—such as those performed by police, legislators, and military personnel. The government is considered to be a purchaser even if it actually produces the goods or services itself. For example, if you are taking your economics course at a public college or university—like Western Illinois University or the City University of New York—then your professor is selling teaching services to a state or city government. His or her salary enters into GDP as part of government purchases.

Finally, it's important to distinguish between government *purchases*—which are counted in GDP—and government *spending* as measured by local, state, and federal budgets and reported in the media. What's the difference? In addition to their purchases of goods and services, government agencies also disburse money for **transfer payments**. These funds are *given* to people or organizations—*not* to buy goods or services from them, but rather to fulfill some social obligation or goal. For example, Social Security payments by the federal government, unemployment insurance and welfare payments by state governments, and money disbursed to homeless shelters and soup kitchens by city governments are all examples of transfer payments. The important thing to remember about transfer payments is this:

Transfer payments represent money redistributed from one group of citizens (taxpayers) to another (the poor, the unemployed, the elderly). While transfers are included in government budgets as spending, they are not purchases of currently produced goods and services, and so are not included in the government purchases or in GDP.



Be *extremely* careful when using the term *investment* in your economics course. In economics, investment refers to capital formation, such as the building of a new factory, home, or hospital, or the production and installation of new capital equipment, or the accumulation of inventories by business firms. In everyday language, however, *investment* has a very different meaning: a place to put your wealth. Thus, in ordinary English, you invest whenever you buy stocks or bonds or certificates of deposit or when you lend money to a friend who is starting up a business. But in the language of economics, you have not invested, but merely changed the form in which you are holding your wealth (say, from checking account balances to stocks or bonds). To avoid confusion, remember that investment takes place only when there is new production of capital goods—that is, only when there is *capital formation*.

Government purchases (G) Spending by federal, state, and local governments on goods and services.

Transfer payment Any payment that is not compensation for supplying goods or services.



The main source of information on U.S. GDP is the Bureau of Economic Analysis. Their Web page can be found at <http://www.bea.doc.gov/>.

Net exports (NX) Total exports minus total imports.

Net Exports. There is one more category of buyer for output produced in the United States: *the foreign sector*. In 1999, for example, purchasers *outside* the nation bought approximately \$996 billion of U.S. goods and services—about 11 percent of our GDP. These exports are part of U.S. production of goods and services and so are included in GDP.

However, once we recognize dealings with the rest of the world, we must correct an inaccuracy in our measure of GDP the way we've reported it so far. Americans buy many goods and services every year that were produced *outside* the United States (Chinese shoes, Japanese cars, Mexican beer, Costa Rican coffee). When we add up the final purchases of households, businesses, and government agencies, we *overcount* U.S. production because we include goods and services produced abroad, which are *not* part of U.S. output. To correct for this overcount, we deduct all *imports* into the United States during the year, leaving us with just output produced in the United States. In 1999, these imports amounted to \$1,253 billion—an amount equal to about 13.5 percent of our GDP.

Let's recap: To obtain an accurate measure of GDP, we must add the part of U.S. production that is purchased by foreigners—total exports. But to correct for including the goods produced abroad, we must subtract Americans' purchases of goods produced outside of the United States—total imports. In practice, we take both of these steps together by adding **net exports (NX)**, which are total exports minus total imports.

To properly account for output sold to, and bought from, foreigners, we must include net exports—the difference between exports and imports—as part of expenditure in GDP.

In 1999, when total exports were \$996 billion and total imports were \$1,253 billion, net exports—as you can see in Table 1—were $996 - 1,253 = -257$ billion. The negative number indicates that the imports we're subtracting from GDP are greater than the exports we're adding.

OTHER APPROACHES TO GDP

In addition to the expenditure approach, in which we calculate GDP as $C + I + G + NX$, there are other ways of measuring GDP. You may be wondering: Why bother? Why not just use one method—whichever is best—and stick to it? Is the Bureau of Economic Analysis just trying to make life difficult for introductory economics students?

Actually, there are two good reasons for measuring GDP in different ways. The first is practical. Each method of measuring GDP is subject to measurement errors. By calculating total output in several different ways, and then trying to resolve the differences, the BEA gets a more accurate measure than would be possible with one method alone. The second reason is that the different ways of measuring total output give us different insights into the structure of our economy. Let's take a look at two more ways of measuring—and thinking about—GDP.

The Value-Added Approach. In the expenditure approach, we record goods and services only when they are sold to their final users—at the end of the production process. But we can also measure GDP by adding up each *firm's* contribution to the product *as it is produced*.

TABLE 2

Firm	Cost of Intermediate Goods	Revenue	Value Added
Lumber Company	\$ 0	\$1.00	\$1.00
Paper Mill	\$1.00	\$1.50	\$0.50
Office Supplies Manufacturer	\$1.50	\$2.25	\$0.75
Wholesaler	\$2.25	\$3.50	\$1.25
Retailer	\$3.50	\$5.00	\$1.50
			Total: \$5.00

VALUE ADDED AT DIFFERENT STAGES OF PRODUCTION

A firm's contribution to a product is called its *value added*. More formally,

A firm's value added is the revenue it receives for its output, minus the cost of all the intermediate goods that it buys.

Look back at Figure 1, which traces the production of a ream of notebook paper. The paper mill, for example, buys \$1.00 worth of wood chips (an intermediate good) from the lumber company and turns it into raw paper, which it sells for \$1.50. The value added by the paper mill is $\$1.50 - \$1.00 = \$0.50$. Similarly, the office-supplies maker buys \$1.50 worth of paper (an intermediate good) from the paper mill and sells it for \$2.25, so its value added is $\$2.25 - \$1.50 = \$0.75$. If we total the value added by each firm, we should get the final value of the notebook paper, as in Table 2.² The total value added is $\$1.00 + \$0.50 + \$0.75 + \$1.25 + \$1.50 = \5.00 , which is equal to the final sales price of the ream of paper. For any good or service, it will always be the case that the sum of the values added by all firms equals the final sales price. This leads to our second method of measuring GDP:

In the value-added approach, GDP is the sum of the values added by all firms in the economy.

Value added The revenue a firm receives minus the cost of the intermediate goods it buys.

Value-added approach Measuring GDP by summing the value added by all firms in the economy.

The Factor Payments Approach. If a bakery sells \$200,000 worth of bread during the year and buys \$25,000 in intermediate goods (flour, eggs, yeast), then its value added—its revenue minus the cost of its intermediate goods—is $\$200,000 - \$25,000 = \$175,000$. This is also the sum that will be *left over* from its revenue after the bakery pays for its intermediate goods.

Where does this \$175,000 go? In addition to its intermediate goods, the bakery must pay for the *resources* it used during the year—the land, labor, and capital that enabled it to add value to its intermediate goods.

Payments to owners of resources are called **factor payments**, because resources are also called the factors of production. Owners of capital (the owners of the firm's buildings or machinery, or those who lend funds to the firm so that it can buy buildings and machinery) receive *interest payments*; owners of land and natural resources receive *rent*; and those who provide labor to the firm receive *wages and salaries*. Finally, there is one additional resource used by the firm: *entrepreneurship*. In every capitalist economy, the entrepreneurs are those who visualize society's needs,

Factor payments Payments to the owners of resources that are used in production.

² To keep our example simple, we assume that the lumber company simply cuts down trees and slices up lumber, using just land, labor, and capital. We thus assume it uses no intermediate goods.

mobilize and coordinate the other resources so that production can take place, and gamble that the enterprise will succeed. The people who provide this entrepreneurship (often the owners of the firms) receive a fourth type of factor payment—*profit*.

Now let's go back to our bakery, which received \$200,000 in revenue during the year. We've seen that \$25,000 of this went to pay for intermediate goods, leaving \$175,000 in value added earned by the factors of production. Let's suppose that \$110,000 went to pay the wages of the bakery's employees, \$10,000 was paid out as interest on loans, and \$15,000 was paid in rent for the land under the bakery. That leaves $\$175,000 - \$110,000 - \$10,000 - \$15,000 = \$40,000$. This last sum—since it doesn't go to anyone else—stays with the owner of the bakery. It, too, is a factor payment—profit—for the entrepreneurship she provides. Thus, when all of the factor payments—including profit—are added together, the total will be $\$110,000 + \$10,000 + \$15,000 + \$40,000 = \$175,000$ —precisely equal to the value added at the bakery. More generally,

In any year, the value added by a firm is equal to the total factor payments made by that firm.

Earlier, we learned that GDP equals the sum of all firms' value added; now we've learned that each firm's value added is equal to its factor payments. Thus, GDP must equal the total factor payments made by all firms in the economy. This gives us our *third* method of measuring GDP:

In the factor payments approach, GDP can be measured by summing all of the factor payments made by all firms in the economy. Equivalently, it can be measured by adding up all of the income—wages and salaries, rent, interest, and profit—earned by all households in the economy.³

The factor payments approach to GDP gives us one of our most important insights into the macroeconomy:

GDP—the total output of the economy—is equal to the total income earned in the economy.

This simple idea—output equals income—follows directly from the factor payments approach to GDP. It explains why macroeconomists use the terms “output” and “income” interchangeably: They are one and the same. If output rises, income rises by the same amount; if output falls, income falls by an equal amount.

MEASURING GDP: A SUMMARY

You've now learned three different ways to calculate GDP:

Expenditure Approach: $GDP = C + I + G + NX$

Value-Added Approach: $GDP = \text{Sum of value added by all firms}$

Factor Payments Approach: $GDP = \text{Sum of factor payments made by all firms}$
 $= \text{Wages and salaries} + \text{interest} + \text{rent} + \text{profit}$
 $= \text{Total household income}$

³ Actually, this is just an approximation. Before a firm pays its factors of production, it first deducts a small amount for depreciation of its plant and equipment, and another small amount for the sales taxes it must pay to the government. Thus, GDP and total factor payments are slightly different. We ignore this difference in the text.

Factor payments approach Measuring GDP by summing the factor payments made by all firms in the economy.

We will use these three approaches to GDP again and again as we study what makes the economy tick. Make sure you understand why each one of them should, in theory, give us the same number for GDP.

REAL VERSUS NOMINAL GDP

Since GDP is measured in dollars, we have a serious problem when we want to track the change in output over time. The problem is that the value of the dollar—its purchasing power—is itself changing. As prices have risen over the past 100 years, the value of the dollar has steadily fallen. Trying to keep track of GDP using dollars in different years is like trying to keep track of a child's height using a ruler whose length changes each year. If we find that the child is three rulers tall in one year and four rulers tall in the next, we cannot know whether the child is really growing taller—or, if so, by how much—until we adjust for the effects of a changing ruler. The same is true for GDP and for any other economic variable measured in dollars: We usually need to adjust our measurements to reflect changes in the value of the dollar.

*When a variable is measured over time with no adjustment for the dollar's changing value, it is called a **nominal variable**. When a variable is adjusted for the dollar's changing value, it is called a **real variable**.*

Nominal variable A variable measured without adjustment for the dollar's changing value.

Most government statistics are reported in both nominal and real terms, but economists focus almost exclusively on real variables. This is because changes in nominal variables don't really tell us much. For example, from 1990 to 1991, nominal GDP increased from \$5,743.8 billion to \$5,916.7 billion. But production actually decreased over that period—the increase in nominal GDP was due entirely to a rise in prices.

Real variable A variable adjusted for changes in the dollar's value.

The distinction between nominal and real values is crucial in macroeconomics. The public, the media, and sometimes even government officials have been confused by a failure to make this distinction. Whenever we want to track significant changes in key macroeconomic variables—such as the average wage rate, wealth, income, or GDP or any of its components—we always use real variables.

Since our economic well-being depends, in part, on the goods and services we can buy, it is important to translate nominal values—which are measured in current dollars—to real values—which are measured in purchasing power.

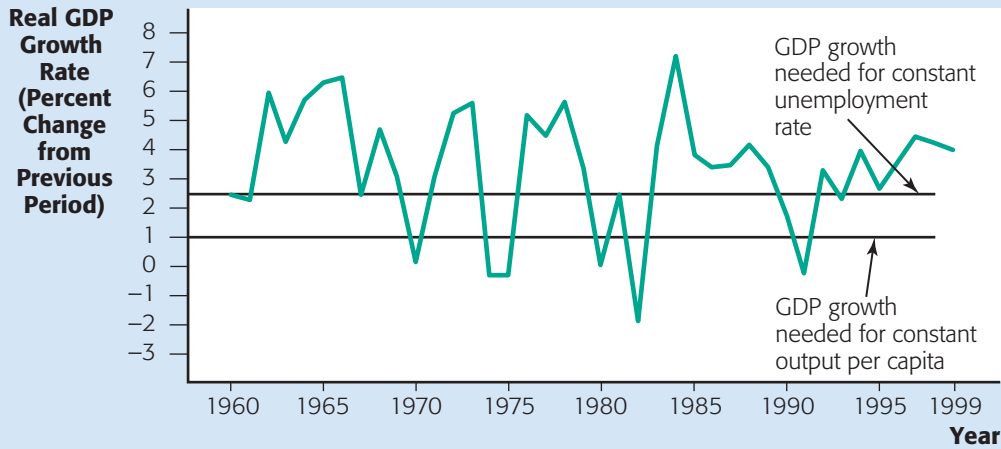
In the next chapter, you'll learn how economists translate nominal variables into real variables.

HOW GDP IS USED

We've come a long way since 1931. In that year—as the United States plummeted into the worst depression in its history—Congress summoned economists from government agencies, from academia, and from the private sector to testify about the state of the economy. They were asked the most basic questions: How much output was the nation producing, and how much had production fallen since 1929? How much income were Americans earning, how much were they spending on goods and services, and how much were they saving? How much profit were businesses earning, and what were they doing with their profits? Had the economy continued to deteriorate in the previous year, or had it finally hit bottom? To the surprise of the members of Congress, no one could answer any of these questions,

FIGURE 2

REAL GDP GROWTH RATE, 1960–1999



Although the growth rate of real GDP has fluctuated over time, it has almost always exceeded the 1 percent rate needed to maintain output per capita. On the average, it has exceeded the 2.5 percent rate needed to prevent a rise in unemployment.

because *no one was keeping track of our national income and output!* The last measurement—which was rather incomplete—had been made in 1929.

Thus began the U.S. system of national income accounts—a system whose value was instantly recognized around the world and was rapidly copied by other countries. Today, the government's reports on GDP are used to steer the economy over both the short run and the long run. In the short run, sudden changes in real GDP can alert us to the onset of a recession or a too-rapid expansion that can overheat the economy. Many (but not all) economists believe that, if alerted in time, the government can design policies to help keep the economy on a more balanced course.

GDP is also used to measure the long-run growth rate of the economy's output. Indeed, we typically define the average *standard of living* as *output per capita*—real GDP divided by the population. In order for output per capita to rise, real GDP must grow faster than the population. Since the U.S. population tends to grow by about 1 percent per year, a real GDP growth rate of 1 percent per year is needed just to *maintain* our output per capita; higher growth rates are needed to increase it.

Look at Figure 2, which shows the annual percentage change in real GDP from 1960 to 1999. The lower horizontal line indicates the 1 percent growth needed to just maintain output per capita. You can see that, on average, real GDP has grown by more than this, so that output per capita has steadily increased over time.

Long-run growth in GDP is also important for another reason: to ensure that the economy generates enough additional jobs for a growing population. In order to prevent the unemployment rate from rising, real GDP must increase even faster than the population. In practice, a growth rate of about 2.5 percent per year—the upper horizontal line in the figure—seems to generate the required number of new jobs each year. You can see that real GDP growth has, on average, been sufficient for this purpose as well.

To sum up: We use GDP to guide the economy in two ways. In the short run, to alert us to recessions and give us a chance to stabilize the economy. In the long run, to tell us whether our economy is growing fast enough to raise output per capita

and our standard of living, and fast enough to generate sufficient jobs for a growing population. You can see that GDP is an extremely useful measure. But it is not without its problems.

PROBLEMS WITH GDP

Our GDP statistics are plagued by some important inaccuracies. One problem is *quality changes*. Suppose a new ballpoint pen comes out that lasts four times as long as previous versions. What *should* happen to GDP? Ideally, each new pen should count the same as four old pens, since one new pen offers the same *writing services* as four old ones. But the analysts at the Bureau of Economic Analysis (BEA) would most likely treat this new pen the same as an old pen and record an increase in GDP only if the total number of pens increased. Why? Because the BEA has a limited budget. While it does include the impact of quality changes for many goods and services (such as automobiles and computers), the BEA simply does not have the resources to estimate quality changes for millions of different goods and services. These include many consumer goods (such as razor blades that shave closer and last longer), medical services (increased surgery success rates and shorter recovery periods), and retail services (faster checkout times due to optical scanners). By ignoring these quality improvements, GDP probably understates true growth from year to year.

A second problem arises from the *underground economy*, which contains hidden economic activity, either because it is illegal (drugs, prostitution, most gambling) or because those engaged in it are avoiding taxes. These activities cannot be measured accurately, so the BEA must estimate them. Many economists believe that the BEA's estimates are too low. As a result, GDP may understate total output. However, since the *relative* importance of the underground economy does not change rapidly, the BEA's estimates of *changes* in GDP from year to year should not be seriously affected.

Finally, except for food grown and consumed by farmers and for housing services, GDP does not include **nonmarket production**—goods and services that are produced, but not sold in the marketplace. All of the housecleaning, typing, sewing, lawn mowing, and child rearing that people do themselves, rather than hiring someone else, are excluded from GDP. Whenever a nonmarket transaction (say, cleaning your apartment) becomes a market transaction (hiring a housecleaner to do it for you), GDP will rise, even though total production (cleaning one apartment) has remained the same. This can exaggerate the growth in GDP over long periods of time. Over the last half-century, much production has, indeed, shifted away from the home and to the market. Parenting, which was not counted in past years' GDP, has become day care, which *does* count, currently contributing several billion dollars annually to GDP. Similarly, home-cooked food has been replaced by takeout, talking to a friend has been replaced by therapy, and the neighbor who watches your house while you're away has been replaced by a store-bought alarm system or an increase in police protection. In all of these cases, real GDP increases, even though production has not.

What do these problems tell us about the value of GDP? That for certain purposes—especially interpreting *long-run* changes in GDP—we must exercise extreme caution. For example, suppose that, over the next 20 years, the growth rate of GDP slows down. Would this mean that something is going wrong with the economy? Would it suggest a need to change course? Not necessarily. It *could* be that the underground economy or unrecorded quality changes are becoming more important. Similarly, if GDP growth accelerates, it could mean that our living standards are rising more rapidly. But it might instead mean that economic activity is shifting out of the home and into the market even more rapidly than in the past.

Nonmarket production Goods and services that are produced, but not sold in a market.

When it comes to *short-term* changes in the economy, however, we can have much more confidence in using GDP. Look back at our discussion of problems with GDP in this section. The distortions we've discussed tend to remain roughly constant over the short run. If GDP suddenly drops, it is extremely unlikely that the underground economy has suddenly become more important, or that there has been a sudden shift from market to nonmarket activities, or that we are suddenly missing more quality changes than usual. Rather, we can be reasonably certain that output and economic activity are slowing down.

Short-term changes in real GDP are fairly accurate reflections of the state of the economy. A significant short-term drop in real GDP virtually always indicates a decrease in production, rather than a measurement problem.

This is why policy makers, businesspeople, and the media pay such close attention to GDP as a guide to the economy from quarter to quarter.

EMPLOYMENT AND UNEMPLOYMENT

When you think of unemployment, you may have an image in your mind that goes something like this: As the economy slides into recession, an anxious employee is called into an office and handed a pink slip by a grim-faced manager. “Sorry,” the manager says, “I wish there were some other way. . . .” Perhaps, in your mind, the worker spends the next few months checking the classified ads, pounding the pavement, and sending out resumes in a desperate search for work. And perhaps, after months of trying, the laid-off worker gives up, spending days at the neighborhood bar, drinking away the shame and frustration, and sinking lower and lower into despair and inertia.

For some people, joblessness begins and ends very much like this—a human tragedy, and a needless one. On one side, we have people who want to work and support themselves by producing something; on the other side is the rest of society, which could certainly use more goods and services. Yet somehow, the system isn't working, and the jobless cannot find work. The result is often hardship for the unemployed and their families, and a loss to society in general.

But this is just one face of unemployment, and there are others. Some instances of unemployment, for example, have little to do with macroeconomic conditions. And frequently, unemployment causes a lot less suffering than in our grim story.



Employment-related information for the United States can be found at the Bureau of Labor Statistics' Web site: <http://stats.bls.gov>.

TYPES OF UNEMPLOYMENT

Economists have found it useful to classify unemployment into four different categories, each arising from a different cause and each having different consequences.

Frictional unemployment Joblessness experienced by people who are between jobs or who are just entering or re-entering the labor market.

Frictional Unemployment. Frictional unemployment is short-term joblessness experienced by people who are between jobs or who are entering the labor market for the first time or after an absence. For example, imagine that you have a job, but that you think you'd be happier at some other firm. Since you can't search for a new job while working full time, you may decide to quit your job and begin looking elsewhere. In an ideal frictionless world, every potential employer would immediately know that you were available, and you would immediately know which job you'd prefer most, so you would become re-employed the instant you quit; you would not be unemployed between jobs. Of course, in the real world, it takes time to find a

job—time to prepare your resume, to decide where to send it, to wait for responses, and then to investigate job offers so you can make a wise choice. It also takes time for employers to consider your skills and qualifications and to decide whether you are right for their firms. During all that time, you will be unemployed: willing and able to work, but not working.

There are other examples of this type of unemployment. A parent reenters the labor force after several years spent raising the children. A 22-year-old searches for a job after graduating from college. In both of these cases, it may take some time to find a job, and during that time, the job seeker is *frictionally* unemployed.

Because frictional unemployment is, by definition, short term, it causes little hardship to those affected by it. In most cases, people have enough savings to support themselves through a short spell of joblessness, or else they can borrow on their credit card or from friends or family to tide them over. Moreover, this kind of unemployment has important benefits: By spending time searching rather than jumping at the first opening that comes their way, people find jobs for which they are better suited and in which they will ultimately be more productive. As a result, workers earn higher incomes, firms have more productive employees, and society has more goods and services.

Seasonal Unemployment. Seasonal unemployment is joblessness related to changes in weather, tourist patterns, or other seasonal factors. For example, most ski instructors lose their jobs every April or May, and many construction workers are laid off each winter.

Seasonal unemployment, like frictional unemployment, is rather benign: It is short term, and, because it is entirely predictable, workers are often compensated in advance for the unemployment they experience in the off-season. Construction workers, for example, are paid higher-than-average hourly wages, in part to compensate them for their high probability of joblessness in the winter.

Seasonal unemployment complicates the interpretation of unemployment data. Seasonal factors push the unemployment rate up in certain months of the year and pull it down in others, even when overall conditions in the economy remain unchanged. For example, each June, unemployment rises as millions of high school and college students—who do not want to work during the school year—begin looking for summer jobs. If the government reported the actual rise in unemployment in June, it would *seem* as if labor market conditions were deteriorating, when in fact, the rise is just a predictable and temporary seasonal change. To prevent any misunderstandings, the government usually reports the *seasonally adjusted* rate of unemployment, a rate that reflects only those changes beyond normal for the month. For example, if the unemployment rate in June is typically one percentage point higher than during the rest of the year, then the seasonally adjusted rate for June will be the actual rate minus one percentage point.

Structural Unemployment. Sometimes, there are jobs available and workers who would be delighted to have them, but job seekers and employers are *mismatched* in some way. For example, in the early 2000s, there have been plenty of job openings in high-tech industries, such as computer hardware and software design, satellite technology, and communications. Many of the unemployed, however, do not have the skills and training to work in these industries—there is a mismatch between the skills they have and those that are needed. The mismatch can also be geographic, as when construction jobs go begging in Northern California, Oregon, and Washington, but unemployed construction workers live in other states.

Unemployment that results from these kinds of mismatches is called **structural unemployment**, because it arises from *structural change* in the economy: when old,

Seasonal unemployment Joblessness related to changes in weather, tourist patterns, or other seasonal factors.

Structural unemployment Joblessness arising from mismatches between workers' skills and employers' requirements or between workers' locations and employers' locations.

ding industries are replaced with new ones that require different skills and are located in different areas of the country. Structural unemployment is generally a stubborn, *long-term* problem, often lasting several years or more. Why? Because it can take considerable time for the structurally unemployed to find jobs—time to relocate to another part of the country or time to acquire new skills. To make matters worse, the structurally unemployed could benefit from financial assistance for job training or relocation, but—because they don't have jobs—they are unable to get loans.

Structural unemployment is a much bigger problem in other countries, especially in Europe, than it is in the United States. In November 1999, when the U.S. unemployment rate was 4.1 percent, the rate in Germany was 9.1 percent, in France 10.5 percent, and in Spain 15.4 percent. All three of those European countries have large groups of lower-skilled workers who are not qualified for the jobs that are available. Even Canada suffers from much more structural unemployment than does the United States—its unemployment rate at the end of 1999 was 6.9 percent, much of it concentrated in the maritime provinces. And within the United States, some areas have higher structural unemployment than others. For example, in early 2000 when the national unemployment rate was 4 percent, the rates in New York City and Los Angeles were closer to 6 percent.

The types of unemployment we've considered so far—frictional, structural, and seasonal—arise from *microeconomic* causes; that is, they are attributable to changes in specific industries and specific labor markets, rather than to conditions in the overall economy. This kind of unemployment cannot be eliminated, as people will always spend some time searching for new jobs, there will always be seasonal industries in the economy, and structural changes will, from time to time, require workers to move to new locations or gain new job skills. Some amount of microeconomic unemployment is a sign of a dynamic economy. It allows workers to sort themselves into the best possible jobs, enables us to enjoy seasonal goods and services like winter skiing and summers at the beach, and permits the economy to go through structural changes when needed.

Nevertheless, many economists feel that the levels of microeconomic unemployment in the United States are too high and that we can continue to enjoy the benefits of a fast-changing and flexible economy with a lower unemployment rate. To achieve this goal, they advocate government programs to help match the unemployed with employers and to help the jobless relocate and learn new skills. Note, however, that these are *microeconomic* policies—government intervention in particular labor markets or to help particular kinds of workers. Since frictional, seasonal, and structural unemployment have microeconomic causes, they need *microeconomic* cures.

Our fourth and last type of unemployment, however, has an entirely *macroeconomic* cause.

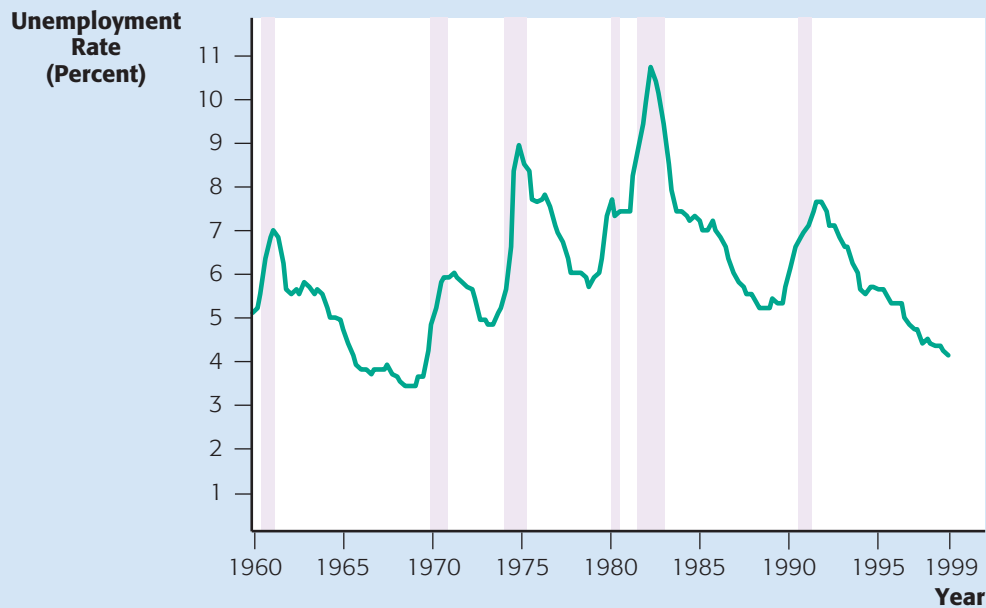
Cyclical Unemployment. When the economy goes into a recession and total output falls, the unemployment rate rises. Many previously employed workers lose their jobs and have difficulty finding new ones. At the same time, there are fewer openings, so new entrants to the labor force must spend more than the usual “frictional” time searching before they are hired. This type of unemployment—because it is caused by the business cycle—is called **cyclical unemployment**.

Look at Figure 3, which shows the unemployment rate in the United States for each quarter since 1960, and notice the rises that occurred during periods of recession (shaded). For example, in the recessions of the early 1980s, the unemployment rate rose from about 6 percent to almost 10 percent; in the more recent recession of 1990–1991, it rose from 5.3 percent to more than 7 percent. These were rises in cyclical unemployment.

Cyclical unemployment Joblessness arising from changes in production over the business cycle.

U.S. QUARTERLY UNEMPLOYMENT RATE, 1960–1999

FIGURE 3



The unemployment rate rises during recessions (shaded) and falls during expansions.

Since it arises from conditions in the overall economy, cyclical unemployment is a problem for *macroeconomic* policy. This is why macroeconomists focus almost exclusively on cyclical unemployment, rather than the other types of joblessness. Reflecting this emphasis, macroeconomists say we have reached **full employment** when we come out of a recession and *cyclical unemployment is reduced to zero*, even though substantial amounts of frictional, seasonal, and structural unemployment may remain:

In macroeconomics, full employment is achieved when cyclical unemployment has been reduced to zero. But the overall unemployment rate at full employment is greater than zero because there are still positive levels of frictional, seasonal, and structural unemployment.

Full employment A situation in which there is no cyclical unemployment.

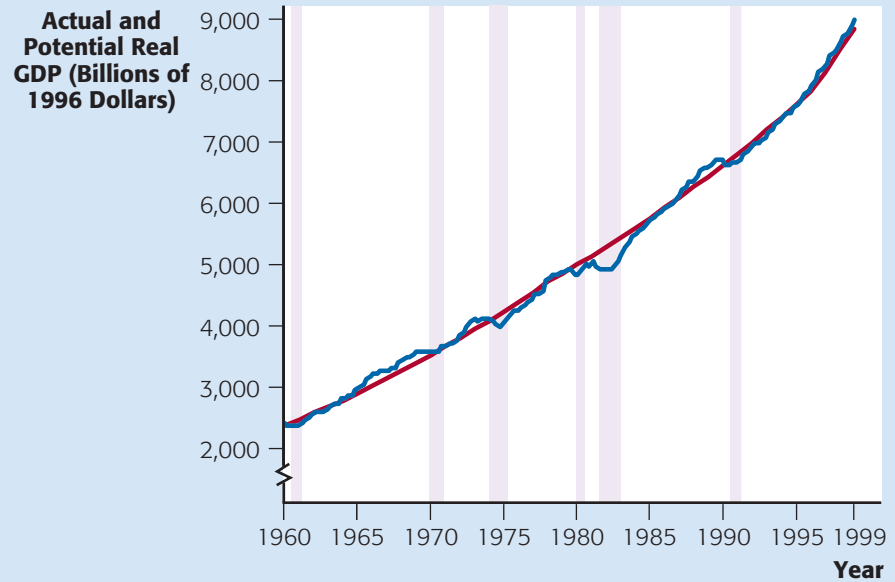
How do we tell how much of our unemployment is cyclical? Many economists believe that today, normal amounts of frictional, seasonal, and structural unemployment account for an unemployment rate of between 4 and 4.5 percent in the United States. Therefore, any unemployment beyond this is considered cyclical unemployment. For example, if the actual unemployment rate were 6 percent, we would identify 1.5 to 2.0 percent of the labor force as cyclically unemployed.

THE COSTS OF UNEMPLOYMENT

Why are we so concerned about achieving a low rate of unemployment? What are the *costs* of unemployment to our society? We can identify two different types of costs: economic costs—those that can be readily measured in dollar terms—and noneconomic costs—those that are difficult or impossible to measure in dollars, but still affect us in important ways.

FIGURE 4

ACTUAL AND POTENTIAL REAL GDP, 1960–1999



Economic Costs. The chief economic cost of unemployment is the *opportunity cost* of lost output—the goods and services the jobless *would* produce if they were working, but do not produce because they cannot find work. This cost must be borne by our society, although the burden may fall more on one group than another. If, for example, the unemployed were simply left to fend for themselves, then *they* would bear most of the cost. If they turned to crime in order to survive, then crime victims would share the burden. In fact, the unemployed are often given government assistance, so that the costs are spread among citizens in general. But there is no escaping this central fact:

When there is cyclical unemployment, the nation produces less output, and so some group or groups within society must consume less output.

One way of viewing the economic cost of cyclical unemployment is illustrated in Figure 4. The blue line shows real GDP over time, while the red line shows the path of our **potential output**—the output we *could* have produced if the economy were operating at full employment.

Notice that actual output is sometimes *above* potential output. At these times, unemployment is *below* the full-employment rate. For example, during the expansion in the late 1960s, cyclical unemployment was eliminated, and the sum of frictional, seasonal, and structural unemployment dropped below 4.5 percent, its normal level for those years. At other times, real GDP is *below* potential output, most often during and immediately following a recession. At these times, unemployment rises above the full-employment rate. In the 1982–83 recession, the unemployment rate remained above 9.5 percent for more than a year

In the figure, you can see that we have spent more of the last 35 years operating *below* our potential than above it. That is, the cyclical ups and downs of the econ-

Potential output The level of output the economy could produce if operating at full employment.

omy have, on balance, led to lower living standards than we would have had if the economy had always operated just at potential output.

Broader Costs. There are also costs of unemployment that go beyond lost output. Unemployment—especially when it lasts for many months or years—can have serious psychological and physical effects. Some studies have found that increases in unemployment cause noticeable rises in the number of heart attack deaths, suicides, and admissions to state prisons and psychiatric hospitals. The jobless are more likely to suffer a variety of health problems, including high blood pressure, heart disorders, troubled sleep, and back pain. There may be other problems—such as domestic violence, depression, and alcoholism—that are more difficult to document. And, tragically, most of those who lose their jobs also lose their health insurance, increasing the likelihood that these problems will have serious consequences.

Unemployment also causes setbacks in achieving important social goals. For example, most of us want a fair and just society where all people have an equal chance to better themselves. But our citizens do not bear the burden of unemployment equally. In a recession, we do not all suffer a reduction in our work hours; instead, some people are laid off entirely, while others continue to work roughly the same hours.

Moreover, the burden of unemployment is not shared equally among different groups in the population, but tends to fall most heavily on minorities, especially minority youth. As a rough rule of thumb, the unemployment rate for blacks is twice that for whites; and the rate for *teenage* blacks is triple the rate for blacks overall. Table 3 shows that the unemployment rates for January 2000 are consistent with this general experience. Notice the extremely high unemployment rate for black teenagers: 23.8 percent. Two years earlier—when the overall unemployment rate was 4.7 percent—the rate for black teenagers was even higher: 36.0 percent. This contributes to a vicious cycle of poverty and discrimination: When minority youths are deprived of that all-important first job, they remain at a disadvantage in the labor market for years to come.

HOW UNEMPLOYMENT IS MEASURED

In January 2000, about 140 million Americans did not have jobs. Were all of these people unemployed? Absolutely not. The unemployed are those *willing and able* to work, but who do not have jobs. Most of the 140 million nonworking Americans were either *unable* or *unwilling* to work. For example, the very old, the very young, and the very ill were unable to work, as were those serving prison terms. Others were able to work, but preferred not to, including millions of college students, homemakers, and retired people.

TABLE 3

UNEMPLOYMENT RATES FOR VARIOUS GROUPS, JANUARY 2000

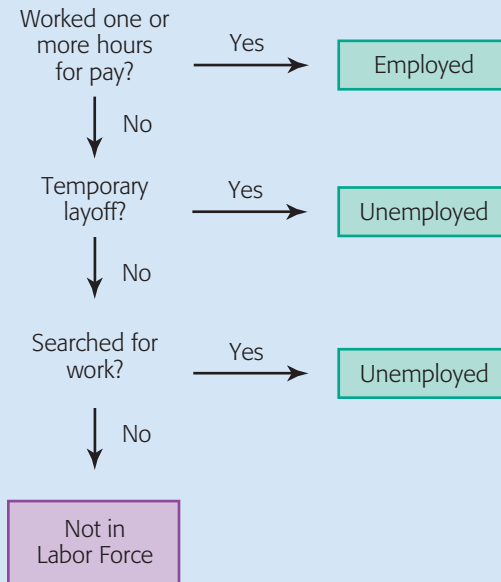
Group	Unemployment Rate
Whites	3.4%
Hispanics	5.6%
Blacks	8.2%
White Teenagers	9.1%
Black Teenagers	23.8%

Source: *The Employment Situation: January 2000*: Bureau of Labor Statistics News Release, February 4, 2000.

FIGURE 5

BLS interviewers ask a series of questions to determine whether an individual is employed, unemployed, or not in the labor force.

HOW BLS MEASURES EMPLOYMENT STATUS



But how, in practice, can we determine who is willing and able? This is a thorny problem, and there is no perfect solution to it. In the United States, we determine whether a person is willing and able to work by his or her *behavior*. More specifically, to be counted as unemployed, you must have recently *searched* for work. But how can we tell who has, and who has not, recently searched for work?

The Census Bureau's Household Survey. Every month, thousands of interviewers from the United States Census Bureau—acting on behalf of the U.S. Bureau of Labor Statistics (BLS)—conduct a survey of 60,000 households across America. This sample of households is carefully selected to give information about the entire population. Household members who are under 16, in the military, or currently residing in an institution like a prison or hospital are excluded. The interviewer will then ask questions to determine what the remaining household members did during the *previous week*.

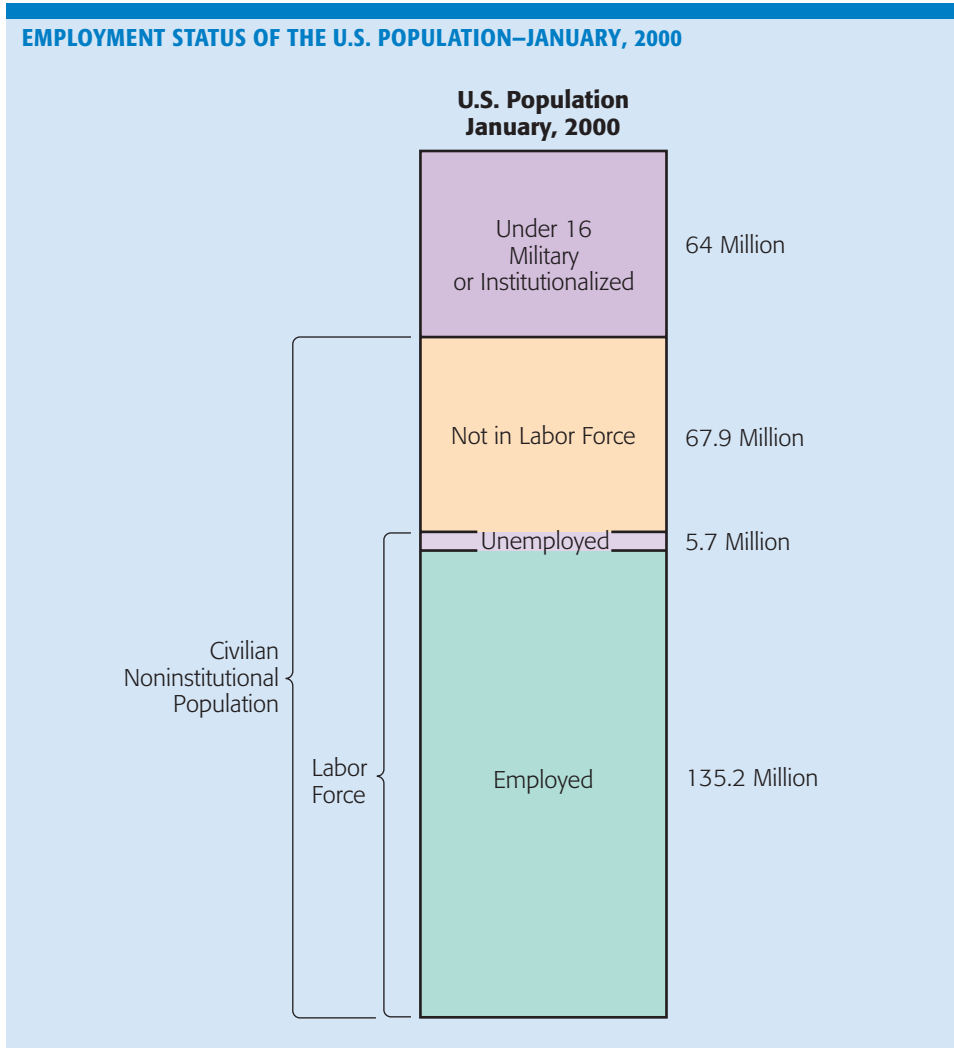
Figure 5 shows roughly how this works. First, the interviewer asks whether the household member has worked one or more hours for pay or profit. If the answer is yes, the person is considered employed; if no, another question is asked: Has she been *temporarily* laid off from a job from which she is waiting to be recalled? A yes means the person is unemployed; a no leads to one more question: Did the person actively *search* for work during the previous four weeks. If yes, the person is unemployed; if no, she is not in the labor force.

Figure 6 illustrates how the BLS, extrapolating from its 60,000-household sample, classified the U.S. population in January 2000. First, note that about 64 million people were ruled out from consideration because they were under 16 years of age, living in institutions, or in the military. The remaining 208.8 million people made up the civilian, noninstitutional population, and of these, 135.2 million were employed, and 5.7 million were unemployed. Adding the employed and unemployed together gives us the **labor force**, equal to 135.2 million + 5.7 million = 140.9 million.

Labor force Those people who have a job or who are looking for one.

EMPLOYMENT STATUS OF THE U.S. POPULATION—JANUARY, 2000

FIGURE 6



Finally, we come to the official **unemployment rate**, which is defined as the percentage of the labor force that is unemployed:

Unemployment rate The fraction of the labor force that is without a job.

$$\text{Unemployment rate} = \frac{\text{Unemployed}}{\text{Labor Force}} = \frac{\text{Unemployed}}{(\text{Unemployed} + \text{Employed})}$$

Using the numbers in Figure 6, the unemployment rate in January 2000 was $5.7/(5.7 + 135.2) = 0.040$ or 4.0 percent. This was the number released to journalists at 8:00 A.M. on the first Friday of February 2000, and the number that made headlines in your local newspaper the next day.

PROBLEMS IN MEASURING UNEMPLOYMENT

The Bureau of the Census earns very high marks from economists for both its sample size—60,000 households—and the characteristics of its sample, which very closely match the characteristics of the U.S. population. Still, the official unemployment rate suffers from some important measurement problems.

Many economists believe that our official measure seriously underestimates the extent of unemployment in our society. There are two reasons for this belief: the treatment of *involuntary part-time workers* and the treatment of *discouraged workers*.

As you can see in Figure 5, anyone working one hour or more for pay during the survey week is treated as employed. This includes many people who would like a full-time job—and may even be searching for one—but who did some part-time work during the week. Some economists have suggested that these people—called **involuntary part-time workers**—should be regarded as partially employed and partially unemployed.

Involuntary part-time workers

Individuals who would like a full-time job, but who are working only part time.

How many involuntary part-time workers are there? In January 2000, the BLS estimated that there were about 3.5 million.⁴ If each of these workers were considered half-employed and half-unemployed, the unemployment rate in that month would have been 5.3 percent, instead of the officially reported 4.0 percent.

Discouraged workers Individuals who would like a job, but have given up searching for one.

Another problem is the treatment of **discouraged workers**—individuals who would like to work but, because they feel little hope of finding a job, have given up searching. Because they are not taking active steps to find work, they are considered “not in the labor force” (see Figure 5). Some observers feel that discouraged workers should be counted as unemployed. After all, these people are telling us that they are willing and able to work, but they are not working. It seems wrong to exclude them just because they are not actively seeking work. Others argue that counting discouraged workers as unemployed would reduce the objectivity of our unemployment measure. Talk is cheap, they believe, and people may *say* anything when asked whether they would like a job; the real test is what people *do*. Yet even the staunchest defenders of the current method of measuring employment would agree that *some* discouraged workers are, in fact, willing and able to work and should be considered unemployed. The problem, in their view, is determining which ones.

How many discouraged workers are there? No one knows for sure. The BLS tries to count them periodically, but defining who is genuinely discouraged is a thorny problem. Using the BLS’s rather strict criteria, there were 339,000 discouraged workers in January 2000. But with a looser, unofficial definition of “discouraged worker”—people who are not working but say they want a job—the count rises to 4.8 million. Including some or all of these people among the unemployed would raise the unemployment rate significantly.

There are also reasons to believe that the unemployment rate overstates the amount of joblessness as we usually think of it. Remember that a person is counted as unemployed if he or she did not work in the past week, but took some active steps to look for work in the past month. Some of those counted as unemployed did work earlier in the month, even though they were not at work in the survey week. Others whose principal activities are outside the labor market—going to school, keeping house, or being retired—are counted as unemployed because they checked the help wanted ads in the past month or talked to friends about what jobs might be available.

Still, the unemployment rate—as currently measured—tells us something important: the number of people who are *searching* for jobs, but have not yet found them. It is not exactly the same as the percentage of the labor force that is jobless even though willing and able to work. But if we could obtain a perfect measure of the latter, the unemployment rate—as currently measured—would be highly correlated with it.

Moreover, the unemployment rate tells us something unique about conditions in the macroeconomy. When the unemployment rate is relatively low—so that few peo-

⁴ This and other information about unemployment in January 2000 comes from *The Employment Situation: January 2000*, Bureau of Labor Statistics News Release, February 4, 2000.

ple are actively seeking work—a firm that wants to hire more workers may be forced to lure them from other firms, by offering a higher wage rate. This puts upward pressure on wages and can lead to future inflation. A high unemployment rate, by contrast, tells us that firms can more easily expand by hiring those who are actively seeking work, without having to lure new workers from another firm and without having to offer higher wages. This suggests little inflationary danger. Later in the book, we will discuss the connection between unemployment and inflation more fully.

SOCIETY'S CHOICE OF GDP

The title of this section might seem absurd: How can we say that society *chooses* its level of GDP? Wouldn't the citizens of any nation want their GDP to be as large as possible—and certainly larger than it currently is? The answer is yes. After all, GDP is certainly important to our economic well-being. Few of us would want to live at the levels of output per capita that prevailed 100, 50, or even 25 years ago. Increased output of medical care, restaurant meals, entertainment, transportation services, and education have all contributed to a higher standard of living and an overall improvement in our economic well-being.

But there is more to economic well-being than *just* GDP. Suppose that, over the next 10 years, real GDP per capita were to double. Further, suppose that our measure is entirely accurate (not plagued by the measurement problems discussed in the previous section). Would the average person be better off in 10 years? Maybe. But maybe not. We cannot say, because our GDP statistic ignores so many *other* things that are important to our economic well-being besides the quantity of goods and services at our disposal, and these things may be changing at the same time that GDP is changing.

What are these other things that affect our economic well-being? They include the leisure time we have to spend with family and friends; the cleanliness of our environment; the safety of our workplaces, homes, and streets; the fairness of our society; and more. None of these are included in GDP, which is, after all, just a measure of our output of goods and services.

But what does this have to do with society's choice of GDP? Remember that economics is the study of choice under conditions of scarcity, and just as individuals are constrained by a scarcity of time or income or wealth, society as a whole is constrained by the resources at its disposal. In many cases, we must choose between using our resources to have more of the output that is included in GDP or more of *other* things we care about that are *not* part of GDP.

For example, look at Figure 7, which shows the familiar production possibility frontier, or PPF, from Chapter 2, but with a new twist. In Chapter 2, we looked at the trade-off between two categories of *goods*—medical care versus everything else. Here, we explore the trade-off between real GDP on the horizontal axis and some other thing that we care about—something *not* in GDP—on the vertical axis. In this example, we've put *leisure time* on the vertical axis.

Why is there a trade-off between real GDP and leisure? Because with a given state of technology for producing output, a given population, and given quantities of other resources, the more labor time we devote to production, the more goods and services we will have. But more labor time means less leisure time: Either more people must become employed, or the employed must work longer hours. In either case, the total amount of leisure time enjoyed by the population will decrease.

Let's first identify the two extremes of the PPF in the figure. The maximum leisure time achievable would occur at point A—zero output. Here, people would

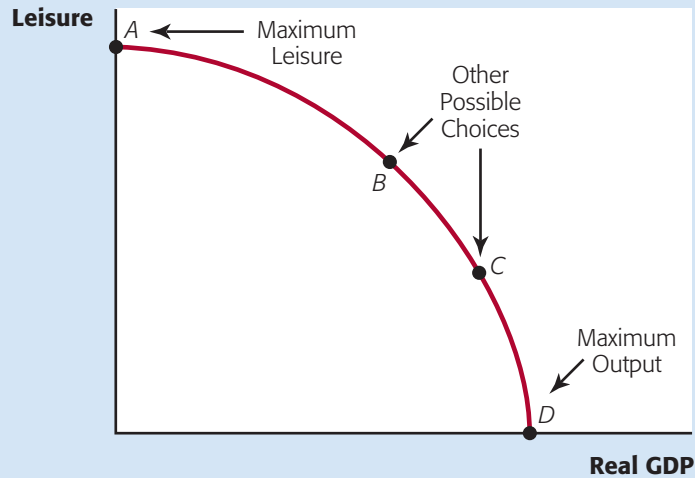
Using the THEORY



FIGURE 7

The production possibilities frontier shows that, for a given population and state of technology, a society must choose between the level of real GDP and the time available for leisure. At point *A*, people devote all their time to leisure, so GDP is zero. Point *D*, by contrast, represents the maximum GDP attainable if everyone works the maximum hours, year-round. *B* and *C* represent intermediate possibilities.

THE TRADE-OFF BETWEEN REAL GDP AND LEISURE



have to survive by eating fruit and nuts that fell to the ground, since even climbing trees or hunting animals would involve work. On the other hand, the maximum GDP achievable would require the lowest possible level of leisure—every able-bodied person working 16 hours per day, 365 days per year. This is indicated by point *D* in the figure. The curve that connects points *A* and *D* is the PPF that shows the maximum combinations of output and leisure achievable. (Why does the curve bow out from the origin? Review the material on PPFs in Chapter 2 if you need to.)

The PPF in Figure 7 makes it clear that society faces a trade-off and that, in any year, we choose our level of GDP subject to the constraints of this trade-off. We could draw a similar PPF illustrating the trade-off between a high GDP on the one hand and a clean environment or workplace safety on the other.

How does society choose its location on this kind of PPF?

In a market economy, the choice is made partly by individual households and firms. Suppose that most workers' tastes began to shift toward having more leisure and that they were willing to sacrifice income in order to have it. For example, workers might prefer a 20 percent cut in work hours, with a 20 percent cut in total compensation. Suppose, too, that there were no loss of efficiency from having people work shorter hours. Then any firm that refused to match these new worker preferences—cutting pay and work hours by 20 percent—would have to pay above-average wages in order to attract workers. With higher labor costs, the firm would have to charge a higher price for its output. Such a firm would not be able to compete with other firms that were offering the more desired, shorter workweek. As more and more firms moved toward a shorter workweek, society as a whole would move from a point like *C* in Figure 7 to a point like *B*—more leisure and a lower GDP.

Thus, at least to some extent, we can expect market pressures to adjust work hours to worker preferences for leisure on the one hand and income on the other. The result of these individual decisions will determine, in large measure, where we will be on the PPF in Figure 7.

Interestingly, the United States is farther down and to the right on this PPF (a point like *C*) than most European countries (which are at points like *B*). For ex-

ample, the average workweek in manufacturing is more than 40 hours in the United States, but around 30 hours in Germany. In addition, the typical U.S. worker takes two weeks of vacation each year, while the typical German worker takes five weeks. To a great extent, these differences in labor hours reflect differences in worker tastes. For example, when Germany introduced Thursday night shopping in 1989, retail workers—who didn't want to work the additional two hours even for additional pay—went on strike. As a result of his greater taste for leisure, the typical German—and the typical French person, Italian, and Spaniard—enjoys more leisure each year than the typical American does. But Europeans pay a cost: a lower GDP, and therefore fewer goods and services per person than they would otherwise have.

Our location on the PPF is also determined by society as a whole, as a matter of public policy. We vote for our representatives, who make rules and regulations under which our firms must operate. If, for example, the majority prefers a higher GDP and less leisure, it can vote for representatives who promise to change work rules. In Germany, for example, it is *illegal* for workers to take another job during their five weeks of annual vacation.⁵

There are also other dimensions to our choice of GDP. For example, with economic growth, a nation can enjoy a greater GDP in the future *and* more of other things—say, workplace safety or leisure. But economic growth comes at a cost as well. We'll examine that cost—and society's choices concerning the rate of economic growth—a few chapters from now.

⁵ Daniel Benjamin and Tony Horwitz, "German View: You Americans Work Too Hard—And For What?" *The Wall Street Journal*, July 14, 1994, p. B1

S U M M A R Y

This chapter discusses how some key macroeconomic aggregates are measured and reported. One important economic aggregate is *gross domestic product*—the total value of all final goods and services produced for the marketplace during a given year, within a nation's borders. GDP is a measure of an economy's total production. It is a flow variable that measures sales, to final users, of newly produced output.

In the *expenditure approach*, GDP is calculated as the sum of spending by households, businesses, governments, and foreigners on domestically produced goods and services. The *value-added approach* computes GDP by adding up each firm's contributions to total product as it is being produced. Value added at each stage of production is the revenue a firm receives minus the cost of the intermediate inputs it uses. Finally, the *factor payments approach* sums the payments to all resource owners. The three approaches reflect three different ways of viewing GDP.

Since nominal GDP is measured in current dollars, it changes when either production or prices change. *Real GDP* is nominal GDP adjusted for price changes; it rises only when production rises.

Real GDP is useful in the short run for giving warnings about impending recessions, and in the long run for indicat-

ing how fast the economy is growing. Unfortunately, it is plagued by important inaccuracies. It does not fully reflect quality changes or production in the underground economy, and it does not include many types of nonmarket production.

When real GDP grows, employment tends to rise and unemployment tends to fall. In the United States, a person is considered unemployed if he or she does not have a job but is actively seeking one. Economists have found it useful to classify unemployment into four different categories. *Fictional unemployment* is short-term unemployment experienced by people between jobs or by those who are just entering the job market. *Seasonal unemployment* is related to changes in the weather, tourist patterns, or other predictable seasonal changes. *Structural unemployment* results from mismatches—in skills or location—between jobs and workers. Finally, *cyclical unemployment* occurs because of the business cycle. Unemployment, particularly the structural and cyclical forms, involves costs. From a social perspective, unemployment means lost production. From the individual viewpoint, unemployment often involves financial, psychological, and physical harm.

KEY TERMS

gross domestic product (GDP)	capital stock	factor payments	cyclical unemployment
intermediate goods	private investment	factor payments approach	full employment
final good	net investment	nominal variable	potential output
flow variable	government purchases	real variable	labor force
stock variable	transfer payment	nonmarket production	unemployment rate
expenditure approach	net exports	frictional unemployment	involuntary part-time workers
consumption	value added	seasonal unemployment	discouraged workers
	value-added approach	structural unemployment	

REVIEW QUESTIONS

- What is the difference between final goods and intermediate goods? Why is it that only the value of final goods and services is counted in GDP?
- Which of the following are stock variables, and which are flow variables?
 - Microsoft's revenues
 - Microsoft's market value (the total value of shares held by its stockholders)
 - A household's spending
 - The value of a household's stock portfolio
- Using the expenditure approach, which of the following would be directly counted as part of U.S. GDP in 2001? In each case, state whether the action causes an increase in *C*, *I*, *G*, or *NX*.
 - A new personal computer produced by IBM, which remained unsold at the year's end
 - A physician's services to a household
 - Produce bought by a restaurant to serve to customers
 - The purchase of 1,000 shares of Disney stock
 - The sale of 50 acres of commercial property
 - A real estate agent's commission from the sale of property
 - A transaction in which you clean your roommate's apartment in exchange for his working on your car
 - An Apple I-Mac computer produced in the United States, and purchased by a French citizen
 - The government's Social Security payments to retired people
- How is the word *investment* used differently in economics than in ordinary language? Explain each of the three categories of investment.
- Describe the different kinds of factor payments.
- What is the difference between nominal and real variables? What is the main problem with using nominal variables to track the economy?
- Discuss the value and reliability of GDP statistics in both short-run and long-run analyses of the economy.
- Real GDP was measured at around \$8.8 trillion in 1999. Was the actual value of goods and services produced in the United States in 1999 likely to have been higher or lower than that? Why?
- What, if anything, could the government do to reduce frictional and structural unemployment?
- Categorize each of the following according to the type of unemployment it reflects. Justify your answers.
 - Workers are laid off when a GM factory closes due to a recession.
 - Workers selling software in a store are laid off when the store goes bankrupt due to competition from on-line software dealers.
 - Migrant farm workers' jobs end when the harvest is finished.
 - Lost jobs result from the movement of textile plants from Massachusetts to the South and overseas.
- Can unemployment ever be good for the economy? Explain.
- What are some of the different types of costs associated with unemployment?
- Discuss some of the problems with the way the Bureau of Labor Statistics computes the unemployment rate. In what ways do official criteria lead to an overestimate or underestimate of the actual unemployment figure?

P R O B L E M S A N D E X E R C I S E S

1. Calculate the total change in a year's GDP for each of the following scenarios:
 - a. A family sells a home, without using a broker, for \$150,000. They could have rented it on the open market for \$700 per month. They buy a 10-year-old condominium for \$200,000; the broker's fee on the transaction is 6 percent of the selling price. The condo's owner was formerly renting the unit at \$500 per month.
 - b. General Electric uses \$10 million worth of steel, glass, and plastic to produce its dishwashers. Wages and salaries in the dishwasher division are \$40 million; the division's only other expense is \$15 million in interest that it pays on its bonds. The division's revenue for the year is \$75 million.
 - c. On March 31, you decide to stop throwing away \$50 a month on convenience store nachos. You buy \$200 worth of equipment, cornmeal, and cheese, and make your own nachos for the rest of the year.
 - d. You win \$25,000 in your state's lottery. Ever the entrepreneur, you decide to open a Ping Pong ball washing service, buying \$15,000 worth of equipment

- e. from SpiffyBall Ltd. of Hong Kong and \$10,000 from Ball-B-Kleen of Toledo, Ohio.
 - e. Tone-Deaf Artists, Inc. produces 100,000 new White Snake CDs that it prices at \$15 apiece. Ten thousand CDs are sold abroad, but, alas, the rest remain unsold on warehouse shelves.
2. The country of Freedonia uses the same method to calculate the unemployment rate as the U.S. Bureau of Labor Statistics uses. From the data below, compute Freedonia's unemployment rate.

Population	10,000,000
Under 16	3,000,000
Over 16	
In military service	500,000
In hospitals	200,000
In prison	100,000
Worked one hour or more in previous week	4,000,000
Searched for work during previous four weeks	1,000,000

C H A L L E N G E Q U E S T I O N

Suppose, in a given year, someone buys a General Motors automobile for \$30,000. That same year, GM produced the car in Michigan, using \$10,000 in parts imported from Japan. However, the parts imported from Japan themselves contained \$3,000 in components produced in the United States.

- a. By how much does U.S. GDP rise?
- b. Using the expenditure approach, what is the change in each component (C, I, G, and NX) of U.S. GDP?
- c. What is the change in Japan's GDP and each of its components?

E X P E R I E N T I A L E X E R C I S E S

1. One criticism of the U.S. national income accounts is that they ignore the effects of environmental pollution. The World Bank's group on environmental economics has been investigating ways of assessing environmental degradation. Take a look at their work on "green accounting" at <http://wbi018.worldbank.org/environment/EEI.nsf/all/Green+Accounting?OpenDocument>. What kinds of problems have they identified, and what proposals have they made to deal with those problems?



2. Data on the Consumer Price Index are released near the middle of each month. (You can find the exact date by consulting the online calendar at <http://www.leggmason.com/CAL/calendar.html>) Data on GDP are released on the last Friday of each month (in preliminary, revised, and then final form). Analysis of these data appears in the first section of the following weekday's *Wall Street Journal*. Look in the "Economy" section to find the story. What do the latest available data tell you about the current rate of inflation and the current rate of GDP growth? Is the economy expanding or contracting?





CHAPTER

19

THE MONETARY SYSTEM, PRICES, AND INFLATION

CHAPTER OUTLINE

The Monetary System

History of the Dollar
Why Paper Currency Is Accepted
as a Means of Payment

Measuring the Price Level and Inflation

Index Numbers
The Consumer Price Index
How the CPI Has Behaved
From Price Index to Inflation Rate
How the CPI Is Used
Real Variables and Adjustment for
Inflation
Inflation and the Measurement of
Real GDP

The Costs of Inflation

The Inflation Myth
The Redistributive Cost of Inflation
The Resource Cost of Inflation

Using the Theory: Is the CPI Accurate?

Sources of Bias in the CPI
The Consequences of Overstating
Inflation
The Future of the CPI

Appendix: Calculating the Consumer Price Index

Unit of value A common unit
for measuring how much
something is worth.

Means of payment Anything
acceptable as payment for
goods and services.

You pull into a gas station deep in the interior of the distant nation of Chaotica. The numbers on the gas pump don't make sense to you, and you can't figure out how much to pay. Luckily, the national language of Chaotica is English, so you can ask the cashier how much the gas costs. He replies, "Here in Chaotica, we don't have any standard system for measuring quantities of gas, and we don't have any standard way to quote prices. My pump here measures in my own unit, called the Slurp, and I will sell you 6 Slurps for that watch you are wearing, or a dozen Slurps for your camera." You spend the next half hour trying to determine how many Slurps there are in a gallon and what form of payment you can use besides your watch and camera.

Life in the imaginary nation of Chaotica would be difficult. People would spend a lot of time figuring out how to trade with each other, time that could otherwise be spent producing things or enjoying leisure activities. Fortunately, in the real world, virtually every nation has a *monetary system* that helps to organize and simplify our economic transactions.

THE MONETARY SYSTEM

A monetary system establishes two different types of standardization in the economy. First, it establishes a **unit of value**—a common unit for measuring how much something is worth. A standard unit of value permits us to compare the costs of different goods and services and to communicate these costs when we trade. The dollar is the unit of value in the United States. If a college textbook costs \$75, while a one-way airline ticket from Phoenix to Minneapolis costs \$300, we know immediately that the ticket has the same value in the marketplace as four college textbooks.

The second type of standardization concerns the **means of payment**—the things we can use as payment when we buy goods and services. In the United States, the means of payment include dollar bills, personal checks, money orders, credit cards like Visa and American Express, and, in some experimental locations, prepaid cash cards with magnetic strips.

These two functions of a monetary system—establishing a unit of value and a standard means of payment—are closely related, but they are not the same thing.

The unit-of-value function refers to the way we *think* about and record transactions; the means-of-payment function refers to how payment is actually made.

The unit of value works in the same way as units of weight, volume, distance, and time. In fact, the same sentence in Article I of the U.S. Constitution gives Congress the power to create a unit of value along with units of weights and measures. All of these units help us determine clearly and precisely what is being traded for what. Think about buying gas in the United States—you exchange dollars for gallons. The transaction will go smoothly and quickly only if there is clarity about both the unit of fluid volume (gallons) *and* the unit of purchasing power (dollars).

The means of payment can be different from the unit of value. For example, in some countries where local currency prices change very rapidly, it is common to use the U.S. dollar as the unit of value—to specify prices in dollars—while the local currency remains the means of payment. Even in the United States, when you use a check to buy something, the unit of value is the dollar, but the means of payment is a piece of paper with your signature on it.

In the United States, the dollar is the centerpiece of our monetary system. It is the unit of value in virtually every economic transaction, and dollar bills are very often the means of payment as well. How did the dollar come to play such an important role in the economy?

HISTORY OF THE DOLLAR

Prior to 1790, each colony had its own currency. It was named the “pound” in every colony, but it had a different purchasing power in each of them. In 1790, soon after the Constitution went into effect, Congress created a new unit of value called the dollar. Historical documents show that merchants and businesses switched immediately to the new dollar, thereby ending the chaos of the colonial monetary systems. Prices began to be quoted in dollars, and accounts were kept in dollars. The dollar rapidly became the standard unit of value.

But the primary means of payment in the United States until the Civil War was paper currency issued by private banks. Just as the government defined the length of the yard, but did not sell yardsticks, the government defined the unit of value, but let private organizations provide the means of payment.

During the Civil War, however, the government issued the first federal paper currency, the greenback. It functioned as both the unit of value and the major means of payment until 1879. Then the government got out of the business of money creation for a few decades. During that time, currency was once again issued by private banks. Then, in 1913, a new institution called the **Federal Reserve System** was created to be the national monetary authority in the United States. The Federal Reserve was charged with creating and regulating the nation’s supply of money, and it continues to do so today.

Federal Reserve System The central bank and national monetary authority of the United States.

WHY PAPER CURRENCY IS ACCEPTED AS A MEANS OF PAYMENT

You may be wondering why people are willing to accept paper dollars as a means of payment. Why should a farmer give up a chicken, or a manufacturer give up a new car, just to receive a bunch of green rectangles with words printed on them? In fact, paper currency is a relatively recent development in the history of the means of payment.

The earliest means of payment were precious metals and other valuable commodities such as furs or jewels. These were called *commodity money* because they had important uses other than as a means of payment. The nonmoney use is what



Today, dollars are not backed by gold or silver, but we accept them as payment because we know that others will accept them from us.

Fiat money Anything that serves as a means of payment by government declaration.

gave commodity money its ultimate value. For example, people would accept furs as payment because furs could be used to keep warm. Similarly, gold and silver had a variety of uses in industry, as religious artifacts, and for ornamentation.

Precious metals were an especially popular form of commodity money. Eventually, to make it easier to identify the value of precious metals, they were minted into coins whose weight was declared on their faces. Because gold and silver coins could be melted down into pure metal and used in other ways, they were still commodity money.

Commodity money eventually gave way to paper currency. Initially, paper currency was just a certificate representing a certain amount of gold or silver held by a bank. At any time, the holder of a certificate could go to the bank that issued it and trade the certificate for the stated amount of gold or silver. People were willing to accept paper money as a means of payment for two reasons. First, the currency could be exchanged for a valuable commodity like gold or silver. Second, the issuer—either a government or a bank—could only print new money when it acquired additional gold or silver. This put strict limits on money printing, so people had faith that their paper money would retain its value in the marketplace.

But today, paper currency is no longer backed by gold or any other physical commodity. If you have a dollar handy, put this book down and take a close look at the bill. You will not find on it any promise that you can trade your dollar for gold, silver, furs, or anything else. Yet we all accept it as a means of payment. Why? A clue is provided by the statement in the upper left-hand corner of every bill: *This note is legal tender for all debts, public and private.* The statement affirms that the piece of paper in your hands will be accepted as a means of payment (you can “tender” it to settle any “debt, public or private”) by any American because the government says so. This type of currency is called **fiat money**. *Fiat*, in Latin, means “let there be,” and fiat money serves as a means of payment by government declaration.

The government need not worry about enforcing this declaration. The real force behind the dollar—and the reason that we are all willing to accept these green pieces of paper as payment—is its long-standing acceptability by *others*. As long as you have confidence that you can use your dollars to buy goods and services, you won’t mind giving up goods and services for dollars. And because everyone else feels the same way, the circle of acceptability is completed.

But while the government can declare that paper currency is to be accepted as a means of payment, it cannot declare the terms. Whether 10 gallons of gas will cost you 1 dollar, 10 dollars, or 20 dollars is up to the marketplace. The value of the dollar—its purchasing power—does change from year to year, as reflected in the changing prices of the things we buy. In the rest of this chapter, we will discuss some of the problems created by the dollar’s changing value and the difficulty economists have measuring and monitoring the changes. We postpone until later chapters the question of *why* the value of the dollar changes from year to year.

MEASURING THE PRICE LEVEL AND INFLATION

One hundred years ago, you could buy a pound of coffee for 15 cents, see a Broadway play for 40 cents, buy a new suit for \$6, and attend a private college for \$200 in yearly tuition.¹ Needless to say, the price of each of these items has gone up considerably since then. Microeconomic causes—changes in individual markets—can

¹ Scott Derks, ed., *The Value of the Dollar: Prices and Incomes in the United States: 1860–1989* (Detroit, MI: Gale Research Inc., 1994), various pages.

explain only a tiny fraction of these price changes. For the most part, these price rises came about because of an ongoing rise in the **price level**—the average level of dollar prices in the economy. In this section, we begin to explore how the price level is measured, and how this measurement is used.

Price level The average level of dollar prices in the economy.

INDEX NUMBERS

Most measures of the price level are reported in the form of an **index**—a series of numbers, each one representing a different period. Index numbers are meaningful only in a *relative* sense: We compare one period's index number with that of another period and can quickly see which one is larger and by how much. The actual number for a particular period has no meaning in and of itself.

Index A series of numbers used to track a variable's rise or fall over time.

In general, an index number for any measure is calculated as

$$\frac{\text{Value of measure in current period}}{\text{Value of measure in base period}} \times 100.$$

Let's see how index numbers work with a simple example. Suppose we want to measure how violence on TV has changed over time, and we have data on the number of violent acts shown in each of several years. We could then construct a TV-violence index. Our first step would be to choose a *base period*—a period to be used as a benchmark. Let's choose 1996 as our base period, and suppose that there were 10,433 violent acts on television in that year. Then our violence index in any current year would be calculated as

$$\frac{\text{Number of violent acts in current year}}{10,433} \times 100.$$

In 1996—the base year—the index will have the value $(10,433/10,433) \times 100 = 100$. Look again at the general formula for index numbers, and you will see that this is always true: *An index will always equal 100 in the base period.*

Now let's calculate the value of our index in another year. If there were 14,534 violent acts in 2000, then the index that year would have the value

$$\frac{14,534}{10,433} \times 100 = 139.3.$$

Index numbers compress and simplify information so that we can see how things are changing at a glance. Our media violence index, for example, tells us at a glance that the number of violent acts in 2000 was 139.3 percent of the number in 1996. Or, more simply, TV violence grew by 39.3 percent between 1996 and 2000.

THE CONSUMER PRICE INDEX

The most widely used measure of the price level in the United States is the **Consumer Price Index (CPI)**. This index—which is designed to track the prices paid by the typical consumer—is compiled and reported by the Bureau of Labor Statistics (BLS).

Consumer Price Index An index of the cost, through time, of a fixed market basket of goods purchased by a typical household in some base period.

Measuring the prices paid by the typical consumer is not easy. Two problems must be solved before we even begin. The first problem is to decide which goods and services we should include in our average. The CPI tracks only *consumer* prices; it excludes goods and services that are not directly purchased by consumers. More specifically, the CPI excludes goods purchased by businesses (such as capital

equipment, raw materials, or wholesale goods), goods and services purchased by government agencies (such as fighter-bombers and the services of police officers) and goods and services purchased by foreigners (U.S. exports). The CPI *does* include newly produced consumer goods and services that are part of consumption spending in our GDP—things such as new clothes, new furniture, new cars, haircuts, and restaurant meals. It also includes some things that are *not* part of our GDP but that are part of the typical family's budget. For example, the CPI includes prices for *used* goods such as used cars or used books, and imports from other countries—for example, French cheese, Japanese cars, and Mexican tomatoes.

The second problem is how to combine all the different prices into an average price level. In any given month, different prices will change by different amounts. The average price of doctor's visits might rise by 1 percent, the price of blue jeans might rise by a tenth of a percent, the price of milk might fall by half a percent, and so on. When prices change at different rates, and when some are rising while others are falling, how can we track the change in the *average* price level? We would not want to use a simple average of all prices—adding them up and dividing by the number of goods. A proper measure would recognize that we spend very little of our incomes on some goods—such as Tabasco sauce—and much more on others—like car repairs or rent.

The CPI's approach is to track the cost of the *CPI market basket*—the collection of goods and services that the typical consumer bought in some base period. If the market basket's cost rises by 10 percent over some period, then the price level, as reported by the CPI, will rise by 10 percent. This way, goods and services that are relatively unimportant in the typical consumer's budget will have little weight in the CPI. Tabasco sauce could triple in price and have no noticeable impact on the cost of the complete market basket. Goods that are more important—such as auto repairs or rent—will have more weight.

In recent years, the base year² for the CPI has been 1983, so, following our general formula for price indexes, the CPI is calculated as

$$\frac{\text{Cost of market basket in current year}}{\text{Cost of market basket in 1983}} \times 100.$$

The appendix to this chapter discusses the calculation of the CPI in more detail.



<http://>

You can find the latest information on the CPI at <http://stats.bls.gov/newsrels.htm>—the Bureau of Labor Statistics Web site.

HOW THE CPI HAS BEHAVED

Table 1 shows the actual value of the CPI for December of selected years. Because it is reported in index number form, we can easily see how much the price level has changed over different time intervals. In December 1999, for example, the CPI had a value of 168.3, telling us that the typical market basket in that year cost 68.3 percent more than it would have cost in the July 1983 base period. In December 1960, the CPI was 29.8, so the cost of the market basket in that year was only 29.8 percent of its cost in July 1983. In July 1983 (not shown), the CPI's value was 100.

² To be more specific: The market basket currently used by the Bureau of Labor Statistics reflects purchasing patterns over the period 1993–95. However, the *base period* used in calculations is July 1983. Thus, a more detailed version of our formula is: CPI in current year = Cost of 1993–95 market basket in current year / Cost of 1993–95 market basket in July 1983 × 100. The denominator requires some careful interpretation: It is what the 1993–95 market basket *would* have cost at July 1983 prices. As you can verify, with this formula, the CPI in July 1983 will be equal to 100. In official BLS statistics, the July 1983 base period is still referred to as the “1982–1984 base period” because a survey of consumer spending patterns had been conducted from 1982 to 1984.

TABLE 1

**CONSUMER PRICE INDEX,
DECEMBER, SELECTED
YEARS, 1960–1999**

Year	Consumer Price Index
1960	29.8
1965	31.8
1970	39.8
1975	55.5
1980	86.3
1985	109.3
1990	133.8
1995	153.5
1999	168.3

FROM PRICE INDEX TO INFLATION RATE

The Consumer Price Index is a measure of the price *level* in the economy. The **inflation rate** measures how fast the price level is changing, as a percentage rate. When the price level is rising, as it almost always is, the inflation rate is positive. When the price level is falling, as it did during the Great Depression, we have a negative inflation rate, which is called **deflation**.

Figure 1 shows the U.S. rate of inflation—as measured by the CPI—since 1950. For each year, the inflation rate is calculated as the percentage change in the CPI from December of the previous year to December of that year. For example, the CPI in December 1998 was 163.9, and in December 1999 it was 168.3. The inflation rate for 1999 was $(168.3 - 163.9)/163.9 = 0.027$ or 2.7 percent. Notice that inflation was low in the 1950s and 1960s, was high in the 1970s and early 1980s, and has been low since then. In later chapters, you will learn what causes the inflation rate to rise and fall.

Inflation rate The percent change in the price level from one period to the next.

Deflation A decrease in the price level from one period to the next.

HOW THE CPI IS USED

The CPI is one of the most important measures of the performance of the economy. It is used in three major ways:

As a Policy Target. In the introductory macroeconomics chapter, we saw that price stability—or a low inflation rate—is one of the nation’s important macroeconomic goals. The measure most often used to gauge our success in achieving low inflation is the CPI.

To Index Payments. A payment is **indexed** when it is set by a formula so that it rises and falls proportionately with a price index. An indexed payment makes up for the loss in purchasing power that occurs when the price level rises. It raises the nominal payment by just enough to keep its purchasing power unchanged. In the United States, millions of government retirees and Social Security recipients have their benefit payments

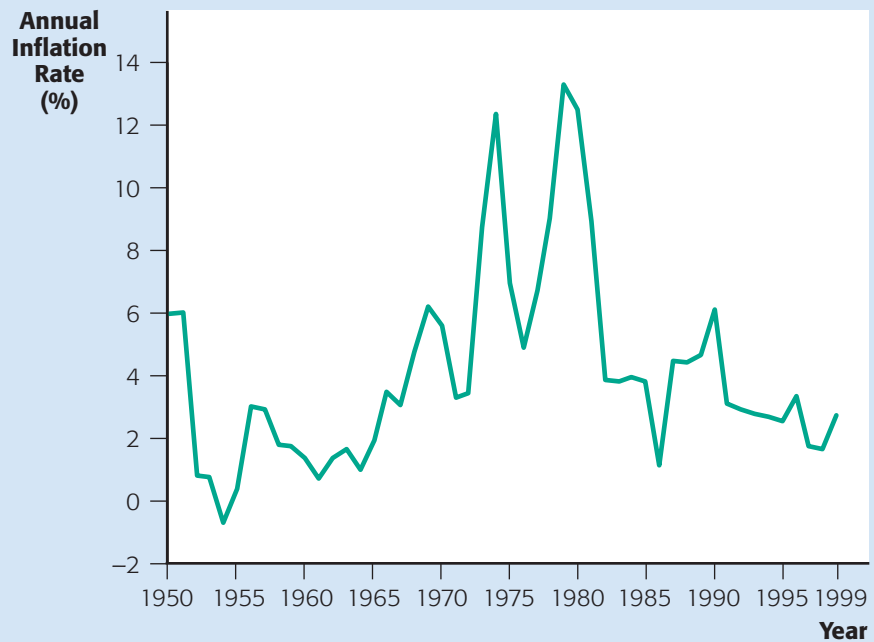
Indexation Adjusting the value of some nominal payment in proportion to a price index, in order to keep the real payment unchanged.



People often confuse the statement “prices are rising” with the statement “inflation is rising,” but they do not mean the same thing. Remember that the inflation rate is the rate of *change* of the price level. To have rising inflation, the price level must be rising by a greater and greater percentage each period. But we can also have rising prices and *falling* inflation. For example, from 1996 to 1998, the CPI rose each year—“prices were rising.” But they rose by a smaller percentage each year than the year before, so “inflation was falling”—from 3.3 percent, to 2.7 percent, and, finally, to 1.6 percent in 1998.

FIGURE 1

THE RATE OF INFLATION USING THE CONSUMER PRICE INDEX



indexed to the CPI. About one-quarter of all union members—more than 5 million workers—have labor contracts that index their wages to the CPI. Since the 1980s, the U.S. income tax has been indexed as well—the threshold income levels at which tax rates change automatically rise at the same rate as the CPI. And the government now sells bonds that are indexed to the CPI. The owner of an indexed bond receives a payment each year to make up for the loss of purchasing power when the CPI rises.

To Translate from Nominal to Real Values. In order to compare economic values from different periods, we must translate **nominal variables**—measured in the number of dollars—into **real variables**, which are adjusted for the change in the dollar’s purchasing power. The CPI is often used for this translation. Since calculating real variables is one of the most important uses of the CPI, we devote the next section to that topic.

Nominal variable A variable measured in current dollars.

Real variable A variable measured in terms of purchasing power.

REAL VARIABLES AND ADJUSTMENT FOR INFLATION

Suppose that from December 2001 to December 2002, your nominal wage—what you are paid in dollars—rises from \$15 to \$30 per hour. Are you better off? That depends. You are earning twice as many dollars. But you should care not about how many green pieces of paper you earn, but how many goods and services you can buy with that paper. How, then, can we tell what happened to your purchasing power? By focusing not on the *nominal wage*—the number of *dollars* you earn—but on the *real wage*—the *purchasing power* of your wage. To track your real wage, we need to look at the number of dollars you earn *relative to the price level*.

Since the “typical worker” and the “typical consumer” are pretty much the same, the CPI is usually the price index used to calculate the real wage. The real-wage formula is as follows:

$$\text{Real wage in any year} = \frac{\text{Nominal wage in that year}}{\text{CPI in that year}} \times 100.$$

To see that this formula makes sense, let's go back to our fictional example: From 2001 to 2002, your nominal wage doubles from \$15 to \$30. Now, suppose the price of everything that you buy doubles at the same time. It is easy to see that in this case, your purchasing power would remain unchanged. And that is just what our formula tells us: If prices double, the CPI doubles as well. With 2001 as our base year, the CPI would increase from 100 in 2001 to 200 in the year 2002. The *real* wage would be $(\$15/100) \times 100 = \15 in 2001 and $(\$30/200) \times 100 = \15 in 2002. The real wage would remain unchanged.

Now suppose that prices doubled between 2001 and 2002, but your nominal wage remained unchanged at \$15. In this case, your purchasing power would be cut in half. You'd have the same number of dollars, but each one would buy half as much as it did before. Our formula gives us a real wage of $(\$15/100) \times 100 = \15 in 2001 and $(\$15/200) \times 100 = \7.50 in 2002. The real wage falls by half.

Now look at Table 2, which shows the average hourly earnings of wage earners over the past four decades. In the first two columns, you can see that the average American wage earner was paid \$4.67 per hour in December 1975, and almost triple that—\$13.46—in December 1999. Does this mean the average hourly worker was paid more in 1999 than in 1975? In *dollars*, the answer is clearly yes. But what about in *purchasing power*? Or, using the new terminology you've learned: What happened to the *real wage* over this period?

Let's see. We know that the *nominal wage* rose from \$4.67 in 1975 to \$13.46 in 1999. But—from the table—we also know that the CPI rose from 55.5 to 168.3 over the same period. Using our formula, we find that

$$\text{Real wage in 1975} = \frac{\$4.67}{55.5} \times 100 = \$8.41.$$

$$\text{Real wage in 1999} = \frac{\$13.46}{168.3} \times 100 = \$7.99.$$

Thus, although the average worker earned more *dollars* in 1999 than in 1975, when we use the CPI as our measure of prices, her purchasing power seems to have fallen

TABLE 2

Year	Nominal Wage, Dollars per Hour	CPI	Real Wage in 1983 Dollars per Hour
1960	2.05	29.8	6.88
1965	2.50	31.8	7.86
1970	3.31	39.8	8.32
1975	4.67	55.5	8.41
1980	6.94	86.3	8.04
1985	8.72	109.3	7.98
1990	10.17	133.8	7.60
1995	11.60	153.5	7.56
1999	13.46	168.3	8.00

NOMINAL AND
REAL WAGES

over those years. *Why* this apparent decline in purchasing power? This is an interesting and important question, and one we'll begin to answer later in the chapter. The important point to remember here is that

when we measure changes in the macroeconomy, we usually care not about the number of dollars we are counting, but the purchasing power those dollars represent. Thus, we translate nominal values into real values using the formula

$$\text{real value} = \frac{\text{nominal value}}{\text{price index}} \times 100.$$

This formula, usually using the CPI as the price index, is how most real values in the economy are calculated. But there is one important exception: To calculate real GDP, the government uses a different procedure, to which we now turn.

INFLATION AND THE MEASUREMENT OF REAL GDP

In the previous chapter, we discussed the difference between nominal GDP and real GDP. After reading this chapter, you might think that real GDP is calculated just like the real wage: dividing nominal GDP by the consumer price index. But the consumer price index is *not* used to translate nominal GDP figures into real GDP figures. Instead, a special price index—which we can call the **GDP price index**—is calculated for GDP.

The most important differences between the CPI and the GDP price index are in the types of goods and services covered by each index. First, the GDP price index *includes* some prices that the CPI ignores. In particular, while the CPI tracks only the prices of goods bought by American *consumers*, the GDP price index must also include the prices of goods and services purchased by the government, investment goods purchased by businesses, and exports, which are purchased by foreigners.

Second, the GDP price index *excludes* some prices that are part of the CPI. In particular, the GDP price index leaves out used goods and imports, both of which are included in the CPI. This makes sense, because while used goods and imports are part of the typical consumer's market basket, they do not contribute to current U.S. GDP.

We can summarize the chief difference between the CPI and the GDP price index this way:

The GDP price index measures the prices of all goods and services that are included in U.S. GDP, while the CPI measures the prices of all goods and services bought by U.S. households.³

THE COSTS OF INFLATION

A high or even moderate rate of inflation—whether it is measured by the CPI or the GDP price index—is never welcome news. What's so bad about inflation? As we've seen, it certainly makes your task as an economics student more difficult: Rather than taking nominal variables at face value, you must do those troublesome calculations to convert them into real variables.

³ The technical name for the GDP price index is the *chain-type annual weights GDP price index*. It differs from the CPI not only in goods covered, but also in its mathematical formula.

GDP price index An index of the price level for all final goods and services included in GDP.

But inflation causes much more trouble than this. It can impose costs on society, and on each of us individually. Yet when most people are asked *what* the cost of inflation is, they come up with an incorrect answer.

THE INFLATION MYTH

Most people think that inflation—merely by making goods and services more expensive—erodes the average purchasing power of income in the economy. The reason for this belief is easy to see: The higher the price level, the fewer goods and services a given income will buy. It stands to reason, then, that inflation—which raises prices—must be destroying the purchasing power of our incomes. Right?

Actually, this statement is mostly wrong.

To see why, remember that every market transaction involves *two* parties—a buyer and a seller. When a price rises, buyers of that good must pay more, but sellers get more revenue when they sell it. The loss in buyers' real income is matched by the rise in sellers' real income. Inflation may *redistribute* purchasing power among the population, but it does not change the *average* purchasing power, when we include both buyers and sellers in the average.

In fact, most people in the economy participate on both sides of the market. On the one hand, they are consumers—as when they shop for food or clothing or furniture. On the other hand, they work in business firms that *sell* products, and may benefit (in the form of higher wages or higher profits) when their firms' incomes rise. Thus, when prices rise, a particular person may find that her purchasing power has either risen or fallen, depending on whether she is affected more as a seller or as a buyer. But regardless of the outcome for individuals, our conclusion remains the same:

Inflation can redistribute purchasing power from one group to another, but it cannot—by itself—decrease the average real income in the economy.

Why, then, do people continue to believe that inflation robs the average citizen of real income? Largely because real incomes sometimes do decline—for *other* reasons. Inflation—while not the *cause* of the decline—will often be the *mechanism* that brings it about. Just as we often blame the messenger for bringing bad news, so too, we often blame inflation for lowering our purchasing power when the real cause lies elsewhere.

Let's consider an example. In Table 2, notice the decline in real wages during the late 1970s. The real wage fell from \$8.41 in 1975 to \$8.04 in 1980—a decline of more than 4 percent. During this period, not only wage earners, but also salaried workers, small-business owners, and corporate shareholders all suffered a decline in their real incomes. What caused the decline?

There were several reasons, but one of the most important was the dramatic rise in the price of imported oil—from \$3 per barrel in 1973 to \$34 in 1981, an increase of more than 1,000 percent. The higher price for oil meant that oil-exporting countries, like Saudi Arabia, Kuwait, and Iraq, got more goods and services for each barrel of oil they supplied to the rest of the world, including the United States. But with these nations claiming more of America's output, less remained for the typical American. That is, the typical American family had to suffer a decline in real income. As always, a rise in price shifted income from buyers to sellers. But in this case, the sellers were foreigners, while the buyers were Americans. Thus, the rise in the price of foreign oil caused average purchasing power in the United States to decline.

But what was the mechanism that brought about the decline? Since real income is equal to $(\text{nominal income}/\text{price index}) \times 100$, it can decrease in one of two ways: a fall in the numerator (nominal income) or a rise in the denominator (the price index).

The decline in real income in the 1970s was all from the denominator. Look back at Figure 1. You can see that this period of declining real wages in the United States was also a period of unusually high inflation; at its peak in 1979, the inflation rate exceeded 13 percent. As a result, most workers blamed *inflation* for their loss of purchasing power. But inflation was not the *cause*; it was just the *mechanism*. The cause was a change in the terms of trade between the United States and the oil exporting countries—a change that resulted in higher oil prices.

To summarize, the common idea that inflation imposes a cost on society by decreasing average real income in the economy is incorrect. But inflation *does* impose costs on society, as the next section shows.

THE REDISTRIBUTIVE COST OF INFLATION

One cost of inflation is that it often redistributes purchasing power *within* society. But because the winners and losers are chosen haphazardly—rather than by conscious social policy—the redistribution of purchasing power is not generally desirable. In some cases, the shift in purchasing power is downright perverse—harming the needy and helping those who are already well off.

How does inflation sometimes redistribute real income? An increase in the price level reduces the purchasing power of any payment that is specified in nominal terms. For example, some workers have contracts that set their nominal wage for two or three years, regardless of any future inflation. The nationally set minimum wage, too, is set for several years and specified in nominal dollars. Under these circumstances, inflation can harm ordinary workers, since it erodes the purchasing power of their pre-specified nominal wage. Real income is redistributed from these workers to their employers, who benefit by paying a lower real wage. But the effect can also work the other way: benefiting ordinary households and harming businesses. For example, many homeowners sign fixed-dollar mortgage agreements with a bank. These are promises to pay the bank the same nominal sum each month. Inflation can reduce the *real* value of these payments, thus redistributing purchasing power away from the bank and toward the average homeowner.

In general,

inflation can shift purchasing power away from those who are awaiting future payments specified in dollars, and toward those who are obligated to make such payments.

But does inflation *always* redistribute income from one party in a contract to another? Actually, no; if the inflation is *expected* by both parties, it should not redistribute income. The next section explains why.

Expected Inflation Need Not Shift Purchasing Power. Suppose a labor union is negotiating a three-year contract with an employer, and both sides agree that each year, workers should get a 3-percent increase in their real wage. Labor contracts, like most other contracts, are usually specified in nominal terms: The firm will agree to give workers so many additional *dollars per hour* each year. If neither side anticipates any inflation, they should simply negotiate a 3-percent *nominal* wage hike. With an unchanged price level, the *real* wage would then also rise by the desired 3 percent.

But suppose instead that both sides anticipate 10-percent inflation each year for the next three years. Then, they must agree to *more* than a 3-percent nominal wage increase in order to raise the real wage by 3 percent. How much more?

We can answer this question with a simple mathematical rule:

Over any period, the percentage change in a real value (% Δ Real) is approximately equal to the percentage change in the associated nominal value (% Δ Nominal) minus the rate of inflation:

$$\% \Delta \text{Real} = \% \Delta \text{Nominal} - \text{Rate of inflation.}$$

If the inflation rate is 10 percent, and the real wage is to rise by 3 percent, then the change in the nominal wage must satisfy the equation

$$3 \text{ percent} = \% \Delta \text{Nominal} - 10 \text{ percent} \implies \% \Delta \text{Nominal} = 13 \text{ percent.}$$

The required nominal wage hike is 13 percent.

You can see that as long as both sides correctly anticipate the inflation, and no one stops them from negotiating a 13-percent nominal wage hike, inflation will *not* affect either party in real terms:

If inflation is fully anticipated, and if both parties take it into account, then inflation will not redistribute purchasing power.

We come to a similar conclusion about contracts between lenders and borrowers. When you lend someone money, you receive a reward—an interest payment—for letting that person use your money instead of spending it yourself. The annual *interest rate* is the interest payment divided by the amount of money you have lent. For example, if you lend someone \$1,000 and receive back \$1,040 one year later, then your interest payment is \$40, and the interest *rate* on the loan is $\$40/\$1,000 = 0.04$, or 4 percent.

But there are actually *two* interest rates associated with every loan. One is the **nominal interest rate**—the percentage increase in the lender's *dollars* from making the loan. The other is the **real interest rate**—the percentage increase in the lender's *purchasing power* from making the loan. It is the *real* rate—the change in purchasing power—that lenders and borrowers should care about.

In the absence of inflation, real and nominal interest rates would always be equal. A 4-percent increase in the lender's *dollars* would always imply a 4-percent increase in her purchasing power. But if there is inflation, it will reduce the purchasing power of the money paid back. Does this mean that inflation redistributes purchasing power? Not if the inflation is correctly anticipated, and if there are no restrictions on making loan contracts.

For example, suppose both parties anticipate inflation of 5 percent and want to arrange a contract whereby the lender will be paid a 4-percent *real* interest rate. What *nominal* interest rate should they choose? Since an interest rate is the *percentage change* in the lender's funds, we can use our approximation rule,

$$\% \Delta \text{Real} = \% \Delta \text{Nominal} - \text{Rate of inflation}$$

which here becomes

$$\% \Delta \text{ in Lender's purchasing power} = \% \Delta \text{ in Lender's dollars} - \text{Rate of inflation}$$

or

$$\text{Real interest rate} = \text{Nominal interest rate} - \text{Rate of inflation.}$$

Nominal interest rate The annual percent increase in a lender's *dollars* from making a loan.

Real interest rate The annual percent increase in a lender's *purchasing power* from making a loan.

In our example, where we want the real interest rate to equal 4 percent when the inflation rate is 5 percent, we must have

$$4 \text{ percent} = \text{Nominal interest rate} - 5 \text{ percent}$$

or

$$\text{Nominal interest rate} = 9 \text{ percent.}$$

Once again, we see that as long as both parties correctly anticipate the inflation rate, and face no restrictions on contracts (that is, they are free to set the nominal interest rate at 9 percent), then no one gains or loses.

When inflation is *not* correctly anticipated, however, our conclusion is very different.

Unexpected Inflation Does Shift Purchasing Power. Suppose that, expecting no inflation, you agree to lend money at a 4-percent nominal interest rate for one year. You and the borrower think that this will translate into a 4-percent real rate. But it turns out you are both wrong: The price level actually rises by 3 percent, so the *real* interest rate ends up being $4\% - 3\% = 1\%$. As a lender, you have given up the use of your money for the year, expecting to be rewarded with a 4-percent increase in purchasing power. But you get only a 1-percent increase. Your borrower was willing to pay 4 percent in purchasing power, but ends up paying only 1 percent. *Unexpected* inflation has led to a better deal for your borrower and a worse deal for you.

That will not make you happy. But it could be even worse. Suppose the inflation rate is higher—say, 6 percent. Then your real interest rate ends up at $4\% - 6\% = -2\%$ —a negative real interest rate. You get back *less* in purchasing power than you lend out—*paying* (in purchasing power) for the privilege of lending out your money. The borrower is *rewarded* (in purchasing power) for borrowing!

Negative real interest rates like this are not just a theoretical possibility. In the late 1970s, when inflation turned out to be higher than expected for several years in a row, many borrowers ending up paying negative rates to lenders.

Now, let's consider one more possibility: Expected inflation is 6 percent, so you negotiate a 10-percent nominal rate, thinking this will translate to a 4 percent real rate. But the actual inflation rate turns out to be zero, so the real interest rate is 10 percent $- 0$ percent = 10 percent. In this case, inflation turns out to be *less* than expected, so the *real* interest rate is higher than either of you anticipated. The borrower is harmed, and you (the lender) benefit.

These examples apply, more generally, to any agreement on future payments: to a worker waiting for a wage payment and the employer who has promised to pay it; to a doctor who has sent out a bill and the patient who has not yet paid it; or to a supplier who has delivered goods and his customer who hasn't yet paid for them.

When inflationary expectations are inaccurate, purchasing power is shifted between those obliged to make future payments and those waiting to be paid. An inflation rate higher than expected harms those awaiting payment and benefits the payers; an inflation rate lower than expected harms the payers and benefits those awaiting payment.

THE RESOURCE COST OF INFLATION

In addition to its possible redistribution of income, inflation imposes another cost upon society. To cope with inflation, we are forced to use up time and other resources as we go about our daily economic activities (shopping, selling, saving) that we could otherwise have devoted to productive activities. Thus, inflation imposes an *opportunity cost* on society as a whole and on each of its members:

When people must spend time and other resources coping with inflation, they pay an opportunity cost—they sacrifice the goods and services those resources could have produced instead.

Let's first consider the resources used up by *consumers* to cope with inflation. Suppose you shop for clothes twice a year. You've discovered that both The Gap and Banana Republic sell clothing of similar quality and have similar service, and you naturally want to shop at the one with the lower prices. If there is no inflation, your task is easy: You shop first at The Gap and then at Banana Republic; thereafter, you rely on your memory to determine which is less expensive.

With inflation, however, things are more difficult. Suppose you find that prices at Banana Republic are higher than you remember them to be at The Gap. It may be that Banana Republic is the more expensive store, or it may be that prices have risen at *both* stores. How can you tell? Only a trip back to The Gap will answer the question—a trip that will cost you extra time and trouble. If prices are rising very rapidly, you may have to visit both stores on the same day to be sure which one is cheaper. Now, multiply this time and trouble by all the different types of shopping you must do on a regular or occasional basis—for groceries, an apartment, a car, concert tickets, compact discs, restaurant meals, and more. Inflation can make you use up valuable time—time you could have spent earning income or enjoying leisure activities. True, if you shop for some of these items on the Internet, you can compare prices in less time, but not zero time. And most shopping is *not* done over the Internet.

Inflation also forces *sellers* to use up resources. First, remember that sellers of goods and services are also buyers of resources and intermediate goods. They, too, must do comparison shopping when there is inflation, and use up hired labor time in the process. Second, each time sellers raise prices, labor is needed to put new price tags on merchandise, to enter new prices into a computer scanning system, to update the HTML code on a web page, or to change the prices on advertising brochures, menus, and so on.

Finally, inflation makes us all use up resources managing our financial affairs. We'll try to keep our funds in accounts that pay high nominal interest rates, in order to preserve our purchasing power, and minimize what we keep as cash or in low-interest checking accounts. Of course, this means more frequent trips to the bank or the automatic teller machine, to transfer money into our checking accounts or get cash each time we need it.

All of these additional activities—inspecting prices at several stores or Web sites, changing price tags or price entries, going back and forth to the automatic teller machine—use up not only time, but other resources too, such as gasoline, paper, or the wear and tear on your computer. From society's point of view, these resources could have been used to produce *other* goods and services that we'd enjoy.

You may not have thought much about the resource cost of inflation, because in recent years, U.S. inflation has been so low—under 3 percent per year in the 1990s. Such a low rate of inflation is often called *creeping inflation*—from week to week



To learn more about the strengths and weaknesses of the CPI, read Allison Wallace and Brian Motley, "A Better CPI" (<http://www.frbsf.org/econsrch/wklyltr/wklytr99/el99-05.html>).

or month to month, the price level creeps up so slowly that we hardly notice the change. The cost of coping with creeping inflation is negligible.

But it has not always been this way. Three times during the last 50 years, we have had double-digit inflation—about 14 percent during 1947–48, 12 percent in 1974, and 13 percent during 1979 and 1980. Going back farther, the annual inflation rate reached almost 20 percent during World War I and rose above 25 percent during the Civil War.

And as serious as these episodes of American inflation have been, they pale in comparison to the experiences of other countries. In Germany in the early 1920s, the inflation rate hit thousands of percent *per month*. And more recently—in the late 1980s—several South American countries experienced inflation rates in excess of 1,000 percent annually. For a few weeks in 1990, Argentina's annual inflation rate even reached 400,000 percent! Under these conditions, the monetary system breaks down almost completely. Economic life is almost as difficult as in Chaotica.

IS THE CPI ACCURATE?

Using the THEORY



The Bureau of Labor Statistics spends millions of dollars gathering data to ensure that its measure of inflation is accurate. To determine the market basket of the typical consumer every 10 years or so, the BLS randomly selects thousands of households and analyzes their spending habits. In the last household survey—completed in 1993–95—each of about 15,000 families kept diaries of their purchases for two weeks.

But that is just the beginning. Every month, the bureau's shoppers visit 23,000 retail stores, 7,000 rental apartments, and 18,000 owner-occupied homes to record 71,000 different prices. Finally, all of the prices are combined to determine the cost of the typical consumer's market basket for the current month.

The BLS is a highly professional agency, typically headed by an economist. Billions of dollars are at stake for each 1-percent change in the CPI, and the BLS deserves high praise for keeping its measurement honest and free of political manipulation. Nevertheless, conceptual problems and resource limitations make the CPI fall short of the ideal measure of inflation. Economists—even those who work in the BLS—widely agree that the CPI overstates the U.S. inflation rate. By how much?

According to a report by an advisory committee of economists appointed by the Senate Finance Committee in 1996, the overall bias has been at least 1.1 percent annually in recent years.⁴ That is, in a typical year, the reported rise in the CPI has been about 1 percentage point greater than the true rise in the price level. The BLS has been working hard to reduce this upward bias, and—especially in the late 1990s—it made some progress. But significant bias remains.

SOURCES OF BIAS IN THE CPI

There are several reasons for the upward bias in the CPI.

Substitution Bias. Until recently, the CPI almost completely ignored a general principle of consumer behavior: People tend to *substitute* goods that have become

⁴ See *Toward a More Accurate Measure of the Cost of Living*, Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index, December 1996.

relatively cheaper in place of those that have become relatively more expensive. For example, in the seven years from 1973 to 1980, the retail price of oil-related products—like gasoline and home heating oil—increased by more than 300 percent, while the prices of most other goods and services rose by less than 100 percent. As a result, people found ways to conserve on oil products. They joined carpools, used public transportation, insulated their homes, and in many cases moved closer to their workplaces to shorten their commute. Yet throughout this period, the CPI basket—based on a survey of buying patterns in 1972–73—assumed that consumers were buying unchanged quantities of oil products.

The treatment of oil products is an example of a more general problem that has plagued the CPI for decades. Until recently, the CPI strictly followed a procedure of using fixed *quantities* to determine the relative importance of each item. That is, it assumed that households continued to buy each good or service in the same quantities at which they bought it during the last household survey. Compounding the problem, the survey to determine spending patterns—and to update the market basket—was taken only about once every 10 years or so. So by the end of each 10-year period, the CPI's assumptions about spending habits could be far off the mark, as they were in the case of oil in the 1970s.

The BLS has *partially* fixed this problem, in two ways.⁵ First, beginning in 2002, it will update the market basket with a household survey every *two* years instead of every 10 years. This is widely considered an important improvement in CPI measurement.

Second, as of January 1999, the CPI no longer assumes that the typical consumer continues to buy the same *quantity* of each good that he bought in the last household “market basket” survey. Instead, the CPI assumes that when a good's relative price rises by 10 percent, typical consumers buy 10 percent less of it, and switch their purchases to other goods whose prices are rising more slowly.

However, this is only a partial fix. The CPI still only recognizes the possibility of such substitution *within* categories of goods, and *not among* them. For example, if the price of steak rises relative to the price of hamburger meat, the CPI now assumes that consumers will substitute away from steak and toward hamburger meat, since both are in the same category: *beef*. However, if the price of all beef products rises relative to chicken and pork, the CPI assumes that there is *no* substitution at all from beef toward chicken and pork. As a result, beef products will be over-weighted in the CPI until the next survey.

Although the BLS has partially fixed the problem, the CPI still suffers from substitution bias. That is, categories of goods whose prices are rising most rapidly tend to be given exaggerated importance in the CPI, and categories of goods whose prices are rising most slowly tend to be given too little importance in the CPI.

New Technologies. Brand-new technologies are another source of upward bias in the CPI. One problem is that goods using new technologies are introduced into the BLS market basket only after a lag. These goods often drop rapidly in price after they are introduced, helping to balance out price rises in other goods. By excluding

⁵ For a discussion of these and other recent changes in the CPI, see “Planned Change in the Consumer Price Index Formula,” Bureau of Labor Statistics, April 16, 1998 (<http://stats.bls.gov/cpigm02.htm>) and “Future Schedule for Expenditure Weight Updates in the Consumer Price Index,” Bureau of Labor Statistics, December 18, 1998 (<http://stats.bls.gov/cpiupdt.htm>).

a category of goods whose prices are dropping, the CPI overstates the rate of inflation. For example, even though many consumers were buying and using cellular phones throughout the 1990s, they were not included in the BLS basket of goods until 1998. As a result, the CPI missed the rapid decline in the price of cell phones. Updating the market basket every two years—instead of every 10—should reduce this source of bias after 2002.

But there is another issue with new technologies: They often offer consumers a lower-cost alternative for obtaining the same service. For example, the introduction of cable television lowered the cost of entertainment significantly by offering a new, cheaper alternative to going out to see movies. This should have registered as a drop in the price of “seeing movies.” But the CPI does not have any good way to measure this reduction in the cost of living. Instead, it treats cable television as an entirely separate service.

The CPI excludes new products that tend to drop in price when they first come on the market. When included, the CPI regards them as entirely separate from existing goods and services, instead of recognizing that they lower the cost of achieving a given standard of living. The result is an overestimate of the inflation rate.

Changes in Quality. Many products are improving over time. Cars are much more reliable than they used to be and require much less routine maintenance. They have features like air bags and antilock brakes that were unknown in the early 1980s. The BLS struggles to deal with these changes. It knows that when cars become more expensive, some of the rise in price is not really inflation, but rather charging more because the consumer is *getting* more. In addition to cars, the BLS has recently developed sophisticated statistical techniques to account for quality improvements in computers and peripherals, clothing, and rental apartments. In January 1999, televisions were added to the list, and in coming years, the BLS hopes to extend the techniques to even more categories. Thus, slowly but surely, the BLS is planning to chip away at the upward bias in inflation caused by unmeasured quality improvements.

But in the meantime, many improvements in quality are still ignored by the CPI. When food prices rise due to better nutritional quality, when VCR prices rise due to better performance and convenience, or when the cost of surgery rises due to more sophisticated techniques that have greater success rates, the CPI merely records a price increase, as if the same thing is costing more.

The CPI still fails to recognize that, in many cases, prices rise because of improvements in quality, not because the cost of living has risen. This causes the CPI to overstate the inflation rate.

Growth in Discounting. The CPI treats toothpaste bought at a high-priced drugstore and toothpaste bought at Wal-Mart or Drugstore.com as different products. And it assumes that we continue to buy from high- and low-priced stores in unchanged proportions. But that is not what has been happening. In fact, Americans are buying more and more of their toothpaste and other products from discounters, but the CPI does not consider this in measuring inflation. The purchasing power you have lost from inflation is not as great as the CPI says if you, like most Americans, are stretching your dollar by going more often to discount outlets, warehouse stores, and Web sites with low prices.

The CPI omits reductions in the prices people pay from more frequent shopping at discount stores and so overstates the inflation rate.

THE CONSEQUENCES OF OVERSTATING INFLATION

The impact of overstating the inflation rate is both serious and wide ranging. First, it means that many real variables have been rising more rapidly than the official numbers suggest. For example, look again at Table 2. It tells us that, from 1975 to 1995, the average real wage *fell* from \$8.41 to \$7.56, a decrease of about 10 percent. This calculation is based on the official CPI. But suppose the CPI overstated the inflation rate by just 1.1 percentage points per year over this period, as the government's advisory commission has suggested. Then the real wage did not fall at all over this period, but actually *rose* by about 11 percent. (See Challenge Question #3 at the end of this chapter.)

Second, remember that low inflation is an important macroeconomic goal. As you'll learn in future chapters, this goal is not always easy to achieve and may require large—if temporary—sacrifices. If the CPI overstates inflation—and continues to do so in the future—we may be making these sacrifices unnecessarily: We may take painful steps to bring inflation down when the real problem is that our official inflation measure is exaggerating the problem.

Finally, since many payments are indexed to the CPI, an overstatement of inflation results in *overindexing*—payments that rise *faster* than the true price level. For example, suppose a Social Security recipient's payment of \$1,000 per month is indexed to the CPI in order to keep the real payment constant as prices rise. Suppose, too, that over 10 years, the CPI reports annual inflation of 3 percent. By the end of the period, the CPI will rise by 35 percent, and the nominal payment will rise to \$1,350.⁶ But what if the CPI is wrong, and the actual inflation rate is just 1.9 percent per year during the period? Then, using the initial year as the base period, an accurate price index will rise from 100 to 120.7. This tells us that the *real* Social Security payment will rise from \$1,000 to $(\$1,350/120.7) \times 100 = \$1,118$ —an increase of about 11 percent. This “overpayment” of \$118 per month at the end of the period may suit the Social Security recipient just fine. But remember that the rest of society pays for the retired person's gain through higher real tax payments. The same general principle applies to union workers, government pensioners, or anyone else who is overindexed due to errors in the CPI:

When a payment is indexed, and the price index overstates inflation, inflation will increase the real payment, shifting purchasing power toward those who are indexed and away from the rest of society.

THE FUTURE OF THE CPI

In the past, the CPI has mostly tracked the cost of a fixed basket of goods, and it has done a reasonably good job of doing so. But it has *not* done a good job tracking what many people call the *cost of living*—the number of dollars a person must pay in order to enjoy a given level of economic satisfaction. When people substitute cheaper goods, take advantage of new technologies, and enjoy quality improvements, they are trying to get more satisfaction for a given cost, or else trying to

⁶ Over 10 years, 3 percent annual inflation raises the CPI by a factor of $(1.03)^{10} = 1.35$.

maintain their level of satisfaction in spite of price hikes. And to some extent, they are successful. In the past, the CPI has ignored our ability to increase and preserve our satisfaction from a given amount of spending; it has *not* tried to tell us what is happening to the cost of *living*.

But this is changing. The repairs that have already been made to the CPI—and others that many economists believe the BLS should make—are moving the index closer to a *cost of living* indicator. Once we try to measure the cost of living, however, we enter into some nebulous territory. How is the cost of living affected when our medical care is provided by an HMO that lowers the price, but gives us fewer options in choosing our own doctors? When the price of new textbooks rises, how much satisfaction do people lose when they substitute cheaper, used textbooks? How should we incorporate falling crime rates that enable us to protect our lives and property at lower cost? And what about other aspects of our society that affect the quality of our lives: leisure time, the state of the environment, the safety of our workplaces, the quality of our culture, and so on? Do we want changes in these aspects of life to affect our cost-of-living measure?

For all of these reasons, fixing the CPI is controversial. Further, some groups—including Social Security recipients, union workers, and pensioners—stand to lose from any fix that will reduce the reported inflation rate. After all, these groups gain from any overestimate of inflation, because their benefits are indexed to the CPI. In fact, many Social Security recipients view suggestions to correct the CPI as a back-door effort to reduce their benefits.

Thus, the CPI has entered the realm of politics. The voices arguing for continued changes to the CPI are getting stronger, but so are the voices of those opposed.

S U M M A R Y

Money serves two important functions. First, it is a *unit of value* that helps us measure how much something is worth and compare the costs of different goods and services. Second, it is a *means of payment* by being generally acceptable in exchange for goods and services. Without money, we would be reduced to barter, a very inefficient way of carrying out transactions.

The value of money is its purchasing power, and this changes as the prices of the things we buy change. The overall trend of prices is measured using a price index. Like any index number, a price index is calculated as: $(\text{value in current period}/\text{value in base period}) \times 100$. The most widely used price index in the United States is the *Consumer Price Index (CPI)*, which tracks the prices paid for a typical consumer's "market basket." The percent change in the CPI is the inflation rate.

The most common uses of the CPI are for indexing payments, as a policy target, and to translate from nominal to real variables. Many nominal variables, such as the nominal wage, can be corrected for price changes by dividing by the CPI and then multiplying by 100. The result is a real variable,

such as the real wage, that rises and falls only when its purchasing power rises and falls. Another price index in common use is the GDP price index. It tracks prices of all final goods and services included in GDP.

Inflation—a rise over time in a price index—is costly to our society. One of inflation's costs is an arbitrary redistribution of income. Unanticipated inflation shifts purchasing power away from those awaiting future dollar payments and toward those obligated to make such payments. Another cost of inflation is the resource cost: People use valuable time and other resources trying to cope with inflation.

It is widely agreed that the CPI has overstated inflation in recent decades—probably by more than one percentage point per year. As a result, the official statistics on real variables may contain errors, and people who are indexed to the CPI have been actually overindexed, enjoying an increase in real income that is paid for by the rest of society. The Bureau of Labor Statistics has been trying to eliminate the upward bias in the CPI, but so far, it has only eliminated part of the problem. Meanwhile, fixing the CPI has become a political issue—and a controversial one.

KEY TERMS

unit of value	price level	deflation	GDP price index
means of payment	index	indexation	nominal interest rate
Federal Reserve System	Consumer Price Index	nominal variable	real interest rate
fiat money	inflation rate	real variable	

REVIEW QUESTIONS

- Distinguish between the *unit-of-value* function of money and the *means-of-payment* function. Give examples of how the U.S. dollar has played each of these two roles.
- How does the price level differ from, say, the price of a haircut or a Big Mac?
- Explain how you might construct an index of bank deposits over time. What steps would be involved?
- What is the CPI? What does it measure? How can it be used to calculate the inflation rate?
- Can the inflation rate be decreasing at the same time the price level is rising? Can the inflation rate be increasing at the same time the price level is falling? Explain.
- What are the main uses of the CPI? Give an example of each use.
- Explain the logic of the formula that relates real values to nominal values.
- What are the similarities between the CPI and the GDP price index? What are the differences?
- What are the costs of inflation?
- Under what circumstances would inflation redistribute purchasing power? How? When would it *not* redistribute purchasing power?
- How is a nominal interest rate different from a real interest rate? Which do you think is the better measure of the rate of return on a loan?

PROBLEMS AND EXERCISES

- Both gold and paper currency have served as money in the United States. What are some of the advantages of paper currency over gold?
- Which would be more costly—a steady inflation rate of 3 percent per year, or an inflation rate that was sometimes high and sometimes low, but that averaged 3 percent per year? Justify your answer.
- Given the following *year-end* data, calculate the inflation rate for years 2, 3, and 4. Calculate the real wage in each year:
- This chapter discusses the costs of inflation. Would there be any costs to a *deflation*—a period of falling prices? If so, what would they be? Give examples.
- Given the following data, calculate the real interest rate for years 2, 3, and 4. (Assume that each CPI number tells us the price level at the *end* of each year.)

Year	CPI	Inflation Rate	Nominal Wage	Real Wage
1	100	—	\$10.00	_____
2	110	_____	\$12.00	_____
3	120	_____	\$13.00	_____
4	115	_____	\$12.75	_____

Year	CPI	Nominal Interest Rate	Real Interest Rate
1	100	—	—
2	110	15%	_____
3	120	13%	_____
4	115	8%	_____

If you lent \$200 to a friend at the beginning of year 2 at the prevailing nominal interest rate of 15 percent, and your friend returned the money—with the interest—at the end of year 2, did you benefit from the deal?

6. Your friend asks for a loan of \$100 for one year and offers to pay you 5 percent interest. Your friend expects the inflation rate over that one-year period to be 6 percent; you expect it to be 4 percent. You agree to make the loan, and the actual inflation rate turns out to be 5 percent. Who benefits and who loses?
7. If there is 5 percent inflation each year for eight years, what is the *total* amount of inflation (i.e., the total percentage rise in the price level) over the entire eight-year period? (*Hint:* The answer is *not* 40 percent.)

C H A L L E N G E Q U E S T I O N S

1. Inflation is sometimes said to be a tax on nominal money holdings. If you hold \$100 and the price level increases by 10 percent, the purchasing power of that \$100 falls by about 10 percent. Who benefits from this inflation tax?
2. During the late nineteenth and early twentieth centuries, many U.S. farmers favored inflationary government policies. Why might this have been the case? (*Hint:* Do farmers typically pay for their land in full at the time of purchase?)
3. Look again at the first paragraph under the heading, “The Consequences of Overstating Inflation.” It says that if the CPI overstated the inflation rate by 1.1 percentage points each year from 1975 to 1995, then the average real wage did not fall by 10 percent as reported, but actually grew by 11 percent. Prove this statement true, using numbers (as needed) from Table 2.

E X P E R I E N T I A L E X E R C I S E

1. How has the U.S. inflation rate compared with rates in other industrial economies in recent years? To explore this question, go to the international economic trends Web page of the Federal Reserve Bank of St. Louis (<http://www.stls.frb.org/publications/iet>). Choose two nations and compare their recent inflation experiences to that of the United States. Why should we be careful in comparing inflation rates internationally?



APPENDIX

CALCULATING THE CONSUMER PRICE INDEX

The Consumer Price Index (CPI) is the government's most popular measure of inflation. It tracks the cost of the collection of goods—called the *CPI market basket*—bought by a typical consumer in some *base period*. This appendix demonstrates how the Bureau of Labor Statistics (BLS) calculates the CPI. To help you follow the steps clearly, we'll do the calculations for a very simple economy with just two goods: hamburger meat and oranges (not a pleasant world, but a manageable one). Table 3 shows prices for each good, and the quantities produced and consumed, in two different periods: December 2002 (the base period) and December 2003. The market basket (measured in the base period) is given in the third column of the table: In December 2002, the typical consumer buys 30 pounds of hamburger and 50 pounds of oranges. Our formula for the CPI in any period t is

CPI in period t

$$= \frac{\text{Cost of market basket at prices in period } t}{\text{Cost of market basket at 2002 prices}} \times 100,$$

where each year's prices are measured in December of that year.

TABLE 3

PRICES AND WEEKLY QUANTITIES IN A TWO-GOOD ECONOMY

	December 2002		December 2003	
	Price (per lb.)	Quantity (lbs.)	Price (per lb.)	Quantity (lbs.)
Hamburger Meat	\$5.00	30	\$6.00	30
Oranges	\$1.00	50	\$1.10	50

Table 4 shows the calculations we must do to determine the CPI in December 2002 and December 2003. In the table, you can see that the cost of the 2002 market basket at 2002 prices is \$200. The cost of the *same* market basket at 2003's higher prices is \$235.

TABLE 4

CALCULATIONS FOR THE CPI

	At December 2002 Prices	At December 2003 Prices
Cost of 30 lbs. of Hamburger	$\$5.00 \times 30 = \150	$\$6.00 \times 30 = \180
Cost of 50 lbs. of Oranges	$\$1.00 \times 50 = \50	$\$1.10 \times 50 = \55
Cost of Entire Market Basket	$\$150 + \$50 = \$200$	$\$180 + \$55 = \$235$

To determine the CPI in December 2002—the base period—we use the formula with period t equal to 2002, giving us

CPI in 2002

$$\begin{aligned} &= \frac{\text{Cost of 2002 basket at 2002 prices}}{\text{Cost of 2002 basket at 2002 prices}} \times 100 \\ &= \frac{\$200}{\$200} \times 100 = 100. \end{aligned}$$

That is, the CPI in December 2002—the base period—is equal to 100. (The formula, as you can see, is set up so that the CPI will always equal 100 in the base period, regardless of which base period we choose.)

Now let's apply the formula again, to get the value of the CPI in December 2003:

CPI in 2003

$$\begin{aligned} &= \frac{\text{Cost of 2002 basket at 2003 prices}}{\text{Cost of 2002 basket at 2002 prices}} \times 100 \\ &= \frac{\$235}{\$200} \times 100 = 117.5. \end{aligned}$$

From December 2002 to December 2003, the CPI rises from 100 to 117.5. The rate of inflation over the year 2003 is therefore 17.5 percent.

Notice that the CPI gives more weight to price changes of goods that are more important in the consumer's budget. In our example, the percentage rise in the CPI (17.5 percent) is closer to the percentage rise in the price of hamburger (20 percent) than it is to the percentage price rise of oranges (10 percent). This is because a greater percentage of our budget is *spent* on hamburger than on oranges, so hamburger carries more weight in the CPI.

But one of the CPI's problems, discussed in the body of the chapter, is *substitution bias*. The CPI recognizes that consumers substitute *within* categories of goods. For example, if we had a third good—steak—the CPI would recognize that consumers will buy more steak if the price of hamburger rises faster than the price of steak. But the CPI assumes there is no substitution *among* categories—between beef products and fruit, for example. No matter how much the relative price of beef products like hamburger rises, the CPI assumes that people will continue to buy the same quantity of it, rather than substitute goods in other categories like oranges. Therefore, as the price of hamburger rises, the

CPI assumes that we spend a greater and greater percentage of our budgets on it; hamburger gets *increasing weight* in the CPI. In our example, spending on hamburger is assumed to rise from $\$150/\$200 = 0.75$, or 75 percent of the typical weekly budget, to $\$180/\$235 = 0.766$, or 76.6 percent. In fact, however, the rapid rise in price would cause people to substitute *away* from hamburger toward other goods whose prices are rising more slowly. This is what occurs in our two-good example, as you can see in the last column of Table 3. In 2003, the quantity of hamburger purchased drops to 10, and the quantity of oranges rises to 100. In an ideal measure, the decrease in the quantity of hamburger would reduce its weight in determining the overall rate of inflation. But the CPI ignores this. Look back at how we've calculated the CPI in this example, and you will see that we have entirely ignored the information in the last column of Table 3, which shows the new quantities purchased in 2003. This failure to correct for substitution bias across categories of goods is one of the reasons the CPI overstates inflation.

THE CLASSICAL LONG-RUN MODEL

Economists sometimes disagree with each other. In news interviews, class lectures, and editorials, they give differing opinions about even the simplest matters. To the casual observer, it might seem that economics is little more than guesswork, where anyone's opinion is as good as anyone else's. But there is actually much more agreement among economists than there appears to be.

Take the following typical example: Two distinguished economists appear on *CNN Moneyline*. In a somber tone, Willow Bay—the anchor—asks each of them what should be done to maintain the health of the economy. “We need to cut taxes,” replies the first economist. “If individuals can keep more of what they earn, they'll have more incentive to work. And if we lower taxes on business, they'll have more incentive to invest and grow.” (Don't worry if this chain of logic isn't clear to you yet—it will be by the end of the next chapter.)

“No, no, no,” the second economist interrupts. “A tax cut would be the *worst* thing we could do right now. The economy is already pumping out just about as many goods and services as it can. A tax cut—which would put more funds into buyers' hands—would only increase spending, overheat the economy, and lead to inflationary dangers that the U.S. Federal Reserve would have to prevent.” (You'll begin learning what's behind this argument a few chapters later.)

Which of these economists is correct? Very likely, *both* of them are correct. But how can this be? Aren't the two responses contradictory? Not really, because each economist is hearing—and answering—a different question. The first economist is addressing the *long-run* impact of a cut in taxes—the impact we can expect after several years have elapsed. The second economist is focusing on the *short-run* impact—the effects we'd see over the next year.

Once the distinction between the long run and the short run becomes clear, many apparent disagreements among macroeconomists dissolve. If Willow Bay had asked our two economists about the long-run impact of cutting taxes, both may well have agreed that it would lead to more jobs and more investment by business firms. If asked about the short-run impact, both may have agreed about the potential danger of inflation. If no time horizon is specified, however, an economist is likely to focus on the horizon he or she feels is most important—something about which economists sometimes *do* disagree. The real dispute, though, is less over how the economy *works* and more about what our priorities should be in guiding it.

CHAPTER OUTLINE

**Macroeconomic Models:
Classical Versus Keynesian**

Assumptions of the Classical Model

How Much Output Will We Produce?

The Labor Market

Determining the Economy's Output

The Role of Spending

Total Spending in a Very Simple Economy

Total Spending in a More Realistic Economy

Leakages and Injections

The Loanable Funds Market

The Supply of Funds Curve

The Demand for Funds Curve

Equilibrium in the Loanable Funds Market

The Loanable Funds Market and Say's Law

**The Classical Model:
A Summary****Using the Theory: Fiscal Policy
in the Classical Model**

Fiscal Policy with a Budget Surplus

Ideally, we would like our economy to do well in both the long run and the short run. Unfortunately, there is often a trade-off between these two goals: Doing better in the short run can require some sacrifice of long-run goals, and vice versa. The problem for policymakers is much like that of the captain of a ship sailing through the North Atlantic. On the one hand, he wants to reach his destination (his long-run goal); on the other hand, he must avoid icebergs along the way (his short-run goal). As you might imagine, avoiding icebergs may require the captain to deviate from an ideal long-run course. At the same time, reaching port might require risking the occasional iceberg.

The same is true of the macroeconomy. If you flip back two chapters and look at Figure 4, you will see that there are two types of movements in total output—the long-run trajectory showing the growth of potential output and the short-run movements around that trajectory, which we call economic fluctuations or business cycles. Macroeconomists are concerned with both types of movements. But, as you will see, policies that can help us smooth out economic fluctuations may prove harmful to growth in the long run, while policies that promise a high rate of growth might require us to put up with more severe fluctuations in the short run.

MACROECONOMIC MODELS: CLASSICAL VERSUS KEYNESIAN

Classical model A macroeconomic model that explains the long-run behavior of the economy, assuming that all markets clear.

The **classical model**, developed by economists in the nineteenth and early twentieth centuries, was an attempt to explain a key observation about the economy: Over periods of several years or longer, the economy performs rather well. That is, if we step back from current conditions and view the economy over a long stretch of time, we see that it operates reasonably close to its potential output. And even when it deviates, it does not do so for very long. Business cycles may come and go, but the economy eventually returns to full employment. Indeed, if we think in terms of decades rather than years or quarters, the business cycle fades in significance much like the waves in a choppy sea disappear when viewed from a jet plane.

In the classical view, this behavior is no accident: Powerful forces are at work that drive the economy toward full employment. Many of the classical economists went even further, arguing that these forces operated within a reasonably short period of time. And even today, an important group of macroeconomists continues to believe that the classical model is useful even in the shorter run.

Until the Great Depression of the 1930s, there was little reason to question these classical ideas. True, output fluctuated around its trend, and from time to time there were serious recessions, but output always returned to its potential, full-employment level within a few years or less, just as the classical economists predicted. But during the Great Depression, output was stuck far below its potential for many years. For some reason, the economy wasn't working the way the classical model said it should.

In 1936, in the midst of the Great Depression, the British economist John Maynard Keynes offered an explanation for the economy's poor performance. His new model of the economy—soon dubbed the *Keynesian model*—changed many economists' thinking.¹ Keynes and his followers argued that, while the classical model

¹ Keynes's attack on the classical model was presented in his book *The General Theory of Employment, Interest and Money* (1936). Unfortunately, it's a very difficult book to read, though you may want to try. Keynes's assumptions were not always clear, and some of his text is open to multiple interpretations. As a result, economists have been arguing for decades about what Keynes really meant.

might explain the economy's operation in the long run, the long run could be a very long time in arriving. In the meantime, production could be stuck below its potential, as it seemed to be during the Great Depression.

Keynesian ideas became increasingly popular in universities and government agencies during the 1940s and 1950s. By the mid-1960s, the entire profession had been won over: Macroeconomics *was* Keynesian economics, and the classical model was removed from virtually all introductory economics textbooks. You might be wondering, then, why we are bothering with the classical model here. After all, it's an older model of the economy, one that was largely discredited and replaced, just as the Ptolemaic view that the sun circled the earth was supplanted by the more modern, Copernican view. Right?

Not really. The classical model is still important, for two reasons. First, in recent decades, there has been an active counterrevolution against Keynes's approach to understanding the macroeconomy. Many of the counterrevolutionary new theories are based largely on classical ideas. In some cases, the new theories are just classical economics in modern clothing, but in other cases significant new ideas have been added. By studying classical macroeconomics, you will be better prepared to understand the controversies centering on these newer schools of thought.

The second—and more important—reason for us to study the classical model is its usefulness in understanding the economy over the long run. Even the many economists who find the classical model inadequate for understanding the economy in the short run find it extremely useful in analyzing the economy in the long run.

While Keynes's ideas and their further development help us understand economic fluctuations—movements in output around its long-run trend—the classical model has proven more useful in explaining the long-run trend itself.

This is why we will use the terms “classical view” and “long-run view” interchangeably in the rest of the book; in either case, we mean “the ideas of the classical model used to explain the economy's long-run behavior.”

ASSUMPTIONS OF THE CLASSICAL MODEL

Remember from Chapter 1 that all models begin with *assumptions* about the world. The classical model is no exception. Many of the assumptions are merely simplifying—they make the model more manageable, enabling us to see the broad outlines of economic behavior without getting lost in the details. Typically, these assumptions involve aggregation, such as ignoring the many different interest rates in the economy and instead referring to a single interest rate, or ignoring the many different types of labor in the economy and analyzing instead a single aggregate labor market. These simplifications are usually harmless—adding more detail would make our work more difficult, but would not add much insight, nor would it change any of the central conclusions of the classical view.

There is, however, one assumption in the classical view that goes beyond mere simplification. This is an assumption about how the world works, and it is critical to the conclusions we will reach in this and the next chapter. We can state it in two words: *markets clear*.

*A critical assumption in the classical model is that **markets clear**: The price in every market will adjust until quantity supplied and quantity demanded are equal.*

Market clearing Adjustment of prices until quantities supplied and demanded are equal.

Does the market-clearing assumption sound familiar? It should: It was the basic idea behind our study of supply and demand. When we look at the economy through the classical lens, we assume that the forces of supply and demand work fairly well throughout the economy and that markets do reach equilibrium. An excess supply of anything traded will lead to a fall in its price; an excess demand will drive the price up.

The market-clearing assumption, which permeates classical thinking about the economy, provides an early hint about why the classical model does a better job over longer time periods (several years or more) than shorter ones. In many markets, prices might not fully adjust to their equilibrium values for many months or even years after some change in the economy. An excess supply or excess demand might persist for some time. Still, if we wait long enough, an excess supply in a market will eventually force the price down, and an excess demand will eventually drive the price up. That is, *eventually*, the market will clear. Therefore, when we are trying to explain the economy's behavior over the long run, market clearing seems to be a reasonable assumption.

In the remainder of the chapter, we'll use the classical model to answer a variety of important questions about the economy in the long run, such as:

- How is total employment determined?
- How much output will we produce?
- What role does total spending play in the economy?
- What happens when things change?

Keep in mind that, in our discussion of the classical model, we will focus on *real* variables: real GDP, the real wage, real saving, and so on. These variables are typically measured in the dollars of some base year, and their numerical values change only when their *purchasing power* changes.

HOW MUCH OUTPUT WILL WE PRODUCE?

Over the last decade, on average, the U.S. economy produced about \$7.5 trillion worth of goods and services per year (valued in 1996 dollars). How was this average level of output determined? Why didn't we produce \$10 trillion per year? Or just \$2 trillion? There are so many things to consider when answering this question—variables you constantly hear about in the news—wages, interest rates, investment spending, government spending, taxes, and more. Each of these concepts plays an important role in determining total output, and our task in this chapter is to show how they all fit together.

But what a task! How can we disentangle the complicated web of economic interactions we see around us? Our starting point will be the first step of our *four-step procedure*, introduced toward the end of Chapter 3. To review, that first step was to *characterize the market*—to decide which market or markets best suit the problem being analyzed, and then identify the buyers and sellers who interact in that market.

But which market should we start with?

The classical approach is to start at the beginning, with the *reason* for all this production in the first place. In the classical view, all production arises from one source: our desires for goods and services. Of course, we cannot buy goods and services if we don't have income. And with that fact comes an important implication:

In order to earn income so we can buy goods and services, we must supply labor and other resources to firms.

THE LABOR MARKET

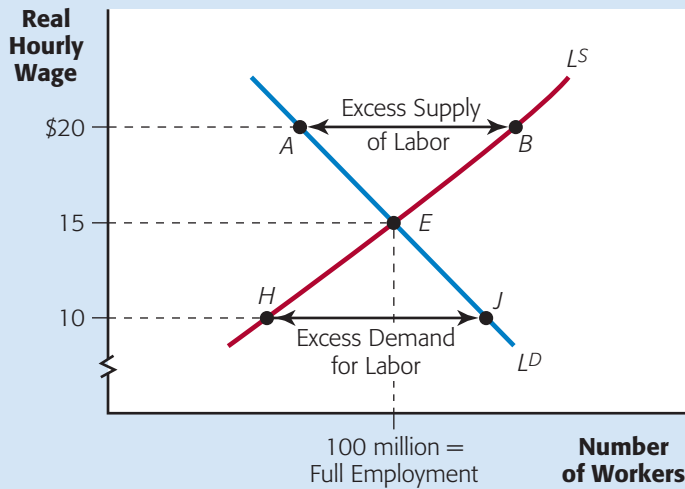


FIGURE 1

The equilibrium wage rate of \$15 per hour is determined at point *E*, where the upward-sloping labor supply curve crosses the downward-sloping labor demand curve. At any other wage, an excess demand or excess supply of labor will cause an adjustment back to equilibrium.

Thus, a logical place to start is with the markets for resources—markets for labor, land, and capital. To keep things simple, however, we'll concentrate our attention on just one type of resource—labor. In our classical world, we assume that firms are making use of all the capital and land that are available in the economy. The only question is: How much *labor* will firms employ to produce goods and services? Moreover, since we are building a *macroeconomic* model, we'll aggregate all the different types of labor—office workers, construction workers, teachers, taxi drivers, waiters, writers, and more—into a single variable, called labor.

THE LABOR MARKET


The classical labor market is illustrated in Figure 1. The number of workers is measured on the horizontal axis, and the real hourly wage rate is measured on the vertical axis. Remember that the *real wage*—which is measured in the dollars of some base year—tells us the amount of goods that workers can buy with an hour's earnings.

Now look at the two curves in the figure. These are supply and demand curves, similar to the supply and demand curves for maple syrup, but there is one key difference: For a *good* such as maple syrup, households are the demanders and firms the suppliers. But for labor, the roles are reversed: Households supply labor, and firms demand it.


The curve labeled L^S is the **labor supply curve** in this market; it tells us how many people will want to work at each wage. The upward slope tells us that the greater the real wage, the greater the number of people who will want to work. Why does the labor supply curve slope upward?

The answer comes from Key Step #2, in which we identify the goals and constraints of decision makers in a market.

Think about your own decision about whether to work—to supply labor. Your goal—at the most general level—is to be as well off as possible. You value both income and leisure time, and in the best of all possible worlds, you'd have a lot of both. However, in the real world, you face a constraint: To earn income, you must go to work and give up leisure. Thus, each of us will want to work only if the income we will earn *at least* compensates us for the leisure that we will give up.

 Characterize the Market

Labor supply curve Indicates how many people will want to work at various wage rates.

 Identify Goals and Constraints

Of course, people differ in the way that they value income and leisure. Thus, for each of us, there is some critical wage rate above which we would decide that we're better off working. Below that wage, we would be better off not working. Thus, in Figure 1,

the labor supply curve slopes upward because—as the wage rate increases—more and more individuals are better off working than not working. Thus, a rise in the wage rate increases the number of people in the economy who want to work—to supply their labor.

The curve labeled L^D is the **labor demand curve**, which shows the number of workers firms will want to hire at any real wage. Why does this curve slope downward?

Identify Goals and Constraints



Labor demand curve Indicates how many workers firms will want to hire at various wage rates.

Once again, we use Key Step #2. In deciding how much labor to hire, a firm's goal is to earn the greatest possible profit—the difference between sales revenue and costs. If a firm's owners could choose, they'd like the firm's revenue to be infinite and its costs to be zero. However, each firm faces a constraint: To earn more revenue, it must produce and sell more output, and this requires it to hire (and pay wages to) more workers. A firm will want to keep hiring additional workers as long as the output produced by those workers adds more to revenue than it adds to costs.

Now think about what happens as the wage rate rises. Some workers that added more to revenue than to cost at the lower wage will now cost more than they add in revenue. Accordingly, the firm will not want to employ these workers at the higher wage.

As the wage rate increases, each firm in the economy will find that—to maximize profit—it should employ fewer workers than before. When all firms behave this way together, a rise in the wage rate will decrease the quantity of labor demanded in the economy. This is why the economy's labor demand curve slopes downward.

Find the Equilibrium



In the classical view, *all markets clear*—including the market for labor. That is, the classical model tells us to apply Key Step #3 in a particular way: The real wage adjusts until the quantities of labor supplied and demanded are equal. In the labor market in Figure 1, the market-clearing wage is \$15 per hour, since that is where the labor supply and labor demand curves intersect. While every worker would prefer to earn \$20 rather than \$15, at \$20 there would be an excess supply of labor equal to the distance AB . With not enough jobs to go around, competition among workers would drive the wage downward. Similarly, firms might prefer to pay their workers \$10 rather than \$15, but at \$10, the excess demand for labor (equal to the distance HJ) would drive the wage upward. When the wage is \$15, however, there is neither an excess demand nor an excess supply of labor, so the wage will neither increase nor decrease. Thus, \$15 is the equilibrium wage in the economy. Reading along the horizontal axis, we see that at this wage, 100 million people will be working.

Notice that, in the figure, labor is fully employed; that is, the number of workers that firms want to hire is equal to the number of people who want jobs. Therefore, everyone who wants a job at the market wage of \$15 should be able to find one. Small amounts of frictional unemployment might exist, since it takes some time for new workers or job switchers to find jobs. And there might be structural unemployment, due to some mismatch between those who want jobs in the market

and the types of jobs available. But there is no *cyclical* unemployment of the type we discussed two chapters ago.

Full employment of the labor force is an important feature of the classical model. As long as we can count on markets (including the labor market) to clear, government action is not needed to ensure full employment; it happens automatically:

In the classical view, the economy achieves full employment on its own.

Automatic full employment may strike you as odd, since it contradicts the cyclical unemployment we sometimes see around us. For example, in the recession of the early 1990s, millions of workers around the country, in all kinds of professions and labor markets, were unable to find jobs for many months. Remember, though, that the classical model takes the long-run view, and over long periods of time, full employment is a fairly accurate description of the U.S. labor market. Cyclical unemployment, by definition, lasts only as long as the current business cycle itself; it is not a permanent, long-run problem.

DETERMINING THE ECONOMY'S OUTPUT

So far, we've focused on the labor market to determine the economy's level of employment. In our example, 100 million people will have jobs. Now we ask: How much output will these 100 million workers produce? The answer depends on two things: (1) the amount of other resources (land and capital) available for labor to use; and (2) the state of *technology*, which determines how much output we can produce with given inputs, as well as the types of inputs available (horse-drawn wagons or trucks; pencil and paper or a laptop computer).

In the classical model, we treat the quantities of land and capital, as well as the state of technology, as fixed during the period we are analyzing. This certainly makes sense in the case of land: Total acreage is pretty much fixed in a country, and there is little that anyone can do to increase it. But what about technology and capital? The state of technology changes with each new invention or discovery. We can already predict, for example, that over the next decade, genetic engineering will lead to completely new drugs and other medical treatments and change the way many existing drugs are produced. And our capital stock changes rapidly as well, since we are constantly producing new capital—more tractors, fiber-optic cable, computers, and factory buildings. How can we treat these as fixed, especially since the classical model is a long-run model?

The answer is: We assume that technology and the capital stock are constant *not* because we believe that they really are, but because doing so helps us understand what happens when they change. We divide our classical analysis of the economy into two questions: (1) What would be the long-run equilibrium of the macroeconomy for a *given* state of technology and a *given* capital stock? and (2) What happens to this equilibrium when capital or technology *changes*? In this chapter, we focus on the first question only. In the next chapter, on economic growth, we'll address the second question. Since we are assuming, for now, a given state of technology, as well as given quantities of land and capital, there is only one variable left that can affect total output: labor. So it's time to explore how changes in total employment affect total production.

The Production Function. The relationship between the quantity of labor employed in the economy and the total quantity of output produced is called the **aggregate production function**:

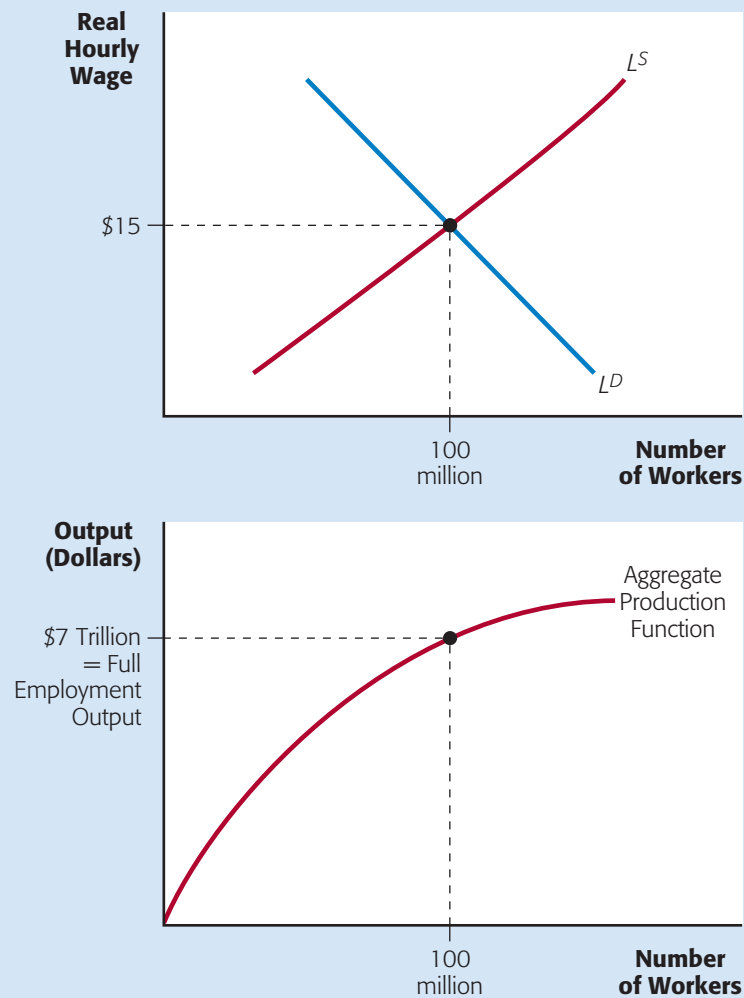
Aggregate production function

The relationship showing how much total output can be produced with different quantities of labor, with land, capital, and technology held constant.

FIGURE 2

In the labor market, the demand and supply curves intersect to determine an employment level of 100 million workers. Given the stock of capital and the current level of technology, the production function shows that those 100 million workers can produce \$7 trillion of real GDP.

OUTPUT DETERMINATION IN THE CLASSICAL MODEL



The aggregate production function shows the total output the economy can produce with different quantities of labor, given constant amounts of land and capital and the current state of technology.

The bottom panel of Figure 2 shows what a nation's aggregate production function might look like. The upward slope tells us that an increase in the number of people working will increase the quantity of output produced. But notice the shape of the production function: It flattens out as we move rightward along it.

The declining slope of the aggregate production function is the result of *diminishing returns to labor*: Output rises when another worker is added, but the rise is smaller and smaller with each successive worker. Why does this happen? For one thing, as we keep adding workers, gains from specialization are harder and harder to come by. Moreover, as we continue to add workers, each one will have less and less capital and land to work with.

Figure 2 also illustrates how the aggregate production function, together with the labor market, determines the economy's total output or real GDP. In our example, the labor market (upper panel) automatically generates full employment of 100 million workers, and the production function (lower panel) tells us that 100 million workers—together with the available capital and land and the current state of technology—can produce \$7 trillion worth of output. Since \$7 trillion is the output produced by a fully employed labor force, it is also the economy's potential output level.

In the classical, long-run view, the economy reaches its potential output automatically.

This last statement is an important conclusion of the classical model and an important characteristic of the economy in the long run: Output tends toward its potential, full-employment level *on its own*, with no need for government to steer the economy toward it. And we have arrived at this conclusion merely by assuming that the labor market clears and observing the relationship between employment and output.

THE ROLE OF SPENDING

Something may be bothering you about the classical view of output determination—a potential problem we have so far carefully avoided: What if business firms are unable to sell all the output produced by a fully employed labor force? Then the economy would not be able to sustain full employment for very long. Business firms will not continue to employ workers who produce output that is not being sold. Thus, if we are asserting that potential output is an equilibrium for the economy, we had better be sure that *total spending* on output is equal to *total production* during the year. But can we be sure of this?

In the classical view, the answer is, absolutely yes! We'll demonstrate this in two stages: first, in a very simple (but very unrealistic) economy, and then, under more realistic conditions.

TOTAL SPENDING IN A VERY SIMPLE ECONOMY

Imagine a world much simpler than our own, a world with just two types of economic units: households and business firms. In this world, households spend all of their income on goods and services. They do not save any of their income, nor do they pay taxes. Such an economy is illustrated in the **circular flow** diagram of Figure 3.

The arrows on the right-hand side show that resources—labor, land, and capital—are supplied by households, and purchased by firms, in *factor markets*. In return, households receive payments—wages, rent, interest, and profit. For example, if you were working part time in a restaurant while attending college, you would be supplying a resource (labor) in a factor market (the market for waiters). In exchange, you would earn a wage. Similarly, the owner of the land on which the restaurant sits is a supplier in a factor market (the market for land) and will receive a payment (rent) in return. The payments received by resource owners are called *factor payments*.

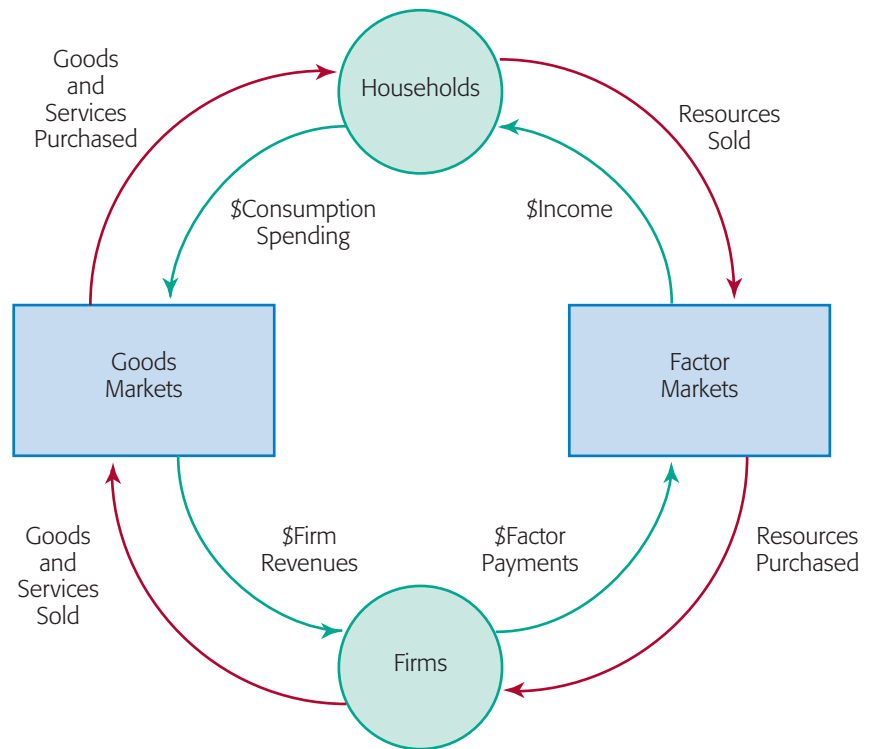
On the left side of the diagram, the outer arrows show the flow of goods and services—food, new clothes, books, movies, and more—that firms supply, and households buy, in various *goods markets*. Of course, households must pay for these goods and services, and their payments provide revenue to firms—as shown by the inner arrows.

Circular flow A diagram that shows how goods, resources, and dollar payments flow between households and firms.

FIGURE 3

THE CIRCULAR FLOW

The outer loop of the diagram shows the flows of goods and resources. Households supply resources to firms, which use them to produce goods. The inner loop shows money flows. Firms' factor payments become income to households. Households use the income to purchase goods from firms.



Now comes an important insight. As you learned two chapters ago, the total output of firms is equal to the total income of households. For example, if the economy is producing \$7 trillion worth of output, then it also creates \$7 trillion in household income. And in this simple economy—in which households spend all of their income—spending would equal \$7 trillion as well.

In general,

In a simple economy with just households and firms, in which households spend all of their income, total spending must be equal to total output.

Say's law The idea that total spending will be sufficient to purchase the total output produced.

This simple proposition is called **Say's law**, after the classical economist Jean Baptiste Say (1767–1832), who popularized the idea. Say noted that each time a good or service is produced, an equal amount of income is created. This income is spent—it comes back to the business sector to purchase its goods and services. In Say's own words:

A product is no sooner created than it, from that instant, affords a market for other products to the full extent of its own value. . . . Thus, the mere circumstance of the creation of one product immediately opens a vent for other products.²

² J. B. Say, *A Treatise on Political Economy*, 4th ed. (London: Longman, 1821), Vol. I, p. 167.

For example, each time a shirt manufacturer produces a \$25 shirt, it creates \$25 in factor payments to households. (Forgot why? Go back two chapters.) But \$25 in factor payments will lead to \$25 in total spending—just enough to buy the very shirt produced. Of course, those households who receive the \$25 in factor payments will not necessarily buy a shirt with it: The shirt manufacturer must still worry about selling its own output. But in the aggregate, we needn't worry about there being sufficient demand for the *total* output produced. Business firms—by producing output—also create a demand for goods and services equal to the value of that output. Or, to put it most simply, *supply creates its own demand*:

Say's law states that by producing goods and services, firms create a total demand for goods and services equal to what they have produced.

Say's law is crucial to the classical view of the economy. Why? Remember that market clearing in resource markets assures us that firms will produce potential output. Say's law then assures us that, in the aggregate, firms will be able to *sell* this output, so that full employment can be sustained.

TOTAL SPENDING IN A MORE REALISTIC ECONOMY

The real world is more complicated than the imaginary one we've just considered. In the real world,

1. Households don't spend *all* their income. Rather, some of their income is saved or goes to *pay taxes*.
2. Households are not the only spenders in the economy. Rather, businesses and the government buy some of the final goods and services we produce.
3. In addition to markets for goods and resources, there is also a *loanable funds* market where household saving is made available to borrowers in the business or government sectors.

All of these details complicate our picture of the economy. Can we have confidence that Say's law will hold under these more realistic conditions?

As you are about to see, yes, we can.

Let's consider the economy of *Classica*—a fictional economy that behaves according to the classical model. *Classica*'s economy in 2002 is described in Table 1. Notice that total output and total income are both equal to \$7 trillion (\$7,000 billion), which is assumed to be the potential output level.

Two entries in the table require a bit of explaining. First, **net taxes** are total tax revenue minus government transfer payments such as unemployment insurance, welfare payments, and Social Security benefits. As discussed two chapters ago, these transfer payments are the part of tax revenue that the government takes from one set of households and gives right back to another set of households. Since transfer

Net taxes Government tax revenues minus transfer payments.

TABLE 1

FLOWS IN THE ECONOMY OF CLASSICA, 2002

Total Output	\$7 trillion
Total Income	\$7 trillion
Consumption Spending (C)	\$4 trillion
Investment Spending (I^P)	\$1 trillion
Government Spending (G)	\$2 trillion
Net Tax Revenue (T)	\$1.25 trillion
Household Saving (S)	\$1.75 trillion

payments stay within the household sector as a whole, we can treat them as if they were never paid to the government at all. Net taxes, then, are the funds that flow from the household sector as a whole to the government in any given year. Letting T represent net taxes, we have

$$T = \text{Total taxes} - \text{Transfer payments.}$$

(Household) saving The portion of after-tax income that households do not spend on consumption goods.

Second, **household saving** (often, just **saving**) is the part of the household sector's income that is left after deducting what it pays to the government in taxes and what it spends on consumption. Using the symbol S for household saving, Y for total income, and C for consumption spending, we can write

$$S = Y - T - C.$$

LEAKAGES AND INJECTIONS

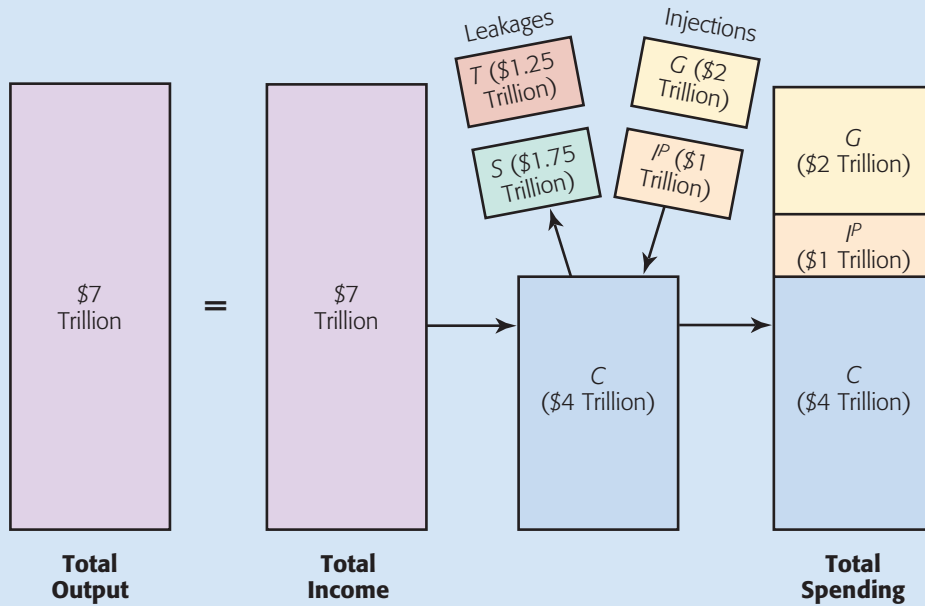
As you can see in Table 1, Classica's households earn \$7 trillion in income during the year, but they spend only \$4 trillion. That leaves \$3 trillion left over from their income after we deduct their consumption spending. Part of this remaining \$3 trillion goes to pay net taxes (\$1.25 trillion), and whatever is left is, by definition, saved (\$1.75 trillion).

Leakages Income earned, but not spent, by households during a given year.

Saving and net taxes are called **leakages** out of the income–spending stream—income that households earn but do not spend. Leakages are important because they seem to threaten Say's law—the classical idea that total spending will always equal output. To see why, look at the rectangles in Figure 4. Total output (the first

FIGURE 4

LEAKAGES AND INJECTIONS



By definition, total output equals total income. Leakages—net taxes and saving—reduce consumption spending below total income. Injections—government purchases plus investment spending—contribute to total spending. When leakages equal injections, total spending equals total output.

rectangle) is, by definition, always equal in value to total income (the second rectangle). As we've seen in Figure 3, if households spent all of this income, then consumption spending would equal total output. But leakages reduce consumption spending below total income, as you can see in the third, lower rectangle. In *Classica*, total leakages = \$1.75 trillion + \$1.25 trillion = \$3 trillion, and this must be subtracted from income of \$7 trillion to get consumption spending of \$4 trillion. Thus, if consumption spending were the only spending in the economy, business firms would be unable to sell their entire potential output of \$7 trillion.

Fortunately, in addition to leakages, there are **injections**—spending from sources *other* than households. Injections boost total spending, and enable firms to produce and sell a level of output greater than just consumption spending.

There are two types of injections in the economy. First is the government's purchases of goods and services. When government agencies—federal, state, or local—buy aircraft, cleaning supplies, cellular phones, or computers, they are buying a part of the economy's output.

The other injection is business firms' investment spending on new capital. We call this **planned investment spending** (or sometimes, just *investment spending*), and represent it with the symbol I^p . Recall, from two chapters ago, that actual investment (I) consists not just of planned investment in new capital, but also the unplanned changes in inventories. While *some* of the change in inventories in any year might be desired and planned by firms, we'll assume that most of the change in inventories comes as a surprise. More specifically, an increase in inventories is usually an unwelcome surprise, while a decrease in inventories is a pleasant surprise. For example, if Calvin Klein produces \$40 million in clothing during the year, but actually ships and sells only \$35 million, the \$5 million in unsold output will be an unplanned increase in inventories—a surprise that will not make Calvin Klein's owners very happy. But if the company sells \$45 million one year—more than it produced—it must have sold some goods out of the inventories it had previously built up. This will generally be good news for the firm.

Why, when we define injections, do we only count *planned* investment spending (I^p), rather than actual investment (I)? Why do we exclude the change in inventories? Because changes in inventories, being unplanned surprises, are basically one-time events. They do not represent a sustainable source of spending for the economy, and therefore do not help us determine the economy's equilibrium.

Injections are the opposite of leakages: Whereas leakages reduce total spending in the economy, injections increase it. In Figure 4, the last rectangle shows how total injections—investment and government purchases—are added to consumption to obtain total spending. As you can see, total spending is the sum of consumption, planned investment, and government purchases.³ In *Classica*, using Table 1, we find that consumption spending (C) is \$4 trillion, investment spending (I) is \$1 trillion, and government purchases (G) are \$2 trillion, giving us total spending of \$7 trillion.

This may strike you as suspiciously convenient: Total spending is exactly equal to total output, just as we would like it to be if we want firms to continue producing their potential output level of \$7 trillion. And, of course, we have cooked the numbers to make them come out that way. But do we have any reason to *expect* this result in an economy over the long run? Actually, we do.

Injections Spending from sources other than households.

Planned investment spending Business purchases of plant and equipment.

³ There is one more source of spending in the economy that we are not considering here: spending by foreigners, on *Classica*'s exports. But as long as exports (an injection) and imports (a leakage) are equal, none of the conclusions that follow are affected in important ways. We'll focus more directly on exports and imports in our short-run macro model, which begins two chapters after this one.

Take another look at the rectangles in Figure 4. Notice that in going from total output to total spending, leakages are subtracted and injections are added. Clearly, total output and total spending will be equal only when leakages and injections are equal as well:

Total spending will equal total output if and only if total leakages in the economy are equal to total injections—that is, only if the sum of saving and net taxes is equal to the sum of investment spending and government purchases.

And here is a surprising result: This condition will automatically be satisfied. To see why, we must first take a detour through another important market. Then we'll come back to the all-important equality between leakages and injections.

THE LOANABLE FUNDS MARKET

Characterize the Market



Loanable funds market Arrangements through which households make their saving available to borrowers.

The **loanable funds market** is where households make their saving available to those who need additional funds. When you save—that is, when you have income left over after paying taxes and buying consumption goods—you can put your surplus funds in a bank, buy a bond or a share of stock, or use the funds to buy a variety of other assets. In each of these cases, you would be a supplier in the loanable funds market.

Households supply funds because they receive a reward for doing so. But the reward comes in different forms. When the suppliers *lend* out funds, the reward is *interest payments*. When the funds are provided through the stock market, the suppliers become part owners of the firm and their payment is called *dividends*. To keep our discussion simple, we'll assume that all funds transferred are *loaned* and that the payment is simply *interest*.

On the other side of the market are those who want to obtain funds—demanders in this market. Business firms are important demanders of funds. When Avis wants to add cars to its automobile rental fleet, when McDonald's wants to build a new beef-processing plant, or when the local dry cleaner wants to buy new dry cleaning machines, it will likely raise the funds in the loanable funds market. It may take out a bank loan, sell bonds, or sell new shares of stock. In each of these cases, a firm's planned investment spending would be equal to the funds it obtains from the loanable funds market.

Aside from households and business firms, the other major player in the loanable funds market is the government. Government participates in the market whenever it runs a budget deficit or a budget surplus.

*When government purchases of goods and services (G) are greater than net taxes (T), the government runs a **budget deficit** equal to $G - T$. When government purchases of goods and services (G) are less than net taxes (T), the government runs a **budget surplus** equal to $T - G$.*

Budget deficit The excess of government purchases over net taxes.

Budget of surplus The excess of net taxes over government purchases.

In our example in Table 1, Classica's government is running a budget deficit: Government purchases are \$2 trillion, while net taxes are \$1.25 trillion, giving us a deficit of \$2 trillion $-$ \$1.25 trillion = \$0.75 trillion. This deficit is financed by borrowing in the loanable funds market. In any year, the government's demand for funds is equal to its deficit.

But surpluses, too, involve the government in the loanable funds market. When the government runs a surplus, it pays back debts that it incurred while running deficits in previous years. For example, the federal government's total unpaid debt is called the **national debt**. When the federal government runs a surplus, it pays back

National debt The total amount of government debt outstanding.

part of the national debt, buying back government bonds that it issued in previous years when it ran deficits. In this sense, it becomes a *supplier* of loanable funds, because it is putting funds into the market, where they can be borrowed by others.

State and local governments, like the federal government, can run deficits and surpluses, requiring them to participate in the loanable funds market. In our classical model, we aggregate all of these levels of government together, and refer only to the government. When the government runs a budget deficit, it demands loanable funds equal to its deficit. When the government runs a budget surplus, it supplies loanable funds equal to its surplus.

We can summarize our view of the loanable funds market so far with these two points:

- The supply of funds is the sum of household saving and the government's budget surplus, if any.
- The demand for funds is the sum of the business sector's planned investment spending and the government sector's budget deficit, if any.

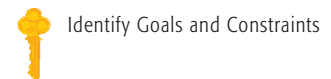
In Classica, the government is running a deficit, not a surplus, so for now, we'll analyze the loanable funds market with a budget deficit only. Then, in the "Using the Theory" section, we'll take up the case of a budget surplus.

THE SUPPLY OF FUNDS CURVE

When the government is running a budget deficit rather than a surplus, households are the only suppliers of funds. Since interest is the reward for saving and supplying funds to the financial market, a rise in the interest rate *increases* the quantity of funds supplied (household saving), while a drop in the interest rate decreases it. This relationship is illustrated by Classica's upward-sloping **supply of funds curve** in Figure 5. If the interest rate is 3 percent, households save \$1.5 trillion, and if the interest rate rises to 5 percent, people save more and the quantity of funds supplied rises to \$1.75 trillion.



When the Stop & Shop Corporation opens a new supermarket, it very likely obtains the funds from the loanable funds market, by issuing bonds, taking out bank loans, or issuing new shares of stock.



Identify Goals and Constraints

Supply of funds curve Indicates the level of household saving at various interest rates.

THE SUPPLY OF FUNDS

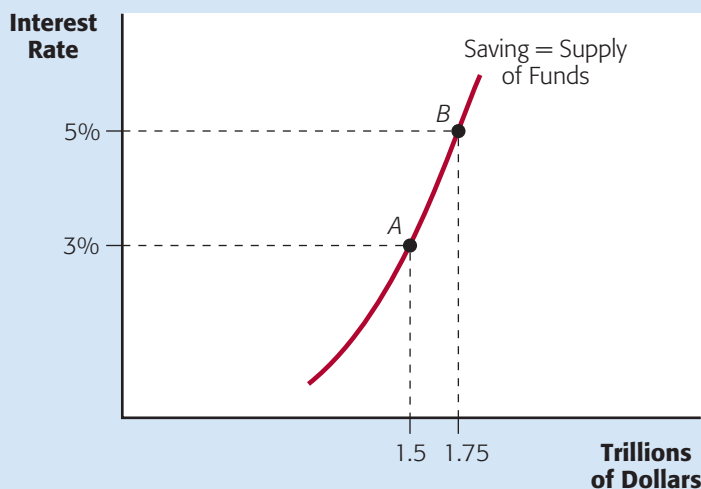


FIGURE 5

Interest is the reward for saving. The upward-sloping supply of funds curve shows that at higher interest rates, households consume less, save more, and supply more funds to the loanable funds market.

The quantity of funds supplied to the financial market depends positively on the interest rate. This is why the saving, or supply of funds, curve slopes upward.

Of course, other things can affect saving besides the interest rate—tax rates, expectations about the future, and the general willingness of households to postpone consumption, to name a few. In drawing the supply of funds curve, we assume each of these variables is constant. In the next chapter, we'll explore what happens when some of these variables change.

Identify Goals and Constraints



THE DEMAND FOR FUNDS CURVE

Like saving, investment also depends on the interest rate. This is because businesses buy plant and equipment when the expected benefits of doing so exceed the costs. Since businesses obtain the funds for their investment spending from the loanable funds market, a key cost of any investment project is the interest rate that must be paid on borrowed funds. As the interest rate rises and investment costs increase, fewer projects will look attractive, and investment spending will decline. This is the logic of the downward-sloping **investment demand curve** in Figure 6. At a 5 percent interest rate, firms would borrow \$1 trillion and spend it on capital equipment; at an interest rate of 3 percent, business borrowing and investment spending would rise to \$1.5 trillion.

Investment demand curve Indicates the level of investment spending firms plan at various interest rates.

When the interest rate falls, investment spending and the business borrowing needed to finance it rise. The investment demand curve slopes downward.

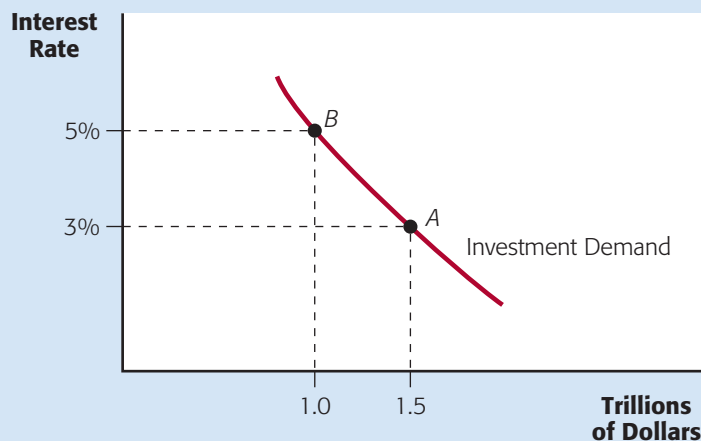
What about the government's demand for funds? Will it, too, be influenced by the interest rate? Probably not very much. Government seems to be cushioned from the cost–benefit considerations that haunt business decisions. Any company president who ignored interest rates in deciding how much to borrow would be quickly out of a job. U.S. presidents and legislators have often done so with little political cost.

For this reason, when government is running a budget deficit, our classical model treats government borrowing as independent of the interest rate: No matter

FIGURE 6

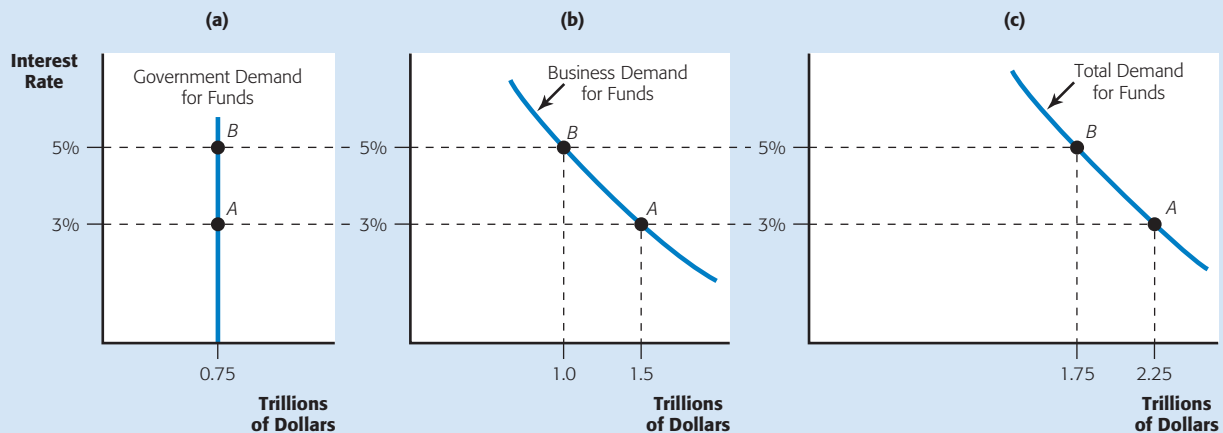
Businesses borrow in order to finance new investment, and the interest rate measures the cost of borrowing. The downward-sloping investment demand curve shows that more new projects will be financially attractive at low interest rates than at high rates.

INVESTMENT SPENDING



THE DEMAND FOR FUNDS

FIGURE 7



In panel (a), the government's demand for funds—to finance the budget deficit—is independent of the interest rate. Businesses' demand for funds—for investment—is inversely related to the interest rate in panel (b). The total demand for funds in panel (c) is the horizontal sum of government and business demand. At lower interest rates, more funds are demanded than at higher rates.

what the interest rate, the government sector's deficit—and its borrowing—remain constant. This is why we have graphed the **government's demand for funds curve** as a vertical line in panel (a) of Figure 7.

The government sector's deficit and, therefore, its demand for funds are independent of the interest rate.

In the figure, the government deficit—and hence the government's demand for funds—is equal to \$0.75 trillion at any interest rate.

In Figure 7, the **total demand for funds curve** is found by horizontally summing the government demand curve (panel (a)) and the business demand curve (panel (b)). For example, if the interest rate is 5 percent, firms demand \$1 trillion in funds, and the government demands \$0.75 trillion, so that the total quantity of loanable funds demanded is \$1.75 trillion. A drop in the interest rate—to 3 percent—increases business borrowing to \$1.5 trillion, while the government's borrowing remains at \$0.75 trillion, so the total quantity of funds demanded rises to \$2.25 trillion.

As the interest rate decreases, the quantity of funds demanded by business firms increases, while the quantity demanded by the government remains unchanged. Therefore, the total quantity of funds demanded rises.

Government demand for funds curve Indicates the amount of government borrowing at various interest rates.

Total demand for funds curve Indicates the total amount of borrowing at various interest rates.

EQUILIBRIUM IN THE LOANABLE FUNDS MARKET

In the classical view, the loanable funds market—like all other markets—is assumed to clear: The interest rate will rise or fall until the quantities of funds supplied and demanded are equal. Figure 8 illustrates the financial market of Clastica, our fictional economy. Equilibrium occurs at point *E*, with an interest rate of

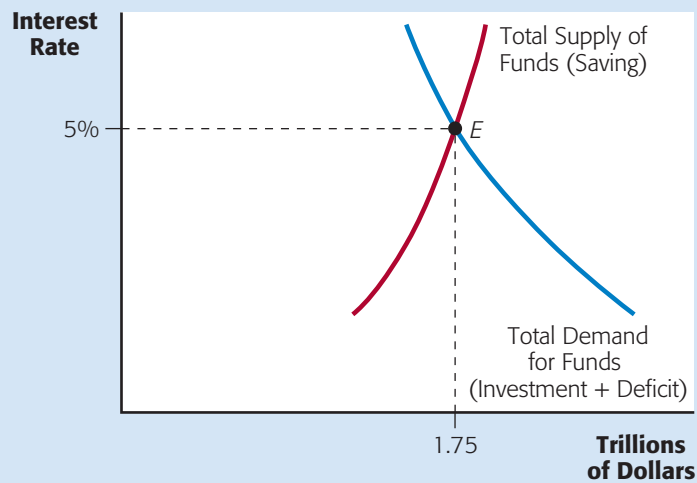


Find the Equilibrium

FIGURE 8

Suppliers and demanders of funds interact to determine the interest rate in the loanable funds market. At an interest rate of 5%, quantity supplied and quantity demanded are both equal to \$1.75 trillion.

LOANABLE FUNDS MARKET EQUILIBRIUM



5 percent and total saving equal to \$1.75 trillion. Of the total saved, \$1 trillion goes to business firms for capital purchases, and \$0.75 trillion goes to the government to cover its deficit.

So far, our exploration of the loanable funds market has shown us how three important variables in the economy are determined: the interest rate, the level of saving, and the level of investment. But it really tells us more. Remember the question that sent us on this detour into the loanable funds market in the first place: Can we be sure that all of the output produced at full employment will be purchased? We now have the tools to answer this question.

THE LOANABLE FUNDS MARKET AND SAY'S LAW

In Figure 4 (flip back 6 pages), you saw that total spending will equal total output if and only if *total leakages* in the economy (saving plus net taxes) are equal to *total injections* (planned investment plus government purchases). Now we can see how this requirement is satisfied automatically. Because the loanable funds market clears, we know that the interest rate—the price in this market—will rise or fall until the quantities of funds supplied (saving) and funds demanded (investment plus the deficit) are equal. Letting S stand for saving, I^p for investment, and $G - T$ for the deficit, we can state that the interest rate will adjust until

$$\underbrace{S}_{\text{Quantity of funds supplied}} = \underbrace{I^p + G - T}_{\text{Quantity of funds demand}}$$

Rearranging this equation by moving T to the left side, we find that, when the loanable funds market clears,

$$\underbrace{S + T}_{\text{Leakages}} = \underbrace{I^p + G}_{\text{Injections}}$$

In other words, market clearing in the loanable funds market *assures us* that total leakages in the economy will equal total injections, which in turn *assures us* that there will be enough spending in the economy to purchase whatever output level is produced. Thus,

as long as the loanable funds market clears, Say's law holds even in a more realistic economy with saving, taxes, investment, and a government deficit.

To see the logic of this conclusion another way, go back again to Figure 4. There, we saw that households spend only part of their income; the rest is either saved or paid as taxes. Now, taxes and saving do not just disappear from the economy: Tax payments go to the government, which spends them. Saving goes to the loanable funds market, where it will be passed along to the government or to business firms. In each case, the funds that households do not spend are simply passed along to another sector of the economy that *does* spend them. As long as the loanable funds market is working properly, income never escapes from the economy. Instead, every dollar in leakages is recycled back into the spending stream in the form of injections.

Figure 9 shows how leakages are transformed into injections. The dollar amounts are for the economy of Classica. In the figure, you can see that by producing \$7 trillion in output, firms create \$7 trillion in payments to inputs. Of this total,

AN EXPANDED CIRCULAR FLOW

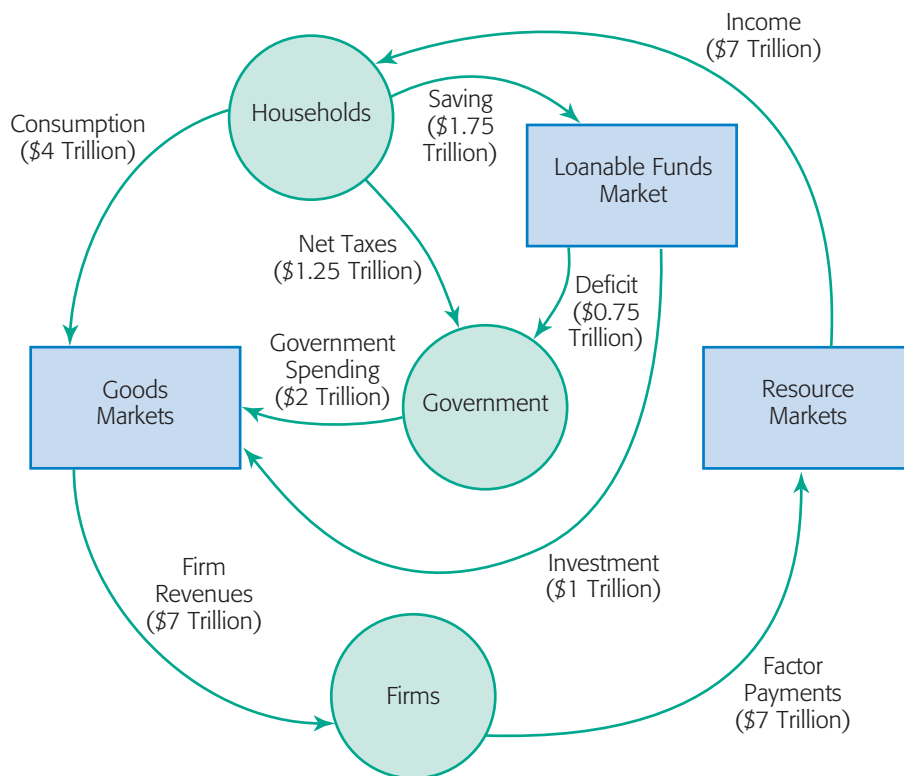


FIGURE 9

Saving is transformed into business and government spending in the loanable funds market. The interest rate adjusts to guarantee that saving plus net taxes will equal government purchases plus investment. As a result, total income will equal total spending. (The dollar numbers—which come from Table 2—are for our hypothetical economy, Classica.)

households spend \$4 trillion. The rest goes to pay net taxes (\$1.25 trillion) or is saved (\$1.75 trillion). But taxes and saving do not escape from the economy: The tax payments of \$1.25 trillion and part of the saving (\$0.75 trillion) are spent by the government, whose purchases are \$2 trillion. The rest of the saving (\$1 trillion) is spent by business firms on new capital. In the end, the entire \$7 trillion in output is purchased, just as Say's law asserts.

Say's law is a powerful concept. But be careful not to overinterpret it. Say's law shows that the *total* value of spending in the economy will equal the *total* value of output, which rules out a general overproduction or underproduction of goods in the economy. It does not promise us that each firm in the economy will be able to sell all of its output. It is perfectly consistent with Say's law that there be excess supplies in some markets, as long as they are balanced by excess demands in other markets.

But lest you begin to think that the classical economy might be a chaotic mess, with excess supplies and demands in lots of markets—don't forget about the *market-clearing* assumption. In each market, prices adjust until supplies and demands are equal. For this reason, the classical, long-run view rules out over- or underproduction in individual markets, as well as the generalized overproduction ruled out by Say's law.

THE CLASSICAL MODEL: A SUMMARY

You've just completed a first tour of the classical model, our framework for understanding the economy in the long run. Before we begin to use this model, this is a good time to go back and review what we've done.

We began with a critical assumption: All markets clear. We then used the first three Key Steps of our four-step procedure to organize our thinking about the economy. First, we focused on an important market—the labor market—and identified the buyers and sellers in that market. We identified the goals and constraints of these buyers and sellers. And then we found the equilibrium in that market by applying the market-clearing assumption.

We went through a similar process with the loanable funds market, identifying the suppliers and demanders, examining how each would be affected by changes in the interest rate, and finding the equilibrium in that market as well. Then, we saw how market clearing in the loanable funds market assures us that total spending will be just sufficient to purchase the potential output level.

In our excursion through the classical model, we've come to some important conclusions. First, we've seen that *the economy will achieve and sustain potential output on its own*. We have also reached an interesting conclusion about the role of spending in the economy: *We need never worry about there being too little or too much spending; Say's law assures us that total spending is always just right to purchase the economy's total output*.

All of this tells us that the government needn't worry much about the economy's level of production: It reaches the right level on its own. But suppose the government wanted to stimulate the economy, and raise the level of economic activity in order to increase employment and output. Could the government accomplish this by engineering an *increase* in total spending? We'll answer that question in our "Using the Theory" section.

FISCAL POLICY IN THE CLASSICAL MODEL

Can the government raise output by raising spending in the economy? It seems like it could, and two ideas come readily to mind. First, the government could simply spend more itself—purchasing more goods, like tanks and police cars, and more services, like those provided by high school teachers and judges. Alternatively, the government could cut taxes so that households would keep more of their income, causing them to spend more on food, clothing, furniture, travel, movies, new cars, and so on. When the government either increases its spending or reduces taxes in order to influence the level of economic activity, it is engaging in *fiscal policy*:

Fiscal policy is a change in government purchases or in net taxes designed to change total spending in the economy and thereby influence the levels of employment and output.

A fiscal policy of increasing government purchases or decreasing net taxes should cause spending to rise, and business firms—able to sell more—would surely hire more workers and produce more goods and services. Right?

In the classical model, this is dead wrong. Fiscal policy is completely ineffective. It cannot change total output or employment in the economy, period. It cannot even change total spending. Moreover, fiscal policy is *unnecessary*, since the economy achieves and sustains full employment on its own.


In the classical view, fiscal policy is both ineffective and unnecessary.

Here, we'll demonstrate this conclusion for the case of an increase in government spending. In a challenge question at the end of this chapter, you are invited to demonstrate the same conclusion for the case of a tax cut.

Let's see what would happen if the government of Classica attempted to increase employment and output by increasing its own purchases. More specifically, suppose its purchases rise from the current \$2 trillion to \$2.5 trillion annually, while net taxes remain unchanged. What will happen?

To answer this, we must first answer another question: Where will Classica's government get the additional \$0.5 trillion it spends? If net taxes are unchanged (as we are assuming), then the government must dip into the loanable funds market to borrow the additional funds. Figure 10 illustrates the effects. Initially, with government purchases equal to \$2 trillion, the demand for funds curve is D_1 , and equilibrium occurs at point A with the interest rate equal to 5 percent. If government purchases increase by \$0.5 trillion, with no change in taxes, the budget deficit increases by \$0.5 trillion, and so does the government's demand for funds. The demand for funds curve shifts rightward by \$0.5 trillion to D_2 , since total borrowing will now be \$0.5 trillion greater at *any* interest rate. After the shift, there would be an excess demand for funds at the original interest rate of 5 percent. The total quantity of funds demanded would be \$2.25 trillion (point H), while the quantity supplied would continue to be \$1.75 trillion (point A). Thus, the excess demand for funds would be equal to the distance AH in the figure, or \$0.5 trillion. This excess demand drives up the interest rate to 7 percent. As the interest rate rises, two things happen.

First, a higher interest rate chokes off some investment spending, as business firms decide that certain investment projects no longer make sense. For example,

 What Happens When Things Change?

Using the
THEORY

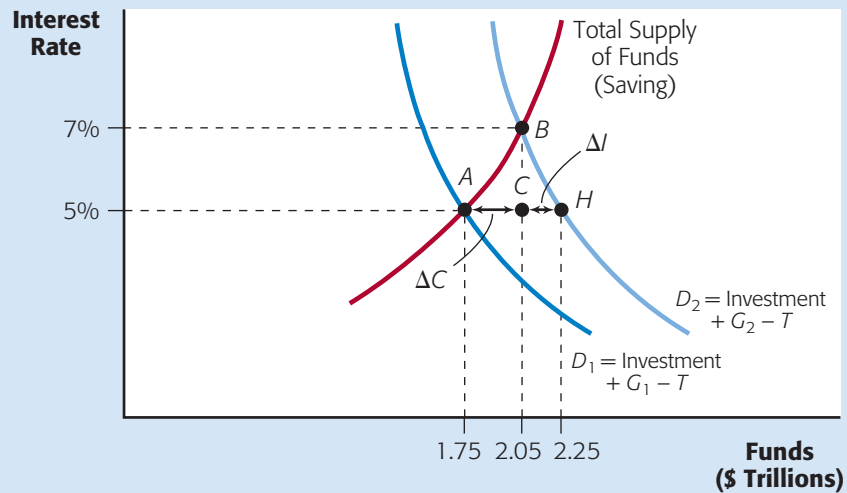


Fiscal policy A change in government purchases or net taxes designed to change total spending and total output.

FIGURE 10

Beginning from equilibrium at point A , an increase in the budget deficit created to finance additional government purchases shifts the demand for funds curve from D_1 to D_2 . At point H , the quantity of funds demanded exceeds the quantity supplied, so the interest rate begins to rise. As it rises, households are led to save more, and business firms invest less. In the new equilibrium at point B , both consumption and investment spending have been completely crowded out by the increased government spending.

CROWDING OUT WITH AN INITIAL BUDGET DEFICIT



the local dry cleaner might wish to borrow funds for that new machine at an interest rate of 5 percent, but not at 7 percent. In the figure, as we move along the new demand-for-funds curve D_2 , from point H to point B , investment declines by \$0.2 trillion (from \$2.25 trillion to \$2.05 trillion). (Question: How do we know that only business borrowing, and not also government borrowing, adjusts as we move from point H to point B ?) Thus, one consequence of the rise in government purchases is a *decrease in investment spending*.

But that's not all: The rise in the interest rate also causes saving to increase. Of course, when people save more of their incomes, they spend less, so another consequence of the rise in government purchases is a *decrease in consumption spending*. In the figure, we move from point A to point B along the saving curve, as saving increases (and consumption decreases) by \$0.3 trillion—rising from \$1.75 trillion to \$2.05 trillion.

Let's recap: As a result of the increase in government purchases, both investment spending and consumption spending decline. The government's purchases have *crowded out* the spending of households (C) and businesses (I).

Crowding out A decline in one sector's spending caused by an increase in some other sector's spending.

Crowding out is a decline in one sector's spending caused by an increase in some other sector's spending.

Complete crowding out A dollar-for-dollar decline in one sector's spending caused by an increase in some other sector's spending.

But we are not quite finished. If we sum the drop in C and the drop in I , we find that total private sector spending has fallen by \$0.3 trillion + \$0.2 trillion = \$0.5 trillion. That is, the drop in private sector spending is precisely equal to the rise in public sector spending, G . Not only is there crowding out, there is **complete crowding out**—each dollar of government purchases causes private sector spending to decline by a full dollar. The net effect is that total spending ($C + I + G$) does not change at all!

In the classical model, a rise in government purchases completely crowds out private sector spending, so total spending remains unchanged.

A closer look at Figure 10 shows that this conclusion always holds, regardless of the particular numbers used or the shapes of the curves. When G increases, the demand-for-funds curve shifts rightward by the same amount that G rises, or the distance from point A to point H . Then the interest rate rises, causing two things to happen. First, the movement along the supply of funds curve, from point A to point B , shows that saving rises (consumption falls) by the distance AC . Second, the movement along the demand for funds curve, from point H to point B , shows that investment spending falls by the amount CH . The impact can be summarized as follows:

- Increase in $G = AH$
- Decrease in $C = AC$
- Decrease in $I = CH$

And since $AC + CH = AH$, we know that the combined decrease in C and I is precisely equal to the increase in G .

Because there is complete crowding out in the classical model, a rise in government purchases cannot change total spending. And the logic behind this result is straightforward. Each additional dollar the government spends is obtained from the financial market, where it would have been spent by someone else if the government hadn't borrowed it. How do we know this? Because the financial market funnels every dollar of household saving—no more and no less—to either the government or business firms. If the government borrows more, it just removes funds that would have been spent by businesses (the drop in I) or by consumers (the drop in C).

An increase in government purchases has no impact on total spending and no impact on total output or total employment.

Of course, the opposite sequence of events would happen if government purchases decreased: The drop in G would shrink the deficit. The interest rate would decline, and private sector spending (C and I) would rise by the same amount that government purchases had fallen. (See if you can draw the graphs to prove this to yourself.) Once again, total spending and total output would remain unchanged.

FISCAL POLICY WITH A BUDGET SURPLUS

Fiscal policy has the same macroeconomic effects whether the government is initially running a budget deficit or a budget surplus. However, in the case of a budget surplus, the graphical analysis is a bit different.

Figure 11 shows equilibrium in the loanable funds market with a budget surplus. Remember that, with a budget surplus, the government *supplies* loanable funds, rather than demands them. Therefore, the total demand for loanable funds in Figure 11 is equal to business investment spending alone. The supply of loanable funds, however, now consists of household saving *plus* the budget surplus. That is, because of the surplus, the total supply of funds curve S_1 lies further to the right than it otherwise would by the amount of the surplus.

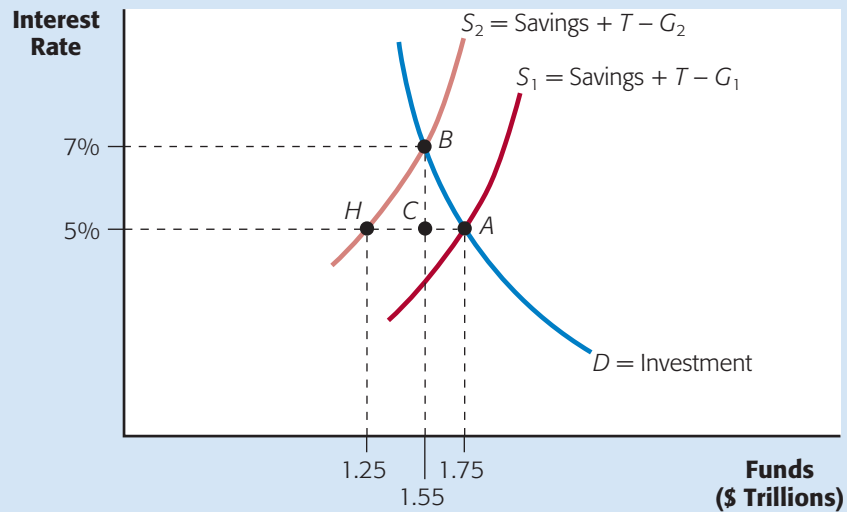
In the initial equilibrium at point A , the interest rate is 5 percent, and the total quantity of funds supplied and demanded are equal, at \$1.75 trillion. If government spending rises by \$0.5 trillion, with no change in taxes, the budget surplus will shrink by \$0.5 trillion, shifting the supply of funds curve leftward by that amount to S_2 . In new equilibrium at point B , the interest rate is higher (7 percent) and the quantity of funds supplied and demanded is lower (\$1.55 trillion).

But that's not all: The rise in the interest rate also causes saving to increase, and consumption spending to decrease. This is represented by the movement from point H

FIGURE 11

Beginning from equilibrium at point *A*, an increase in government purchases causes a decrease in the budget surplus, shifting the supply of funds curve from S_1 to S_2 . At point *H*, the quantity of funds demanded exceeds the quantity supplied, so the interest rate begins to rise. As it rises, households are led to save more, and business firms invest less. In the new equilibrium at point *B*, both consumption and investment spending have been completely crowded out by the increased government spending.

CROWDING OUT WITH AN INITIAL BUDGET SURPLUS



to point *B* along the new supply of funds curve, which causes saving to rise (consumption to decrease) by \$0.3 trillion, or the distance *HC*. The rise in the interest rate also causes investment spending to decrease along the total demand for funds curve, from point *A* to point *B*. Investment spending falls by \$0.2 trillion, or the distance *AC*. Once again, we see that the rise in government spending has completely crowded out consumption and investment spending: A \$0.5 trillion rise in government spending has caused consumption and investment spending to decrease by a total of \$0.5 trillion. Total spending remains unchanged, and the fiscal policy is completely ineffective.

Our exploration of fiscal policy shows us that, in the long run, government efforts to change total output by changing government spending or taxes are not only unnecessary, but also ineffective. What, then, *should* a government do to help manage the macroeconomy in the long run? And what *can* it do? These are questions we explore in the next chapter, where we use the classical model to analyze how the economy grows and what governments can do to help or hinder that growth.

SUMMARY

The classical model is an attempt to explain the behavior of the economy over long time periods. Its most critical assumption is that markets clear—that prices adjust in every market to equate quantities demanded and supplied. The labor market is perhaps the most important part of the classical model. When the labor market clears, we have full employment, and the economy produces the potential level of output.

Another important concept is the production function. It shows the total output the economy can produce with different quantities of labor and for given amounts of land and capital and a given state of technology. When the labor market is at full employment, the production function can be used to determine the economy's potential level of output.

According to Say's law, total spending in the economy will always be just sufficient to purchase the amount of total output produced. By producing and selling goods and services, firms create a total demand equal to what they have produced. If households do not spend their entire incomes, the excess is channeled—as saving—into the loanable funds market, where it is borrowed and spent by businesses and government.

In the loanable funds market, the quantity of funds supplied equals household saving, which depends positively on the interest rate, plus the government budget surplus, if there is one. The quantity of funds demanded equals business investment, which depends negatively on the interest rate, and any government budget deficit, if there is one. The interest

rate adjusts so that the quantity of funds supplied always equals the quantity demanded. Equivalently, it adjusts so that saving (S) equals the sum of investment spending (I) and the government budget deficit ($G - T$).

Fiscal policy cannot affect total output in the classical model. An increase in government purchases results in complete crowding out of investment and consumption spending, leaving total spending and total output unchanged.

KEY TERMS

classical model	Say's law	loanable funds market	government demand for funds curve
market clearing	net taxes	budget deficit	total demand for funds curve
labor supply curve	(household) saving	budget surplus	fiscal policy
labor demand curve	leakages	national debt	crowding out
aggregate production function	injections	supply of funds curve	complete crowding out
circular flow	planned investment spending	investment demand curve	

REVIEW QUESTIONS

- Discuss the critical assumption on which the classical model is based. How does it relate to the length of time over which we are analyzing the economy?
- Describe how, in the classical model, the economy reaches full employment automatically. Is this a "realistic" depiction of how the economy behaves?
- Why does the classical model treat technology and the capital stock as constant?
- Explain why the slope of the aggregate production function diminishes as more labor is employed.
- "According to Say's law, all markets always clear." True or false? Explain.
- What is the difference between net taxes and total tax revenue? Why is the distinction important?
- Who are the two major groups on the demand side of the loanable funds market? Why does each seek funds there? What is the "price" of these funds?
- What is the source of funds supplied to the loanable funds market? Explain why the supply of funds curve slopes upward, and why the curve depicting business demand for funds slopes downward.
- How will the slope of the demand for funds curve be affected if the government runs a budget deficit? Why?
- Why does Say's law hold even after household saving and taxes are taken into account?
- Explain the implications of the classical model for government economic policy. What are the two consequences of an increase in government spending that the model predicts?
- A senator asserts that deficit spending reduces business investment dollar for dollar—every dollar the government borrows means that business investment must fall by a dollar. Is he correct? Why or why not?

PROBLEMS AND EXERCISES

- Use a diagram similar to Figure 2 to illustrate the effect—on aggregate output and the real hourly wage—of (a) an increase in labor demand, and (b) an increase in labor supply.
- The following data give a complete picture of the household, business, and government sectors for 2001 in the small nation of Sylvania. (All dollar figures are in billions.)

Consumption spending	\$50
Capital stock (end of 2000)	\$100
Capital stock (end of 2001)	\$103
Government welfare payments	\$5
Government unemployment insurance payments	\$2
Government payroll	\$3
Government outlays for equipment and material	\$2
Depreciation rate	7%
Interest rate	6%

- a. Assuming the government budget for 2001 was in balance, calculate total investment, government purchases, real GDP, total saving, and net taxes for this economy.
 - b. Calculate total leakages and total injections.
 - c. Now suppose, instead, that the government increased its spending by \$2 billion for the year with no change in taxes. Explain how the variables from (a) will be affected (i.e., will they increase or decrease?).
 - d. Draw a graph depicting the situation in the loanable funds market and reflecting the assumption of a balanced budget. Clearly label the equilibrium interest rate, saving, and demand for funds. Now, add another curve reflecting any change that occurs when the government runs a deficit; show what happens to the variables you discussed in (c).
 - e. Under the assumption in (c), suppose Sylvania has a usury law that prohibits interest rates from going above 6 percent. Explain what will happen now in the loanable funds market, and in the economy as a whole.
3. Using a three-panel graph similar in style to Figure 7, illustrate how the *supply* of funds curve is obtained when the government is running a budget surplus.
 4. Show that Say's law still holds when the government is running a surplus, rather than a deficit. (*Hint*: Use an argument similar to the one in the section titled "The Loanable Funds Market and Say's Law.")
 5. Use graphs to depict the effect on saving, investment, and the interest rate of a *decrease* in government spending when the government is running a budget surplus.

C H A L L E N G E Q U E S T I O N S

1. Using an analysis similar to the one in the "Using the Theory" section, show that a tax cut cannot increase total spending in the economy, under each of the following two assumptions:
 - a. Initially, *none* of the tax cut is saved, so that consumption spending rises by an amount equal to the tax cut.
 - b. Initially, the *entire* tax cut is saved, causing the supply of funds curve to shift rightward by an amount equal to the tax cut.
2. Assume the loanable funds market is in equilibrium. Influential media pundits begin to warn about impending economic doom—recession, layoffs, and so forth. Using graphs, discuss what might happen to the equilibrium interest rate and the equilibrium quantity of funds. Assume that the government budget is in balance—neither a deficit nor a surplus. (*Hint*: How would these warnings separately affect household and business behavior in the loanable funds market?)

E X P E R I E N T I A L E X E R C I S E

1. Use the *Wall Street Journal*, or Infotrac, to locate a recent article about U.S. fiscal policy. More specifically, look for an article that mentions both the interest rate and the rate of economic growth. Once you have found such an article, try to translate the argument into graphs similar to those you have encountered in this chapter. Is the story consistent with what you have learned? If yes, explain how. If not, how might you account for the discrepancy?

ECONOMIC GROWTH AND RISING LIVING STANDARDS

CHAPTER

21

Economist Thomas Malthus, writing in 1798, came to a striking conclusion: “Population, when unchecked, goes on doubling itself every twenty-five years, or increases in a geometrical ratio. . . . The means of subsistence . . . could not possibly be made to increase faster than in an arithmetic ratio.”¹ From this simple logic, Malthus forecast a horrible fate for the human race. There would be repeated famines and wars to keep the rapidly growing population in balance with the more slowly growing supply of food and other necessities. The prognosis was so pessimistic that it led Thomas Carlyle, one of Malthus’s contemporaries, to label economics “the dismal science.”

But history has proven Malthus wrong . . . at least in part. In the industrialized nations, living standards have increased beyond the wildest dreams of anyone alive in Malthus’s time. Economists today are optimistic about these nations’ long-run material prospects. At the same time, living standards in many of the less-developed countries have remained stubbornly close to survival level and, in some cases, have fallen below it.

What are we to make of this? Why have living standards steadily increased in some nations but not in others? And what, if anything, can governments do to speed the rise in living standards? These are questions about economic growth—the long-run increase in an economy’s output of goods and services.

In this chapter, you will learn what makes economies grow. Our approach will make use of the classical model, focusing on Key Step #4: What Happens When Things Change? As you’ll see, growth arises from *shifts* of the curves of the classical model. And by the end of this chapter, you will know why increasing the rate of economic growth is not easy. While nations can take measures to speed growth, each measure carries an opportunity cost. More specifically,

achieving a higher rate of growth in the long run generally requires some sacrifice in the short run.

CHAPTER OUTLINE

The Importance of Growth

What Makes Economies Grow?

Growth in Employment

How to Increase Employment
Employment Growth and Productivity

Growth of the Capital Stock

Investment and the Capital Stock
How to Increase Investment
Human Capital and Economic Growth

Technological Change

The Cost of Economic Growth

Budgetary Costs
Consumption Costs
Opportunity Costs of Workers’ Time
Sacrifice of Other Social Goals

Using the Theory: Economic Growth in the Less-Developed Countries

¹ Thomas Robert Malthus, *Essay on the Principle of Population*, 1798.

THE IMPORTANCE OF GROWTH

Why should we be concerned about economic growth? For one simple reason:

Average standard of living Total output (real GDP) per person.

When output grows faster than the population, GDP per capita—which we call the average standard of living—will rise. When output grows more slowly than the population, the average standard of living will fall.

Measuring the standard of living by GDP per capita may seem limiting. After all, as we saw two chapters ago, many important aspects of our quality of life are not captured in GDP. Leisure time, workplace safety, good health, a clean environment—we care about all of these. Yet they are not considered in GDP.

Still, many aspects of our quality of life *are* counted in GDP: food, housing, medical care, education, transportation services, and movies and video games, to name a few. It is not surprising, then, that economic growth—measured by increases in GDP—remains a vital concern in every nation.

Economic growth is especially important in countries with income levels far below those of Europe, Japan, and the United States. The average standard of living in some third-world nations is so low that many families can barely acquire the basic necessities of life, and many others perish from disease or starvation. Table 1 lists GDP per capita, infant mortality rates, life expectancies, and adult literacy rates for some of the richest and poorest countries. The statistics for the poor countries are grim enough, but even they capture only part of the story. Unsafe and unclean workplaces, inadequate housing, and other sources of misery are part of daily life for most people in these countries. Other than emigration, economic growth is their only hope.

Growth is a high priority in prosperous nations, too. As we know, resources are scarce, and we cannot produce enough of everything to satisfy all of our desires simultaneously. We want more and better medical care, education, vacations, enter-

TABLE 1

SOME INDICATORS OF ECONOMIC WELL-BEING IN RICH AND POOR COUNTRIES, 1997

Country	Real GDP per Capita	Infant Mortality Rate (per 1,000 Live Births)	Life Expectancy at Birth	Adult Literacy Rate
RICH COUNTRIES				
United States	\$29,010	6.6	76.7	Greater than 99%
Japan	\$24,070	4.4	80.0	Greater than 99%
France	\$22,030	6.0	78.1	Greater than 99%
United Kingdom	\$20,730	6.3	77.2	Greater than 99%
Italy	\$20,290	6.8	78.2	98.3%
POOR COUNTRIES				
Ghana	\$1,640	78.9	60.0	66.4%
Pakistan	\$1,560	95.1	64.0	40.9%
Azerbaijan	\$1,550	73.9	69.9	96.3%
Cambodia	\$1,290	106.0	53.4	66.0%
Sierra Leone	\$ 410	na	37.2	33.3%

Sources: United Nations Development Programme, *Human Development Report 1999* (available at <http://www.undp.org/hdro/report.html>), Table 1; U.S. Bureau of the Census, *Statistical Abstract of the United States, 1997* (available at <http://www.census.gov/prod/www/statistical-abstract-us.html>), Table 1336.

tainment . . . the list is endless. When output per capita is growing, it's at least *possible* for everyone to enjoy an increase in material well-being without anyone having to cut back. We can also accomplish important social goals—helping the poor, improving education, cleaning up the environment—by asking those who are doing well to sacrifice part of the rise in their material well-being, rather than suffer a drop.

But when output per capita stagnates, material gains become a fight over a fixed pie: The more purchasing power my neighbor has, the less is left for me. With everyone struggling for a larger piece of this fixed pie, conflict replaces cooperation. Efforts to help the less fortunate, wipe out illiteracy, reduce air pollution—all are seen as threats, rather than opportunities.

In the 1950s and 1960s, economic growth in the wealthier nations seemed to be taking care of itself. Economists and policy makers focused their attention on short-run movements around full-employment output, rather than on the growth of full-employment output itself. The real payoff for government seemed to be in preventing recessions and depressions—in keeping the economy operating as close to its potential as possible.

All of that changed starting in the 1970s, and economic growth became a national and international preoccupation. Like most changes in perception and thought, this one was driven by experience. Table 2 tells the story. It gives the average yearly growth rates of real GDP per capita for the United States and some of our key trading partners.

Over most of the postwar period, output in the more prosperous industrialized countries (such as the United States, the United Kingdom, and Canada) grew by 2 or 3 percent per year, while output in the less wealthy ones—those with some catching up to do—grew even faster. But beginning in the mid-1970s, all of these nations saw their growth rates slip.

In the late 1990s, only the United States and the United Kingdom returned to their previous high rates of growth, while the other industrialized countries continued to grow more slowly than their historical averages.

Looking at the table, you might think that this slowing in growth was rather insignificant. Does the tiny difference between the pre-1972 and the post-1972 growth rates in the United States really matter? Indeed, it does. Recall our example a few chapters ago in which an increase in the growth rate of around 1 percentage point over the past 26 years would mean that, today, our GDP per capita would be \$10,000 greater. Seemingly small differences in growth rates matter a great deal.

TABLE 2

Country	1948–1972	1972–1988	1988–1995	1995–1999
United States	2.2%	1.7%	1.0%	2.6%
United Kingdom	2.4	2.1	0.9	2.1
Canada	2.9	2.6	0.6	1.9
France	4.3	2.1	1.2	2.1
Italy	4.9	2.8	1.6	1.4
West Germany	5.7	2.2	1.3	0.9
Japan	8.2	3.3	2.1	1.5

**AVERAGE ANNUAL
GROWTH RATE OF
OUTPUT PER CAPITA**

Sources: Angus Maddison, *Phases of Capitalist Development* (Oxford: Oxford University Press, 1982); U.S. Census Bureau IDB Summary Demographic Data (<http://www.census.gov/ipc/www/idbsum.html>); and *Economic Report of the President*, 2000, Table B-110, and various World Bank publications. Note: Data for Germany includes West Germany only through 1995, and all of Germany from 1995–1999.

WHAT MAKES ECONOMIES GROW?

Today we understand much more about economic growth than we did in the days of Thomas Malthus. Yet virtually all of our modern ideas about growth are based on the classical model you studied in the previous chapter—and for good reason: Economic growth is a *long-run* phenomenon. The classical model is particularly well suited to analyze long-run economic problems, including the problem of growth.

From the classical model, we know that the economy tends to operate at its full-employment output level over the long run. When we think about the causes of economic growth, then, we should think about changes that would cause full-employment output to increase. In virtually all countries enjoying economic growth, the three most important causes are increases in employment, increases in the capital stock, and changes in technology. In the next several pages, we'll look at each of these in turn.

What Happens When
Things Change?



GROWTH IN EMPLOYMENT

In the long run, as the classical model shows, the economy tends to generate a job for just about everyone who wants to work at prevailing wage rates. Therefore, total employment will rise whenever the *labor force*—the number of people who have or want jobs—increases. But what causes the labor force to increase?

One possibility is an increase in labor *supply*: a rise in the number of people who would like to work at any given wage. This is illustrated in Figure 1 by a rightward shift in the labor supply curve. We'll discuss *why* the labor supply curve might shift later; here, we'll concentrate on the consequences of the shift.

Before the shift, the labor supply curve is L_1^S , the market clears at a wage of \$15 per hour, and the fully employed labor force is 100 million workers. The aggregate production function tells us that, with the given amounts of capital and land in the economy, and the given state of technology, 100 million workers can produce \$7 trillion in goods and services—the initial value of full-employment output. When the labor supply curve shifts to L_2^S , the market-clearing wage drops to \$12. Business firms—finding labor cheaper to hire—increase the number of workers employed along the labor demand curve, from point *A* to point *B*. The labor force increases to 120 million workers, and full-employment output rises to \$8 trillion.

But growth in employment can also arise from an increase in labor demand: a rise in the number of workers firms would like to hire at any given wage. Once again, we'll consider the *causes* of labor demand changes momentarily; here, we focus on the *consequences*.

Graphically, an increase in labor demand is represented by a rightward shift in the labor demand curve, as in Figure 2. As the wage rate rises from \$15 to its new equilibrium of \$17, we move along the labor supply curve from point *A* to point *B*. More people decide they want to work as the wage rises. Equilibrium employment once again rises from 100 million to 120 million workers, and full-employment output rises from \$7 trillion to \$8 trillion. Thus,

growth in employment can arise from an increase in labor supply (a rightward shift in the labor supply curve) or an increase in labor demand (a rightward shift of the labor demand curve).

You may have noticed one very important difference between the labor market outcomes in Figures 1 and 2: When labor *supply* increases, the wage rate falls (from \$15 to \$12 in Figure 1); when labor *demand* increases, the wage rate rises (from

AN INCREASE IN LABOR SUPPLY

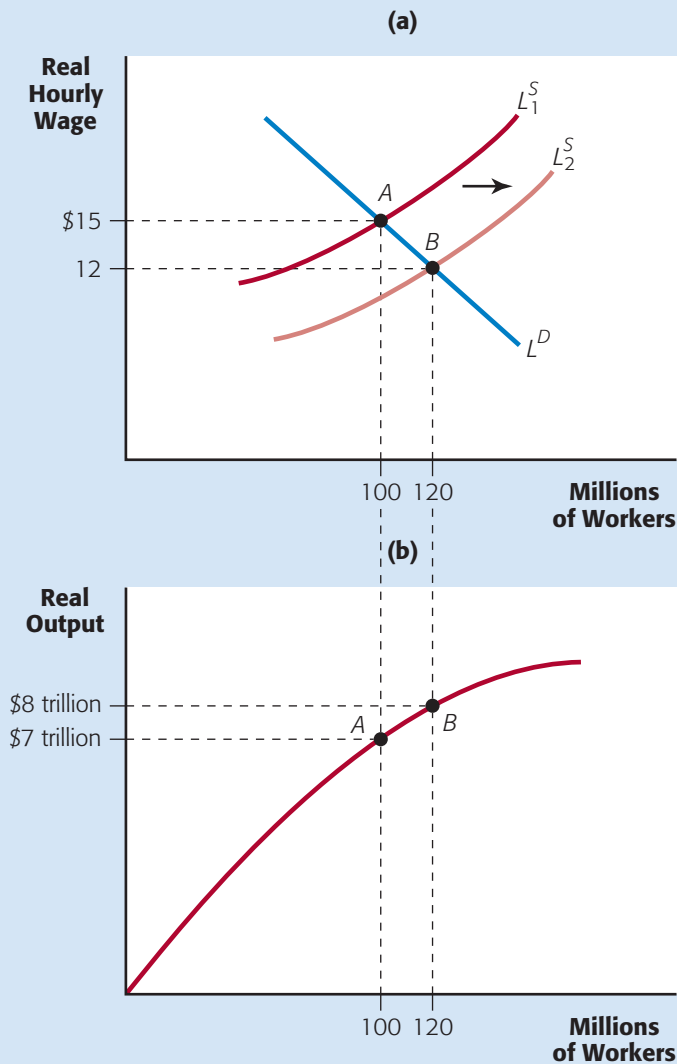


FIGURE 1

At point *A*, labor supply and demand determine an employment level of 100 million workers, and real GDP of \$7 trillion. An increase in labor supply will raise employment to 120 million (at point *B*), although with a lower wage rate. With more people working, real GDP rises to \$8 trillion.

\$15 to \$17 in Figure 2). Which of the figures describes the actual experience of the U.S. labor market?

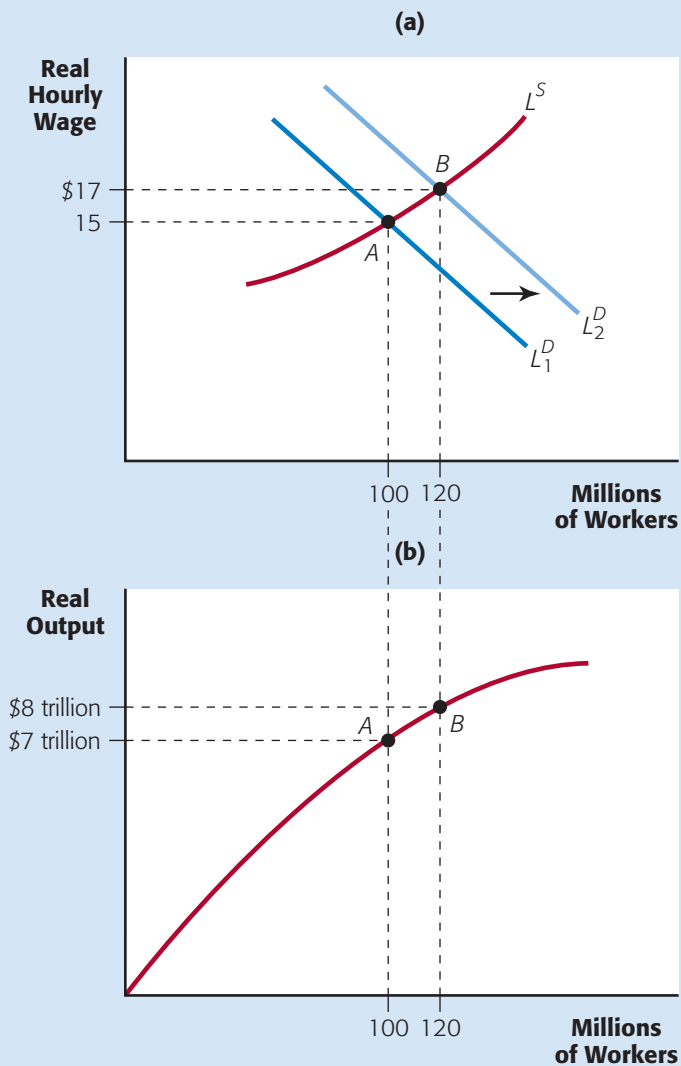
Actually, a combination of both: Over the past 50 years, the U.S. labor supply curve has shifted steadily rightward, sometimes slowly, sometimes more rapidly. Why the shift in labor supply? In part, the reason has been steady population growth: The more people there are, the more will want to work at any wage. But another reason has been an important change in tastes: an increase in the desire of women (especially married women) to work.

Over the past 50 years, as the labor supply curve has shifted rightward, the labor demand curve has shifted rightward as well. Why? Throughout this period, firms have been acquiring more and better capital equipment for their employees to use. Managers and accountants now keep track of inventories and other important accounts with lightning-fast computer software instead of account ledgers,

FIGURE 2

If firms demand more labor, employment will increase—from 100 million to 120 million—while the wage rate rises. With more people working, real GDP increases from \$7 trillion to \$8 trillion.

AN INCREASE IN LABOR DEMAND



supermarket clerks use electronic scanners instead of hand-entry cash registers, and college professors or their research assistants now gather data by searching for a few hours on the Web instead of a few weeks in the library. At the same time, workers have become better educated and better trained. These changes have increased the amount of output a worker can produce in any given period, so firms have wanted to hire more of them at any wage.²

In fact, over the past century, increases in labor demand have outpaced increases in labor supply, so that, on balance, the average wage has risen and employment has increased. This is illustrated in Figure 3, which shows a shift in the labor supply curve from L_1^S to L_2^S , and an even greater shift in the labor demand curve from L_1^D to L_2^D .

² These changes in physical and human capital have also shifted the economy's production function, but we'll consider that in the next section.

THE U.S. LABOR MARKET OVER A CENTURY

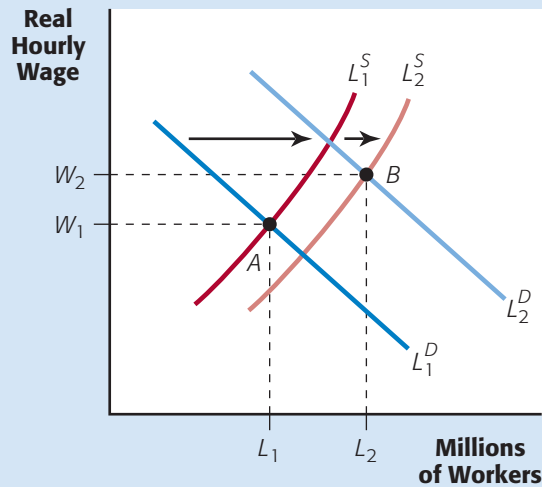


FIGURE 3

Over the past century, increases in labor demand have outpaced increases in supply. As a result, both the level of employment and the average wage have risen.

The impact of these changes on total employment has been dramatic. Between 1947 and 1999, the *labor force participation rate*—the fraction of the adult population that is either working or looking for work—rose from 58.3 percent to 67.1 percent. The increased participation rate was due partly to women’s increased tastes for working, as mentioned, and partly to the increase in the average wage rate that made work more rewarding. Together, growth in the population and in the participation rate have increased the U.S. labor force from 59.4 million workers in 1947 to 139.4 million workers in 1999.

Currently, the U.S. Bureau of Labor Statistics predicts employment growth of 1 percent per year until the year 2010. Is there anything we can do to make employment grow even faster, and thus increase our rate of economic growth? Can we speed up the rightward shifts in labor supply and labor demand? Yes, we can. But as you read on, keep in mind that these measures to increase employment are not necessarily socially desirable. These measures would, most likely, accomplish the goal, but they would also have costs—costs that Americans may or may not be willing to pay. Later, we’ll discuss these costs.

HOW TO INCREASE EMPLOYMENT

One set of policies to increase employment focuses on changing labor supply. And an often-proposed example of this type of policy is a decrease in income tax rates. Imagine that you have a professional degree in accounting, physical therapy, or some other field, and you are considering whether to take a job. Suppose the going rate for your professional services is \$30 per hour. If your average tax rate is 33 percent, then one-third of your income will be taxed away, so your take-home pay would be only \$20 per hour. But if your tax rate were cut to 20 percent, you would take home \$24 per hour. Since you care about your take-home pay, you will respond to a tax cut in the same way you would respond to a wage increase—even if the wage your potential employer pays does not change at all. If you would be willing to take a job that offers a take-home pay of \$24, but not one that offers \$20, then the tax cut would be just what was needed to get you to seek work.

When we extend your reaction to the population as a whole, we can see that a cut in the income tax rate can convince more people to seek jobs at any given wage, shifting the labor supply curve rightward. This is why economists and politicians who focus on the economy's long-run growth often recommend lower taxes on labor income to encourage more rapid growth in employment. They point out that many American workers must pay combined federal, state, and local taxes of more than 40 cents out of each additional dollar they earn, and that this may be discouraging work effort in the United States.

In addition to tax rate changes, some economists advocate changes in government transfer programs to speed the growth in employment. They argue that the current structure of many government programs creates disincentives to work. For example, families receiving welfare payments, food stamps, unemployment benefits, and Social Security retirement payments all face steep losses in their benefits if they go to work or increase their work effort. Redesigning these programs might therefore stimulate growth in labor supply.

This reasoning was an important motive behind the sweeping reforms in the U.S. welfare system passed by Congress, and signed by President Clinton, in August 1996. Among other things, the reforms reduced the number of people who were eligible for benefits, cut the benefit amount for many of those still eligible, and set a maximum coverage period of five years for most welfare recipients. Later in this chapter, we'll discuss some of the costs of potentially growth-enhancing measures like this. Here, we only point out that changes in benefit programs have the potential to change labor supply.

A cut in tax rates increases the reward for working, while a cut in benefits to the needy increases the hardship of not working. Either policy can cause a greater rightward shift in the economy's labor supply curve than would otherwise occur and speed the growth in employment and output.

Government policies can also affect the labor demand curve. In recent decades, subsidies for education and training, such as government-guaranteed loans for college students or special training programs for the unemployed, have helped to increase the skills of the labor force and made workers more valuable to potential employers. Government also subsidizes employment more directly—by contributing part of the wage when certain categories of workers are hired—the disabled, college work-study participants, and, in some experimental programs, inner-city youth. By enlarging these programs, government could increase the number of workers hired at any given wage and thus shift the labor demand curve to the right:

Government policies that help increase the skills of the workforce or that subsidize employment more directly shift the economy's labor demand curve to the right, increasing employment and output.

Efforts to speed employment growth are controversial. In recent decades, those who prefer an activist government have favored policies to increase labor *demand* through government-sponsored training programs, more aid to college students, employment subsidies to firms, and similar programs. Those who prefer a more *laissez-faire* approach have generally favored policies to increase the labor *supply* by *decreasing* government involvement—lower taxes or a less generous social safety net.

EMPLOYMENT AND LABOR PRODUCTIVITY

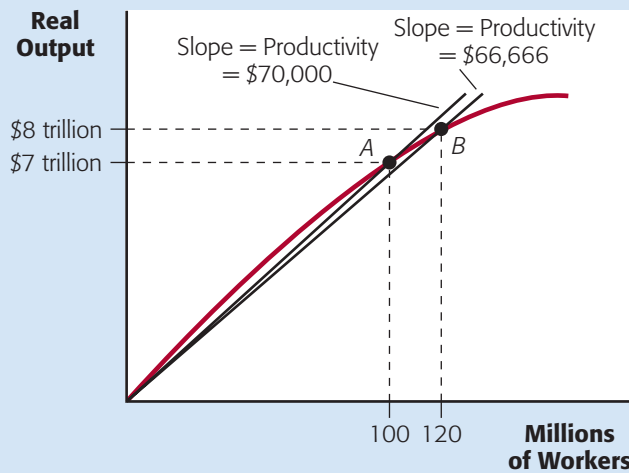


FIGURE 4

At any level of employment, labor productivity can be calculated by dividing total output by total employment. This is also shown by the slope of a line from the origin to a point on the production function. At point A, productivity is \$70,000 per worker. With more employment at point B, productivity is lower—\$66,666 per worker.

EMPLOYMENT GROWTH AND PRODUCTIVITY

Increases in employment have been an important source of economic growth in the United States and many other countries. But growth from this source has a serious drawback: It does not necessarily raise a nation's standard of living. Indeed, it can even cause living standards to fall. Why? Because living standards are closely tied to **labor productivity** (sometimes just called **productivity**)—the nation's total output divided by the total number of workers that produce it. Productivity is the output produced by the average worker in a year.³

Figure 4 illustrates the relationship between labor productivity and the economy's production function. At any level of employment, productivity is calculated by dividing total yearly output (on the vertical axis) by the total number of workers (on the horizontal axis):

$$\text{Productivity} = \frac{\text{output}}{\text{employment}} = \frac{\text{vertical measure}}{\text{horizontal measure}}$$

For example, in the figure, 100 million workers can produce \$7 trillion in output. Productivity at this level of employment is thus \$7 trillion/100 million = \$70,000 per worker, which is the slope of the line drawn from the origin to point A on the production function.⁴

Now look at what happens when employment rises to 120 million workers: Labor productivity falls to \$8 trillion/120 million = \$66,666 per worker, the slope of

Labor productivity Total output (real GDP) per worker.

³ Productivity is more often defined as total output divided by total *labor hours*—the output produced by the average worker in an hour. But our calculations will be easier if we use the definition given in the text. As long as the typical worker's hours remain unchanged, the two definitions of productivity—output per hour or output per worker per year—will rise or fall by the same percentage.

⁴ The slope of a straight line is always “rise over run,” or the change along the vertical axis divided by the change along the horizontal axis between any two points. Since our straight line begins at the origin, we can use the origin as our first point, so that the change in the vertical axis is just total output and the change in the horizontal axis is total employment. This gives us total output/total employment as the slope of the line.



Paul Bauer's "Are We in a Productivity Boom?" provides a more in-depth exploration of recent U.S. productivity experience. It's available at <http://www.clev.frb.org/research/com99/index.htm>.

the line drawn from the origin to point *B* on the production function. In fact, as you can see in the figure, as employment rises, labor productivity drops.

Why? The answer lies with the assumption that the production function remains unchanged. As we move rightward along a given production function, like the one in Figure 4, we are assuming that the nation's capital stock is constant. As a consequence, as employment increases, each worker has less and less capital equipment with which to work, and the average worker's output falls. If 100 ditchdiggers have 100 shovels, then each has his own shovel. If we double the number of ditchdiggers, but hold constant the number of shovels, then each worker must share his shovel with another and digs fewer ditches in any period. Labor productivity decreases.

When employment increases, while the capital stock remains constant, the amount of capital available to the average worker will decrease, and labor productivity will fall.

Falling labor productivity is bad news for a society. If output per worker falls, then the average standard of living will ordinarily fall as well. What can be done to prevent the fall in labor productivity as employment grows? Or—even better—can anything be done to *increase* labor productivity even as more people are working? The answer is yes, as you'll see in the next section.

What Happens When
Things Change?



GROWTH OF THE CAPITAL STOCK

The key to increasing labor productivity is to increase the nation's stock of capital. Has your college or university acquired more computers, desks, or campus-patrol vehicles in the past year? Did it install a new phone system? Build a new classroom or dormitory? If the answer to any of these questions is yes, then your school has helped create growth of the U.S. capital stock. With more capital—more assembly lines, bulldozers, computers, factory buildings, and the like—a given number of workers can produce more output than before, so the production function will *shift upward*.

Figure 5 shows the shift. With the initial amount of capital, the economy operates at point *A* on the lower aggregate production function, where 100 million workers produce \$7 trillion in output. The increase in capital shifts the production function upward, and—with the same employment level—the economy now operates at point *D*, where 100 million workers produce \$8 trillion in output.⁵

Looking back to Figure 4, and comparing it with Figure 5, you'll notice that output increases by the same amount in both cases; but the consequences for productivity are very different. In Figure 4, an increase in employment causes labor productivity to fall; in Figure 5, an increase in capital causes labor productivity to rise. (How do we know that productivity rises in Figure 5? *Hint*: Compare the slopes of the line through point *A* and the line through point *D*.)

An increase in the capital stock causes labor productivity and living standards to increase.

⁵ In order to focus on the pure effects of an increase in capital, Figure 5 holds the level of employment constant. But as you learned earlier in this chapter, an increase in capital will make workers more productive, and firms will want to hire more of them at any given wage. Thus, a complete analysis of capital growth would show the labor demand curve shifting rightward at the same time as the production function shifts upward.

CAPITAL ACCUMULATION AND LABOR PRODUCTIVITY

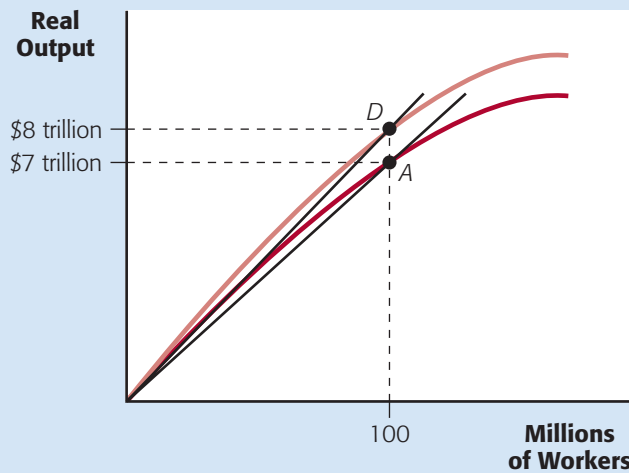


FIGURE 5

An increase in the capital stock shifts the production function upward. At point *A*, 100 million workers could produce \$7 trillion of real GDP; labor productivity is \$70,000 per worker. With more capital, those same workers could produce \$8 trillion of real GDP; productivity is then higher, at \$80,000 per worker.

To summarize, when the labor force grows (with a constant capital stock), labor productivity falls; and when the capital stock grows (with a constant labor force), productivity rises. These are interesting hypothetical cases. But in the real world, both the capital stock and the labor force grow from year to year. What happens to labor productivity when both changes occur simultaneously? That depends on what happens to **capital per worker**—the total quantity of capital divided by total employment. Greater capital per worker means greater productivity: You can dig more ditches with a shovel than with your bare hands, and even more with a backhoe.

Capital per worker The total capital stock divided by total employment.

If the capital stock grows faster than employment, then capital per worker will rise, and labor productivity will increase along with it. But if the capital stock grows more slowly than employment, then capital per worker will fall, and labor productivity will fall as well.

In the United States and most other developed countries, the capital stock has grown more rapidly than the labor force. As a result, labor productivity has risen over time. But in some developing countries, the capital stock has grown at about the same rate as, or even more slowly than, the population, and labor productivity has remained stagnant or fallen. We will return to this problem in the “Using the Theory” section of this chapter.

INVESTMENT AND THE CAPITAL STOCK

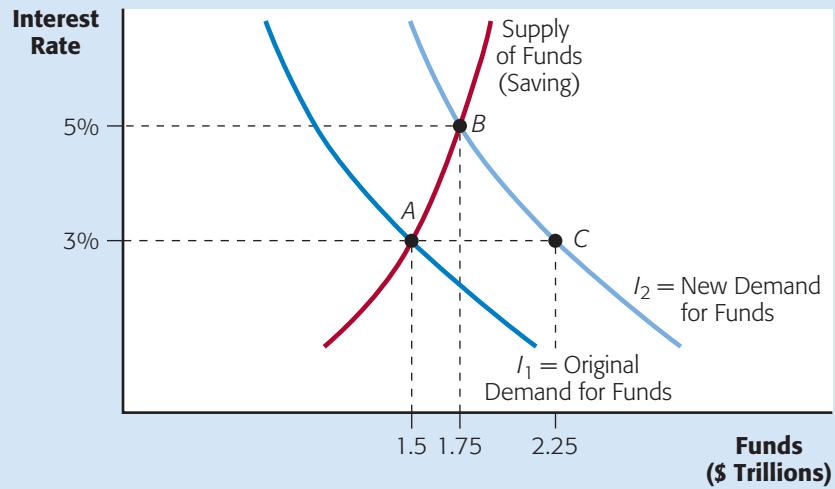
Now you can see why an increase in the capital stock plays such a central role in economists’ thinking about growth: It works by raising labor productivity and thus unambiguously helps to raise living standards. But how does a nation’s capital stock grow?

To answer this question, it’s important to realize that capital is a *stock variable*. As you learned a few chapters ago, a stock variable measures a quantity at a moment in time. More specifically, the capital stock is a measure of total plant and equipment in the economy at any moment. Planned investment, on the other hand, is a *flow variable*—it measures a process that takes place over a period of time. In this case, the flow is the rate at which we are producing *new* plant and equipment over some period. The

FIGURE 6

Government policies that make investment more profitable will increase investment spending at each interest rate. The resulting rightward shift of the investment demand curve leads to a higher level of investment spending, at point *B*.

AN INCREASE IN INVESTMENT SPENDING



relationship between the capital stock and the flow of investment is similar to that between the flow of water into a bathtub and the total amount of water in the tub itself. As long as investment is greater than depreciation (more water flows into the tub than drains out), the total stock of capital (the quantity of water in the tub) will rise. Moreover, the greater the flow of investment, the faster will be the rise in the capital stock.

HOW TO INCREASE INVESTMENT

A government seeking to spur investment has more than one weapon in its arsenal. It can direct its efforts toward businesses themselves, toward the household sector, or toward its own budget.

Targeting Businesses: Increasing the Incentive to Invest. One kind of policy to increase investment targets the business sector itself, with the goal of increasing planned investment spending. Figure 6 shows how this works. The figure shows a simplified view of the loanable funds market where—to focus on investment—we assume that there is no budget deficit, so there is no government demand for funds. The initial equilibrium in the market is at point *A*, where household saving (the supply of funds) and investment (the demand for funds) are both equal to \$1.5 trillion and the interest rate is 3 percent. Now suppose that the government takes steps to make investment more profitable, so that—at any interest rate—firms will want to purchase \$0.75 trillion more in capital equipment than before. Then the investment curve would shift rightward by \$0.75 trillion—from I_1 to I_2 , and the interest rate would rise from 3 percent to 5 percent. Note that, as the interest rate rises, some—but not all—of the original increase in planned investment is choked off. In the end, investment rises from \$1.5 trillion to \$1.75 trillion, and so each year \$0.25 trillion more is added to the capital stock than would otherwise be added.

These are the mechanics of a rightward shift in the investment curve. But what government measures would *cause* such a shift in the first place? That is, how could the government help to make investment spending more profitable for firms?

One such measure would be a reduction in the **corporate profits tax**, which would allow firms to keep more of the profits they earn from investment projects.

Corporate profits tax A tax on the profits earned by corporations.

AN INCREASE IN SAVING

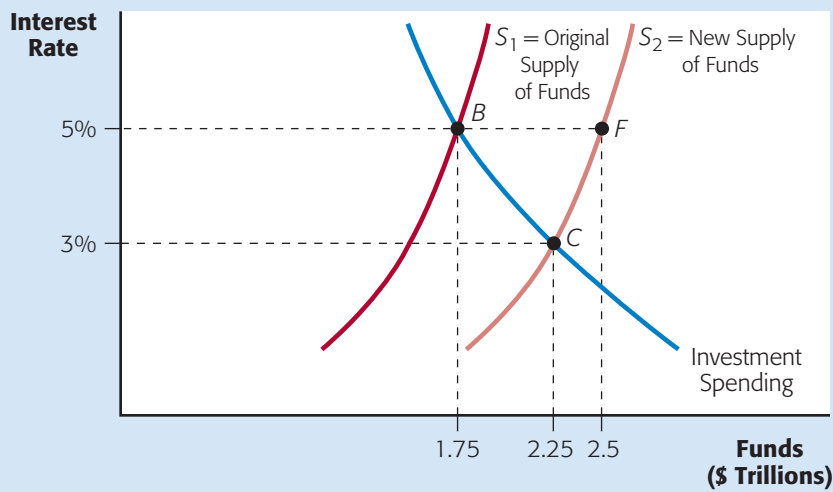


FIGURE 7

If households decide to save more of their incomes, the supply of funds will increase. With more funds available, the interest rate will fall. Businesses will respond by increasing their borrowing, and investment will increase from \$1.75 trillion to \$2.25 trillion.

Another, even more direct, policy is an **investment tax credit**, which subsidizes corporate investment in new capital equipment.

Reducing business taxes or providing specific investment incentives can shift the investment curve rightward, thereby speeding growth in physical capital, and increasing the growth rate of living standards.

Of course, the same reasoning applies in reverse: An *increase* in the corporate profits tax or the *elimination* of an investment tax credit would shift the investment curve to the left, slowing the rate of investment, the growth of the capital stock, and the rise in living standards.

Targeting Households: Increasing the Incentive to Save. While firms make decisions to purchase new capital, it is largely households that supply the firms with funds, via personal saving. Thus, an increase in investment spending can originate in the household sector, through an increase in the desire to save. This is illustrated in Figure 7. If households decide to save more of their incomes at any given interest rate, the supply of funds curve will shift rightward, from S_1 to S_2 . The increase in saving drives down the interest rate, from 5 percent to 3 percent, which, in turn, causes investment to increase. With a lower interest rate, NBC might decide to borrow funds to build another production studio, or the corner grocery store may finally decide to borrow the funds it needs for a new electronic scanner at the checkout stand. In this way, an increase in the desire to save is translated—via the financial market—into an increase in investment and faster growth in the capital stock.

What might cause households to increase their saving? The answer is found in the reasons people save in the first place. And to understand these reasons, you needn't look farther than yourself or your own family. You might currently be saving for a large purchase (a car, a house, a vacation, college tuition) or to build a financial cushion in case of hard times ahead. You might even be saving to support yourself during retirement, though this is a distant thought for most college students. Given these motives, what would make you save more? Several things: greater uncertainty about

Investment tax credit A reduction in taxes for firms that invest in certain favored types of capital.

your economic future, an increase in your life expectancy, anticipation of an earlier retirement, a change in tastes toward big-ticket items, or even just a change in your attitude about saving. Any of these changes—if they occurred in many households simultaneously—would shift the saving curve (the supply of funds curve) to the right, as in Figure 7.

Capital gains tax A tax on profits earned when a financial asset is sold at more than its acquisition price.

But government policy can increase household saving as well. One often-proposed idea is to decrease the **capital gains tax**. A capital gain is the profit you earn when you sell an asset, such as a share of stock or a bond, at a higher price than you paid for it. By lowering the special tax rate for capital gains, households would be able to keep more of the capital gains they earn. As a result, stocks and bonds would become more rewarding to own, and you might decide to reduce your current spending in order to buy them. If other households react in the same way, total saving would rise, and the supply of funds to the financial market would increase.

Consumption tax A tax on the part of their income that households spend.

Another frequently proposed measure is to switch from the current U.S. income tax—which taxes all income whether it is spent or saved—to a **consumption tax**, which would tax only the income that households spend. A consumption tax could work just like the current income tax, except that you would deduct your saving from your income and pay taxes on the remainder. This would increase the reward for saving since, by saving, you would earn additional interest on the part of your income that would have been taxed away under an income tax. Individual retirement accounts, or IRAs, allow households to deduct limited amounts of saving from their incomes before paying taxes. A general consumption tax would go much further and allow *all* saving to be deducted.

Another proposal to increase household saving is to restructure the U.S. Social Security system, which provides support for retired workers who have contributed funds to the system during their working years. Because Social Security encourages people to rely on the government for income during retirement, they have less incentive to save for retirement themselves. The proposed restructuring would link workers' Social Security benefits to their actual contributions to the system, whereas under the current system some people receive benefits worth far more than the amount they have contributed.

Government can alter the tax and transfer system to increase incentives for saving. If successful, these policies would make more funds available for investment, speed growth in the capital stock, and speed the rise in living standards.

(Do any of these methods of increasing saving disturb you? Remember, we are not advocating any measures here; rather, we are merely noting that such measures would increase saving and promote economic growth. We'll discuss the *costs* of growth-promoting measures later.)

Shrinking the Government's Budget. A final pro-investment measure is directed at the government sector itself. The previous chapter showed that an increase in government purchases, financed by borrowing in the financial market, completely crowds out consumption and investment. A *decrease* in government purchases has the opposite effect: raising consumption and investment.

Figure 8 reintroduces the government to the financial market to show how this works. Initially, the government is running a deficit of \$0.75 trillion, equal to the distance *EA*. The total demand for funds is now the sum of investment and the government's budget deficit, given by the curve labeled "Investment Spending + Deficit." The demand for funds curve intersects the supply of funds curve at point *A*, creating an equilibrium interest rate of 5 percent and equilibrium saving of \$1.75

DEFICIT REDUCTION AND INVESTMENT SPENDING

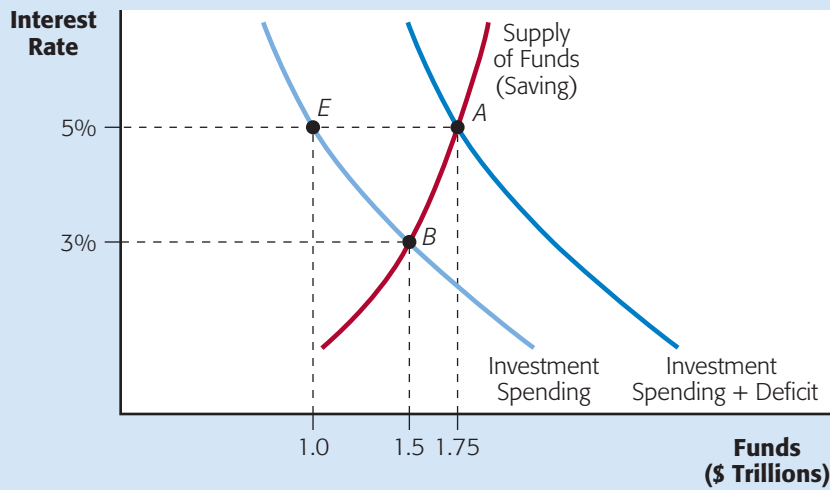


FIGURE 8

Eliminating the government's budget deficit will reduce government borrowing in the loanable funds market. As a result, the total demand for funds will fall, as will the interest rate. At a lower interest rate, businesses will increase their investment spending from \$1 trillion (point *E*) to \$1.5 trillion (point *B*).

trillion. At this interest rate, investment spending is only \$1 trillion. The part of saving not going to finance investment spending (\$1.75 trillion – \$1 trillion = \$0.75 trillion) is being used to finance the budget deficit.

Now consider what happens if the government eliminates the deficit—say, by reducing its purchases by \$0.75 trillion. The demand for funds would consist of investment spending only. Since there would be no other borrowing, the new equilibrium would be point *B*, with an interest rate of 3 percent and investment equal to \$1.5 trillion—greater than before. By balancing its budget, the government no longer needs to borrow in the loanable funds market, which frees up funds to flow to the business sector instead. Initially, this creates a surplus of funds. But—as the loanable funds market clears—the interest rate drops, and the surplus of funds disappears. (Why does a drop in the interest rate make the surplus disappear? *Hint*: What happens to saving and to investment as the interest rate declines?)

The link between the government budget, the interest rate, and investment spending is the major reason why the U.S. government, and governments around the world, try to reduce and, if possible, eliminate budget deficits. They have learned that

a shrinking deficit or a rising surplus tends to reduce interest rates and increase investment, thus speeding the growth in the capital stock.

In the 1990s, Congress set strict limits on the growth of government spending, and the budget deficit began shrinking. In 1998, the federal budget turned from deficit to surplus, and was projected to remain in surplus for at least a decade. These surpluses are helping to keep interest rates low, which in turn leads to greater business investment spending. The hope is that this will lead to a higher capital stock and greater productivity than we would have without the budget surpluses.

An Important Proviso About the Government Budget. A reduction in the deficit or an increase in the surplus—even if they stimulate private investment—are not *necessarily* pro-growth measures. It depends on *how* the budget changes. By an increase in taxes? A cut in government spending? And if the latter, which government

programs will be cut? Welfare? National defense? Highway repair? The answers can make a big difference to the impact on growth.

For example, in our discussions of the capital stock so far, we've ignored government capital—roads, communication lines, bridges, and dams. To understand the importance of government capital, just imagine what life would be like without it. How would factories obtain their raw materials or distribute their goods if no one repaired the roads? How would contracts between buyers and sellers be enforced if there were no public buildings to house courts and police departments? Government capital supports private economic activity in more ways than we can list here.

Government investment in new capital and in the maintenance of existing capital makes an important contribution to economic growth.

This important observation complicates our view of deficit reduction. It is still true that a decrease in government spending will lower the interest rate and increase private investment. But if the budget cutting falls largely on government investment, the negative effect of smaller public investment will offset some of the positive impact of greater private investment. Shrinking the deficit will then alter the *mix* of capital—more private and less public—and the effect on growth could go either way. A society rife with lawlessness, deteriorating roads and bridges, or an unreliable communications network might benefit from a shift toward public capital. For example, a study of public budgets in African nations—which have poor road conditions—found that each one-dollar-per-year cut in the road-maintenance budget increased vehicle operating costs by between \$2 and \$3 per year, and in one case, by as much as \$22 per year.⁶ This is an example where a cut in government spending—even if it reduces the deficit—probably hinders growth. By contrast, a stable society (Sweden comes to mind) with a fully developed and well-maintained public infrastructure might be able to have faster growth by shifting the mix away from public and toward private capital.

The impact of deficit reduction on economic growth depends on which government programs are cut. Shrinking the deficit by cutting government investment will not stimulate growth as much as would cutting other types of government spending.

HUMAN CAPITAL AND ECONOMIC GROWTH

So far, the only type of capital we've discussed is physical capital—the plant and equipment workers use to produce output. But when we think of the capital stock most broadly, we include *human capital* as well. **Human capital**—the skills and knowledge possessed by workers—is as central to economic growth as is physical capital. After all, most types of physical capital—computers, CAT scanners, and even shovels—will contribute little to output unless workers know how to use them. And when more workers gain skills or improve their existing skills, output rises just as it does when workers have more physical capital:

An increase in human capital works like an increase in physical capital to increase output: It causes the production function to shift upward, raises productivity, and increases the average standard of living.

Human capital Skills and knowledge possessed by workers.

⁶ This World Bank study was cited in *The Economist*, June 10, 1995, p. 72.

There is another similarity between human and physical capital: Both are *stocks* that are increased by *flows* of investment. The stock of human capital increases whenever investment in new skills during some period, through education and training, exceeds the depreciation of existing skills over the same period, through retirement, death, or deterioration. Therefore, greater investment in human capital will speed the growth of the human capital stock, the growth in productivity, and the growth in living standards.

Human capital investments are made by business firms (when they help to train their employees), by government (through public education and subsidized training), and by households (when they pay for general education or professional training). Human capital investments have played an important role in recent U.S. economic growth. Can we do anything to increase our rate of investment in human capital?

In part, we've already answered this question: Some of the same policies that increase investment in *physical* capital also work to raise investment in human capital. For example, a decrease in the budget deficit would lower the interest rate and make it cheaper for households to borrow for college loans and training programs. A change in the tax system that increases the incentive to save would have the same impact, since this, too, would lower interest rates. And an easing of the tax burden on business firms could increase the profitability of *their* human capital investments, leading to more and better worker training programs.

But there is more: Human capital, unlike physical capital, cannot be separated from the person who provides it. If you own a building, you can rent it out to one firm and sell your labor to another. But if you have training as a doctor, your labor and your human capital must be sold together, as a package. Moreover, your wage or salary will be payment for both your labor and your human capital. This means that income tax reductions—which we discussed earlier as a means of increasing labor supply—can also increase the profitability of human capital to households, and increase their rate of investment in their own skills and training. For example, suppose an accountant is considering whether to attend a course in corporate financial reporting, which would increase her professional skills. The course costs \$4,000, and will increase the accountant's income by \$1,000 per year for the rest of her career. With a tax rate of 40 percent, her take-home pay would increase by \$600 per year, so her annual rate of return on her investment would be $\$600/\$4,000 = 15$ percent. But with a lower tax rate—say, 20 percent—her take-home pay would rise by \$800 per year, so her rate of return would be $\$800/\$4,000 = 20$ percent. The lower the tax rate, the greater is the rate of return on our accountant's human capital investment, and the more likely she will be to acquire new skills. Thus,

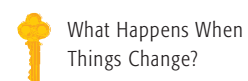
many of the pro-growth policies discussed earlier—policies that increase employment or increase investment in physical capital—are also effective in promoting investment in human capital.

TECHNOLOGICAL CHANGE

So far, we've discussed how economic growth arises from greater quantities of resources—more labor, more physical capital, or more human capital. But another important source of growth is **technological change**—the invention or discovery of new inputs, new outputs, or new methods of production. Indeed, it is largely because of



College-level courses are one important way that countries increase the stock of human capital and shift up their production function.



Technological change The invention or discovery of new inputs, new outputs, or new production methods.

technological change that Malthus's horrible prediction (cited at the beginning of this chapter) has not come true. In the last 60 years, for example, the inventions of synthetic fertilizers, hybrid corn, and chemical pesticides have enabled world food production to increase faster than population.

New technology affects the economy in much the same way as do increases in the capital stock. Flip back 7 pages to Figure 5. There, you saw that an increase in the capital stock would shift the production function upward and increase output. New technology, too, shifts the production function upward, since it enables any given number of workers to produce more output. In many cases, the new technology requires the acquisition of physical and human capital before it can be used. For example, a new technique for destroying kidney stones with ultrasound, rather than time-consuming surgery, can make doctors more productive—but not until they spend several thousand dollars to buy the ultrasound machine and take a course on how to use it. In other cases, a new technology can be used without any additional equipment or training, as when a factory manager discovers a more efficient way to organize workers on the factory floor. In either case, technological change will shift the production function upward and increase productivity. It follows that

the faster the rate of technological change, the greater the growth rate of productivity, and the faster the rise in living standards.

It might seem that technological change is one of those things that just happens. Thomas Edison invents electricity, or Steve Jobs and Steve Wozniak develop the first practical personal computer in their garage. But the pace of technological change is not as haphazard as it seems. The transistor was invented as part of a massive research and development effort by AT&T and intended to improve the performance of communications electronics. Similarly, the next developments in computer technology, transportation, and more will depend on how much money is spent on research and development (R&D) by the leading technology firms:

The rate of technological change in the economy depends largely on firms' total spending on R&D. Policies that increase R&D spending will increase the pace of technological change.

What can the government do to increase spending on R&D? First, it can increase its own direct support for R&D by carrying out more research in its own laboratories or increasing funding for universities and tax incentives to private research labs.

Second, the government can enhance **patent protection**, which increases rewards for those who create new technology by giving them exclusive rights to use it or sell it. For example, when the DuPont Corporation discovered a unique way to manufacture Spandex, it obtained a patent to prevent other firms from copying its technique. This patent has enabled DuPont to earn millions of dollars from its invention. Without the patent, other firms would have copied the technique, competed with DuPont, and taken much of its profit away. Hundreds of thousands of new patents are issued every year in the United States: to pharmaceutical companies for new prescription drugs, to telecommunications companies for new cellular technologies, and to the producers of a variety of household goods ranging from can openers to microwave ovens.

Since patent protection increases the rewards that developers can expect from new inventions, it encourages them to spend more on R&D. By broadening patent protection—issuing patents on a wider variety of discoveries—or by lengthening

Patent protection A government grant of exclusive rights to use or sell a new technology.

patent protection—increasing the number of years during which the developer has exclusive rights to market the invention—the government could increase the expected profits from new technologies. That would increase total spending on R&D and increase the pace of technological change. Currently in the United States, patents give inventors and developers exclusive marketing rights over their products for a period of about 20 years. Increasing patent protection to 30 years would certainly increase R&D spending at many firms.

Finally, R&D spending is in many ways just like other types of investment spending: The funds are drawn from the financial market, and R&D programs require firms to buy something now (laboratories, the services of research scientists, materials to build prototypes) for the uncertain prospect of profits in the future. Therefore, almost any policy that stimulates investment spending in general will also increase spending on R&D. Cutting the tax rate on capital gains or on corporate profits, or lowering interest rates by encouraging greater saving or by reducing the budget deficit, can each help to increase spending on R&D and increase the rate of technological change.

THE COST OF ECONOMIC GROWTH

So far in this chapter, we've discussed a variety of policies that could increase the rate of economic growth and speed the rise in living standards. Why don't all nations pursue these policies and push their rates of economic growth to the maximum? For example, why did the U.S. standard of living (output per capita) grow by 2.6 percent per year between 1995 and 1999? Why not 4 percent per year? Or 6 percent? Or even more?

The answer hinges on one of the basic principles of economics:

Government policy is constrained by the reactions of private decision makers. As a result, policy makers face trade-offs: Making progress toward one goal often requires some sacrifice of another goal.

Economics is famous for making the public aware of policy trade-offs. One of the most important things you will learn in your introductory economics course is that there are no costless solutions to society's problems. Just as individuals face an opportunity cost when they take an action (they must give up something else that they value), so, too, policy makers face an opportunity cost whenever they pursue a policy: They must compromise on achieving some other social goal.

Economic models can help us identify the trade-offs associated with different policy choices. Although confronting a trade-off is rarely pleasant, doing so helps us formulate wiser policies and avoid unpleasant surprises. In this section, you will see that while a variety of policies can increase a nation's rate of economic growth, each of these policies involves a trade-off: It imposes a cost on some group or requires some sacrifice of other social goals.

Promoting economic growth involves unavoidable trade-offs: It requires some groups, or the nation as a whole, to give up something else that is valued. In order to decide how fast we want our economy to grow, we must consider growth's costs as well as its benefits.

What are the costs of growth?

BUDGETARY COSTS

If you look back over this chapter, you'll see that many of the pro-growth policies we've analyzed involve some kind of tax cut. Cutting the income tax rate will likely increase the labor supply. Cutting taxes on capital gains or corporate profits will increase investment directly. And cutting taxes on saving will increase household saving, lower interest rates, and thus increase investment spending indirectly. Unfortunately, implementing any of these tax cuts would force the government to choose among three unpleasant alternatives: increase some other tax to regain the lost revenue, cut government spending, or permit the budget deficit to rise.

Who will bear the burden of this budgetary cost? That depends on which alternative is chosen. Under the first option—increasing some other tax—the burden falls on those who pay the other tax. For example, if income taxes are cut, real estate taxes might be increased. A family might pay lower income taxes, but higher property taxes. Whether it comes out ahead or behind will depend on how much income the family earns relative to how much property it owns.

The second option, cutting government spending, imposes the burden on those who currently benefit from government programs. These include not only those who directly benefit from a program—like welfare recipients or farmers—but also those who benefit from government spending more indirectly. Even though you may earn your income in the private sector, if government spending is cut, you may suffer from a deterioration of public roads, decreased police protection, or poorer schools for your children.

The third option—a larger budget deficit or a smaller budget surplus—is more complicated. Suppose a tax cut causes the government to end up with a larger deficit. Then greater government borrowing will increase the total amount of government debt outstanding—called the national debt—and lead to greater interest payments to be made by future generations, in the form of higher taxes. The same is true even if the government is running a budget surplus. In that case, a tax cut will *reduce* the size of the surplus, and reduce the amount of the national debt the government pays back each year. Once again, the tax cut raises the interest payments that future generations must bear.

But that is not all. From the previous chapter, we know that a rise in the budget deficit (by increasing the demand for funds) or a drop in the budget surplus (by decreasing the supply of funds) drives up the interest rate. The higher interest rate will reduce investment in physical capital by businesses, as well as investment in human capital by households, and both effects will work to decrease economic growth. It is even possible that so much private investment will be crowded out that the tax cut, originally designed to boost economic growth, ends up slowing growth instead. At best, the growth-enhancing effects of the tax cut will be weakened. This is why advocates of high growth rates usually propose one of the other options—a rise in some other tax or a cut in government spending—as part of a pro-growth tax cut.

In sum,

properly targeted tax cuts can increase the rate of economic growth, but will force us to either redistribute the tax burden or cut government programs.

CONSUMPTION COSTS

Any pro-growth policy that works by increasing investment—private or government, in physical capital, human capital, or R&D—requires a sacrifice of current consumption spending. The land, labor, and capital we use to produce new cloth-

CONSUMPTION, INVESTMENT, AND ECONOMIC GROWTH

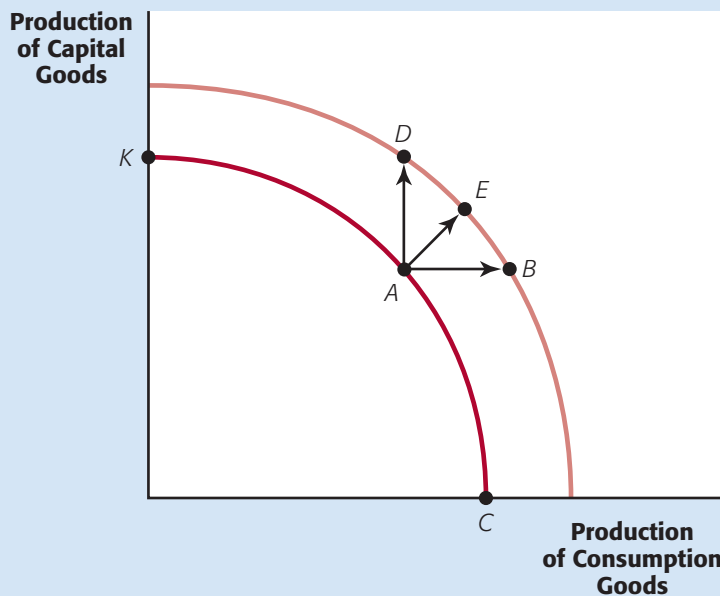


FIGURE 9

In the current period, a nation can choose to produce only consumer goods (point C), or it can produce some capital goods by sacrificing some current consumption, as at point A. If investment at point A exceeds capital depreciation, the capital stock will grow, and the production possibilities frontier will shift outward. After it does, the nation can produce more consumption goods (point B), more capital goods (point D), or more of both (point E).

cutting machines, oil rigs, assembly lines, training facilities, college classrooms, or research laboratories could have been used instead to produce clothing, automobiles, video games, and other consumer goods. In other words, we face a trade-off: The more capital goods we produce in any given year, the fewer consumption goods we can enjoy in that year.

The role of this trade-off in economic growth can be clearly seen with a familiar tool from Chapter 2: the production possibilities frontier (PPF). Figure 9 shows the PPF for a nation with some given amount of land, labor, and capital that must be allocated to the production of two types of output: capital goods and consumption goods. At point K, the nation is using all of its resources to produce capital goods and none to produce consumption goods. Point C represents the opposite extreme—all resources used to produce consumption goods and none for capital goods. Ordinarily, a nation will operate at an intermediate point such as A, where it is producing both capital and consumption goods.

Now, as long as capital production at point A is greater than the depreciation of existing capital, the capital stock will grow. In future periods, the economy—with more capital—can produce more output, as shown by the outward shift of the PPF in the figure. If a nation can produce more output, then it can produce more consumption goods for the same quantity of capital goods (moving from point A to point B) or more capital goods for the same quantity of consumption goods (from point A to point D) or more of both (from point A to point E).

Let's take a closer look at how this sacrifice of current consumption goods might come about. Suppose that some change in government policy—an investment tax credit or a lengthening of the patent period for new inventions—successfully shifts the investment curve to the right. (Go back to Figure 6.) What will happen? Businesses—desiring more funds for investment—will drive up the interest rate, and households all over the country will find that saving has become more attractive.

As families increase their saving, we move rightward along the economy's supply of funds curve. In this way, firms get the funds they need to purchase new capital. But a decision to *save more* is also a decision to *spend less*. As current saving rises, current consumption spending necessarily falls. By driving up the interest rate, *the increase in investment spending causes a voluntary decrease in consumption spending by households*. Resources are freed from producing consumption goods and diverted to producing capital goods instead.

Although this decrease in consumption spending is voluntary, it is still a cost that we pay. And in some cases, a painful cost: Some of the increase in the household sector's net saving results from a decrease in borrowing by households that—at higher interest rates—can no longer afford to finance purchases of homes, cars, or furniture. In sum,

greater investment in physical capital, human capital, or R&D will lead to faster economic growth and higher living standards in the future, but we will have fewer consumer goods to enjoy in the present.

OPPORTUNITY COSTS OF WORKERS' TIME

Living standards will also rise if a greater fraction of the population works or if those who already have jobs begin working longer hours. In either case, there will be more output to divide among the same population.⁷ But this increase in living standards comes at a cost: a decrease in time spent in nonmarket activities. For example, with a greater fraction of the population working, a smaller fraction is spending time at home. This might mean that more students have summer jobs instead of studying, more elderly workers are postponing their retirement, or more previously nonworking spouses are entering the labor force. Similarly, an increase in average working hours means that the average worker will have less time for other activities—less time to watch television, read novels, garden, fix up the house, teach his or her children, or do volunteer work.

Thus, when economic growth comes about from increases in employment, we face a trade-off: On the one hand, we can enjoy higher incomes and more goods and services; on the other hand, we will have less time to do things other than work in the market. In a market economy, where choices are voluntary, the value of the income gained must be greater than the value of the time given up. No one forces a worker to re-enter the labor force or to increase her working hours. Any worker who takes either of these actions must be better off for doing so. Still, we must recognize that *something* of value is always given up when employment increases:

An increase in the fraction of the population with jobs or a rise in working hours will increase output and raise living standards, but also requires us to sacrifice time previously spent in nonmarket activities.

SACRIFICE OF OTHER SOCIAL GOALS

Rapid economic growth is an important social goal, but it's not the only one. Some of the policies that quicken the pace of growth require us to sacrifice other goals that we

⁷ You might be wondering how a rise in average hours would be represented in the classical model we've been using. This is left to you as an exercise. But here's a hint: An increase in average hours enables the same number of workers to produce more output.

care about. For example, you've seen that restructuring Social Security benefits would increase saving, leading to more investment and faster growth. But such a move would cut the incomes of those who benefit from the current system and increase the burden on other social programs, such as welfare and food stamps. Extending patent protection would increase incentives for research and development. But it would also extend the monopoly power exercised by patent holders and force consumers to pay higher prices for drugs, electronic equipment, and even packaged foods.

Of course, the argument cuts both ways: Just as government policies to stimulate investment require us to sacrifice other goals, so, too, can the pursuit of other goals impede investment spending and economic growth. Most of us would like to see a cleaner environment and safer workplaces. But government safety and environmental regulations have increased in severity, complexity, and cost over time, reducing the rate of profit on new capital and shrinking investment spending.

Does this mean that business taxes and government regulations should be reduced to the absolute minimum? Not at all. As in most matters of economic policy, we face a trade-off:

We can achieve greater worker safety, a cleaner environment, and other social goals, but we may have to sacrifice some economic growth along the way. Alternatively, we can achieve greater economic growth, but we will have to compromise on other things we care about.

When values differ, people will disagree on just how much we should sacrifice for economic growth or how much growth we should sacrifice for other goals.

ECONOMIC GROWTH IN THE LESS-DEVELOPED COUNTRIES

In most countries, Malthus's dire predictions have not come true. An important part of the reason is that increases in the capital stock have raised productivity and increased the average standard of living. Increases in the capital stock are even more important in the less-developed countries (LDCs), which have relatively little capital to begin with and where even small increases in capital formation can have dramatic effects on living standards.

But how does a nation go about increasing its capital stock? As you've learned, there are a variety of measures, all designed to accomplish the same goal: shifting resources away from consumer-goods production toward capital-goods production. A very simple formula.

Some countries that were once LDCs—like the four Asian tigers (Hong Kong, Singapore, South Korea, and Taiwan)—have applied the formula very effectively. Output per capita in these countries has grown by an average of 6 percent per year over the past two decades. They were able to shift resources from consumption goods into capital goods in part by pursuing many of the growth-enhancing measures discussed in this chapter: large subsidies for human and physical capital investments, pro-growth tax cuts to encourage saving and investment, and the willingness to sacrifice other social goals—especially a clean environment—for growth.⁸ These economies gave up large amounts of potential consumption during a period of intensive capital formation.

Using the
THEORY



⁸ The Asian tigers also had some special advantages—such as a high level of human capital to start with.

TABLE 4

**ECONOMIC GROWTH IN
SELECTED POOR
COUNTRIES**

Country	Average Annual Growth Rate of Output per Capita	
	1975–85	1985–1997
Pakistan	3.5%	2.9%
Bangladesh	2.1%	2.8%
Ghana	–2.2%	1.8%
Kenya	0.6%	0.5%
Benin	1.9%	–0.3%
Democratic Republic of the Congo	–3.1%	–6.8%
Sierra Leone	–1.2%	–3.5%

Source: United Nations Development Programme, *Human Development Report 1999* (available at <http://www.undp.org/hdro/report.html>), Table 6.

But other LDCs have had great difficulty raising living standards. Table 4 shows growth rates for several of them. In some cases—such as Pakistan, Bangladesh, and more recently, Ghana—slow but consistent growth has given cause for optimism. In other cases—such as Kenya and Benin—living standards have barely budged over the past few decades. In still other cases—for example, the Democratic Republic of the Congo and Sierra Leone—output per capita has been falling steadily. Why do some LDCs have such difficulty achieving economic growth?

Much of the explanation for the low growth rates of many LDCs lies with three characteristics that they share:

1. *Very low current output per capita.* Living standards are so low in some LDCs that they cannot take advantage of the trade-off between producing consumption goods and producing capital goods. In these countries, pulling resources out of consumption would threaten the survival of many households. In the individual household, the problem is an inability to save: Incomes are so low that households must spend all they earn on consumption.

2. *High population growth rates.* Low living standards and high population growth rates are linked together in a cruel circle of logic. On the one hand, population growth by itself tends to reduce living standards; on the other hand, a low standard of living tends to increase population growth. Why? First, the poor are often uneducated in matters of family planning. Second, high mortality rates among infants and children encourage families to have many offspring, to ensure the survival of at least a few to care for parents in their old age. As a result, while the average woman in the United States will have fewer than two children in her lifetime, the average woman in Haiti will have about five children, and the average woman in Rwanda will have more than six.

3. *Poor infrastructure.* Political instability, poor law enforcement, corruption, and adverse government regulations make many LDCs unprofitable places to invest. Low rates of investment mean a smaller capital stock and lower productivity. Infrastructure problems also harm worker productivity in another way: Citizens must spend time guarding against thievery and trying to induce the government to let them operate businesses—time they could otherwise spend producing output.

These three characteristics—low current production, high population growth, and poor infrastructure—interact to create a vicious circle of continuing poverty, which we can understand with the help of the familiar PPF between capital goods and consumption goods. Look back at Figure 9, and now imagine that it applies to



<http://>

The World Bank Economic Growth Project's Web site is a comprehensive source of information about economic growth (http://www.worldbank.org/html/prdmg/grthweb/growth_t.htm).

a poor, developing country. In this case, an outward shift of the PPF does not, in itself, guarantee an increase in the standard of living. In the LDCs, the population growth rate is often very high, and—with a constant labor force participation ratio—employment grows at the same rate as the population. If employment grows more rapidly than the capital stock, then even though the PPF is shifting outward, capital per worker will decline. The result is falling labor productivity and a general decline in living standards.

In order to have a rising living standard, a nation's stock of capital must not only grow, but grow faster than its population.

Point N in Figure 10 shows the minimum amount of investment needed to increase capital per worker, labor productivity, and living standards for a given rate of population growth. For example, if the population is growing at 4 percent per year, then point N indicates the investment needed to increase the total capital stock by 4 percent per year. If investment is just equal to N , then capital per worker—and living standards—remains constant. If investment exceeds N , then capital per worker—and living standards—will rise. Of course, the greater the growth in population, the higher point N will be on the vertical axis, since greater investment will be needed just to keep up with population growth.

The PPF in Figure 10 has an added feature: Point S shows the minimum acceptable level of consumption—the amount of consumer goods the economy *must* produce in a year. For example, S might represent the consumption goods needed to prevent starvation among the least well off, or to prevent unacceptable social consequences, such as violent revolution.

Now we can see the problem faced by the most desperate of the less-developed economies. Output is currently at a point like H in Figure 10, with investment just equal to N . The capital stock is not growing fast enough to increase capital per worker, and so labor productivity and living standards are stagnant. In this situation, the PPF shifts outward each year, but not quickly enough to improve people's

LDC GROWTH AND LIVING STANDARDS

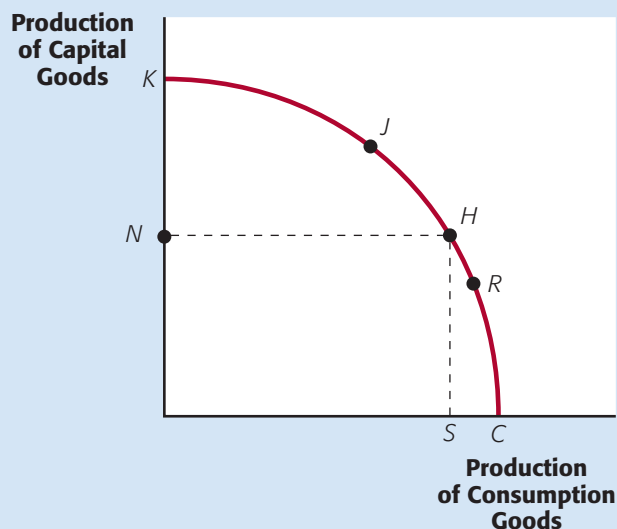


FIGURE 10

In order to increase capital per worker when population is growing, yearly investment spending must exceed some minimum level N . In any year, there is a minimum level of consumption, S , needed to support the population. If output is currently at point H , capital per worker and living standards are stagnant. But movement to a point like J would require an unacceptably low level of consumption.

lives. It could be even worse: Convince yourself that, at a point like *R*, the average standard of living declines even though the capital stock is growing—that is, even though the PPF will shift outward in future periods.

The solution to this problem appears to be an increase in capital production beyond point *N*—a movement *along* the PPF from point *H* to a point such as *J*. As investment rises above *N*, capital per worker rises, and the PPF shifts outward rapidly enough over time to raise living standards. In a wealthy country, like the United States, such a move could be engineered by changes in taxes or other government policies. But in the LDCs depicted here, such a move would be intolerable: At point *H*, consumption is already equal to *S*, the lowest acceptable level. Moving to point *J* would require reducing consumption *below S*.

The poorest LDCs are too poor to take advantage of the trade-off between consumption and capital production in order to increase their living standards. Since they cannot reduce consumption below current levels, they cannot produce enough capital to keep up with their rising populations.

In recent history, countries have attempted several methods to break out of this vicious circle of poverty. During the 1930s, the dictator Joseph Stalin simply *forced* the Soviet economy from a point like *H* to one like *J*. His goal was to shift the Soviet Union's PPF outward as rapidly as possible. But, as you can see, this reduced consumption below the minimum level *S*, and Stalin resorted to brutal measures to enforce his will. Many farmers were ordered into the city to produce capital equipment. With fewer people working on farms, agricultural production declined, and there was not enough food to go around. Stalin's solution was to confiscate food from the remaining farmers and give it to the urban workforce. Of course, this meant starvation for millions of farmers. Millions more who complained too loudly, or who otherwise represented a political threat, were rounded up and executed.

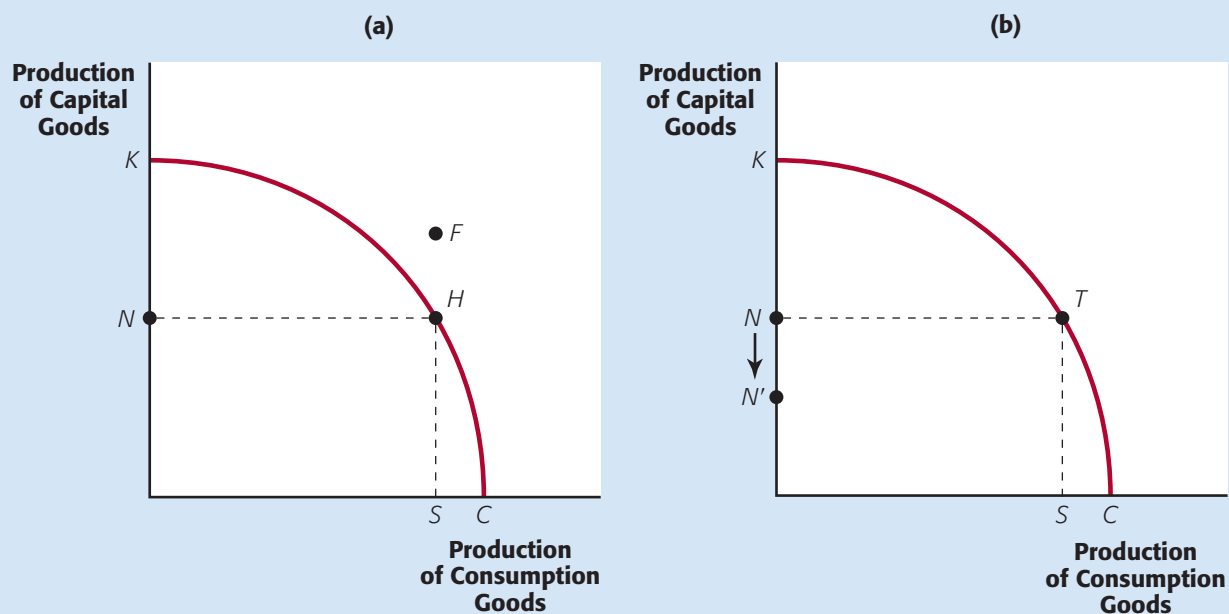
A less-brutal solution to the problem of the LDCs is to make the wealthy bear more of the burden of increasing growth. If the decrease in consumption can be limited to the rich, then total consumption can be significantly reduced—freeing up resources for investment—without threatening the survival of the poor. This, however, is not often practical, since the wealthy have the most influence with government in LDCs. Being more mobile, they can easily relocate to other countries, taking their savings with them. This is why efforts to shift the sacrifice to the wealthy are often combined with restrictions on personal liberties, such as the freedom to travel or to invest abroad. These moves often backfire in the long run, since restrictions on personal and economic freedom are remembered long after they are removed and make the public—especially foreigners—hesitant to invest in that country.

A third alternative—and the one used increasingly since the 1940s—is *foreign investment* or *foreign assistance*. If the wealthier nations—individually or through international organizations such as the World Bank or the International Monetary Fund—provide the LDCs with capital, then the capital *available* to them can increase, with *no* cutbacks in consumption. This permits an LDC to make *use* of capital and consumption goods at a point like *F* in Figure 11(a), even though its *production* remains—for the moment—at point *H*.

A variation on this strategy is for foreign nations to provide consumer goods so that the poorer nation can shift its *own* resources out of producing them (and into capital production) without causing consumption levels to fall. Once again, if capital production exceeds point *N* during the year, capital per worker will grow, setting the stage for continual growth to higher standards of living.

GROWTH OPTIONS FOR LDCs

FIGURE 11



Panel (a) shows an LDC producing at point H , where the available consumption goods are just sufficient to meet minimum standards (point S). If the nation can obtain goods externally—through foreign investment or foreign assistance—it can *make use* of capital and consumption goods at a point like F —outside of its PPF.

Panel (b) shows a case where capital production at point T is just sufficient to keep up with a rising population, but not great enough to raise capital per worker and living standards. If this nation can reduce its population growth rate, then the same rate of capital production will increase capital per worker and raise the standard of living.

Finally, there is a fourth alternative. Consider a nation producing at point T in Figure 11(b). Capital production is just sufficient to keep up with a rising population, so the PPF shifts outward each year, but not rapidly enough to raise living standards. If this nation can reduce its population growth rate, however, then less capital production will be needed just to keep up with population growth. In the figure, point N will move downward to N' . If production remains at point T , the PPF will continue to shift outward as before, but now—with slower population growth—productivity and living standards will rise. Slowing the growth in population has been an important (and successful) part of China's growth strategy, although it has required severe restrictions on the rights of individual families to have children. Policy trade-offs, once again.

SUMMARY

The growth rate of real GDP is a key determinant of economic well-being. If output grows faster than the population, then the average standard of living will rise. Output can grow because of increases in employment, increases in capital, and improvements in technology.

Employment will increase if there is an increase in either labor supply or labor demand. Labor supply is determined by the size of the working-age population and by individuals'

willingness to forego leisure in return for a wage. Population growth is something that occurs naturally, but the amount of work effort supplied by a given population is sensitive to after-tax labor earnings. A decrease in the income tax rate would stimulate labor supply.

Labor demand is influenced by productivity. Any factor that makes labor more productive will increase the demand for labor, raise employment, and contribute to economic

growth. If employees become better trained, or if they are given more capital to work with, their productivity will increase.

An increase in the capital stock will shift the production function upward and contribute to economic growth. Whenever investment exceeds depreciation, the capital stock will grow. And if the capital stock grows faster than the labor force, then labor productivity will rise.

Investment can be encouraged by government policies. If the government reduces its budget deficit, the demand for loanable funds will fall, the interest rate will decline, and investment will increase. Investment can also be stimulated directly through reductions in the corporate profits tax or through subsidies to new capital. Finally, policies that encourage household saving can also lower the interest rate and contribute to capital formation.

The third factor that contributes to economic growth is technological change—the application of new inputs or new methods of production. Technological change increases productivity and raises living standards by permitting us to produce more output from a given set of inputs. Technological improvements can be traced back to spending on research and development, either by the government or by private firms.

Economic growth is not costless. Government policies that stimulate employment, capital formation, or technological progress require either tax increases, cuts in other spending programs, or an increase in the national debt. More broadly, any increase in investment requires the sacrifice of consumption today. Any increase in employment from a given population requires a sacrifice of leisure time and other non-market activities.

KEY TERMS

average standard of living
labor productivity
capital per worker

corporate profits tax
investment tax credit
capital gains tax

consumption tax
human capital
technological change

patent protection

REVIEW QUESTIONS

- Discuss the three ways a country can increase its equilibrium level of output.
- Why can population growth be a mixed blessing in terms of economic growth?
- Explain how a tax cut could lead to *slower* economic growth.
- If a country's PPF is shifting outward, is it necessarily the case that the country's standard of living is rising? Why or why not?
- Why did Malthus's dire prediction fail to materialize? Do you think it could still come true? Explain your reasoning.
- "Faster economic growth can benefit everyone and need not harm anyone. That is, there is no policy trade-off when it comes to economic growth." True or false? Explain.
- Explain the following statement: "In some LDCs, it can be said that a significant cause of continued poverty is poverty itself."
- Describe four ways in which LDCs might improve their growth performance. Discuss the opportunity cost that must be borne in each case and identify the group that is most likely to bear it.

PROBLEMS AND EXERCISES

- Discuss the effect (holding everything else constant) each of the following would have on full-employment output, productivity, and the average standard of living. Use the appropriate graphs (e.g., labor market, loanable funds market, production function), and state your assumptions when necessary.
 - Increased immigration
 - An aging of the population with an increasing proportion of retirees
 - A baby boom
 - A decline in the tax rate on corporate profits
 - Reduction of unemployment compensation benefits

- g. Expanding the scope of the federal student loan program
 - h. Easier access to technical information on the Internet
2. Below are GDP and growth data for the United States and four other countries:

	1950 per Capita GDP (in Constant Dollars)	1990 per Capita GDP (in Constant Dollars)	Average Yearly Growth Rate
United States	\$9,573	\$21,558	2.0%
France	\$5,221	\$ 17,959	3.0%
Japan	\$1,873	\$19,425	5.7%
Kenya	\$ 609	\$ 1,055	1.3%
India	\$ 597	\$ 1,348	2.0%

Source: Angus Maddison, *Monitoring the World Economy, 1820–1992*. Paris, OECD, 1995.

- a. For both years, calculate each country’s per capita GDP as a percentage of U.S. per capita GDP. Which countries appear to be catching up to the United States, and which are lagging behind?

- b. If these countries continue to grow at the average growth rates given, how long will it take France to catch up to the United States? How long will it take India? Kenya?
3. Below are data for the country of Barrovia, which has long been concerned with economic growth.

	Population (Millions)	Employment (Millions)	Labor Productivity	Total Output
1997	100	50	\$ 9,500	_____
1998	104	51	\$ 9,500	_____
1999	107	53	\$ 9,750	_____
2000	108	57	\$ 9,750	_____
2001	110	57	\$10,000	_____

- a. Fill in the entries for total output in each of the five years.
- b. Calculate the following for each year (except 1997):
 - (1) Population growth rate (from previous year)
 - (2) Growth rate of output (from previous year)
 - (3) Growth rate of per capita output (from previous year)

C H A L L E N G E Q U E S T I O N S

- 1. Economist Amartya Sen has argued that famines in underdeveloped countries are not simply the result of crop failures or natural disasters. Instead, he suggests that wars, especially civil wars, are linked to most famine episodes in recent history. Using a framework similar to Figure 11, discuss the probable effect of war on a country’s PPF. Explain what would happen if the country were initially operating at or near a point like S, the minimum acceptable level of consumption.
- 2. All else equal, why might someone prefer to invest in physical capital in a less-developed country with a small capital stock than in a more developed country that already has much capital? When wealth holders look for a place to invest and compare prospects in these two types of countries, is all else (other than existing capital stock) really equal? Explain.

E X P E R I E N T I A L E X E R C I S E S

- 1. Technological change is an important drive of economic growth. Refer to the “Technology” column in the Marketplace section of a recent *Wall Street Journal*. Find a story about some technological innovation that seems interesting to you. How will this innovation affect the U.S. production function? Does it seem likely to affect employment as well? If so, which types of workers will benefit, and which will be harmed?

- 2. Investment in computing technology is an oft-cited source of economic growth. To learn more, read Adam Zaretsky’s “Have Computers Made Us More Productive? A Puzzle.” It’s available from the Federal Reserve Bank of St. Louis at <http://www.stls.frb.org/publications/re/1998/d/re19998d3.html>. Based on what you’ve learned, use the model developed in this chapter to show how improvements in information technology will affect the U.S. economy in the long run. Then, make a list of who will benefit and who will be harmed by these changes. How would you expect each group to respond to the changes?





CHAPTER

22

ECONOMIC FLUCTUATIONS

CHAPTER OUTLINE

Can the Classical Model Explain Economic Fluctuations?

- Shifts in Labor Demand
- Shifts in Labor Supply
- Verdict: The Classical Model Cannot Explain Economic Fluctuations

Economic Fluctuations: A More Realistic View

- Opportunity Cost and Labor Supply
- Firms' Benefits from Hiring:
 - The Labor Demand Curve
- The Meaning of Labor Market Equilibrium
- The Labor Market When Output Is Below Potential
- The Labor Market When Output Is Above Potential

What Triggers Economic Fluctuations?

- A Very Simple Economy
- The Real-World Economy
- Shocks That Push the Economy Away from Equilibrium

The Economics of Slow Adjustment

- Adjustment in a Boom
- Adjustment in a Recession
- The Speed of Adjustment

Where Do We Go from Here?

Boom A period of time during which real GDP is above potential GDP.

If you are like most college students, you will be looking for a job when you graduate, or you will already have one and want to keep it for a while. In either case, your fate is not entirely in your own hands. Your job prospects will depend, at least in part, on the overall level of economic activity in the country.

If the classical model of the previous two chapters described the economy at every point in time, you'd have nothing to worry about. Full employment would be achieved automatically, so you could be confident of getting a job at the going wage for someone with your skills and characteristics. Unfortunately, this is not how the world works: Neither output nor employment grows as smoothly and steadily as the classical model predicts. Instead, as far back as we have data, the United States and similar countries have experienced *economic fluctuations*.

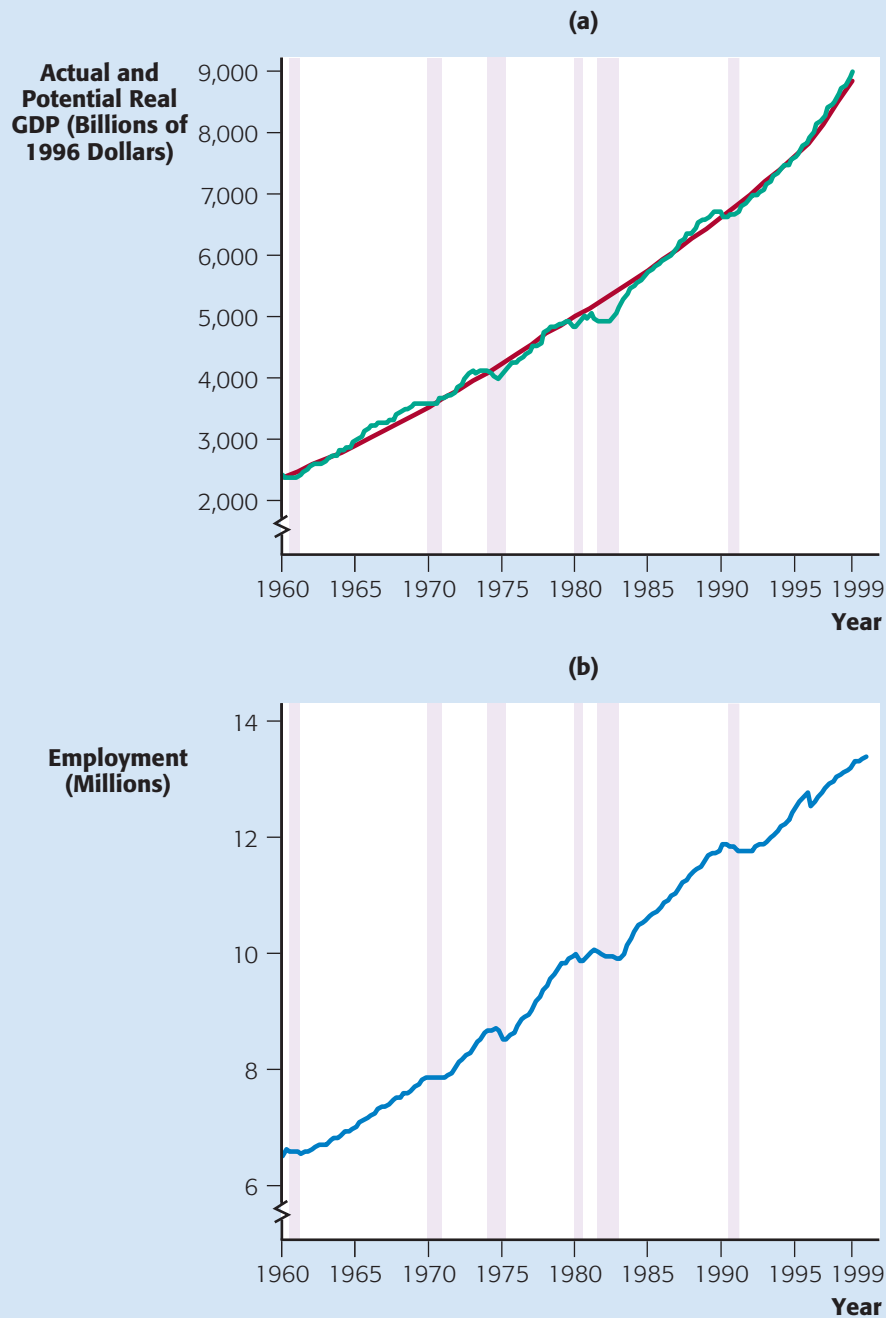
In Figure 1, look first at the red line in panel (a). It shows full-employment or potential output since 1960—the level of real GDP predicted by the classical model. As a result of technological change and growth in the capital stock and population, full-employment output rises steadily. But now look at the blue line, which shows *actual* output. You can see that actual GDP fluctuates above and below the classical model's predictions. During *recessions*, which are shaded in the figure, output declines, occasionally sharply. During *expansions* (the unshaded periods) output rises quickly—usually faster than potential output is rising. Indeed, in the later stages of an expansion, output often *exceeds* potential output—a situation that economists call a **boom**.

Panel (b) shows another characteristic of expansions and contractions: fluctuations in employment. During expansions, such as the period from 1983 to 1990, employment grows rapidly. During recessions (shaded), such as 1990–91, employment declines. Moreover, as we go through a cycle, the causal relationship between output and employment seems to go in the opposite direction to what the classical model predicts. Instead of changes in employment causing changes in output, it seems that—over the business cycle—it is changes in output that cause firms to change their employment levels. For example, in a recession, many business firms lay off workers. If asked why, they would answer that they are reducing employment *because* they are producing less output.

Finally, look at Figure 2, which presents the unemployment rate over the same period as in Figure 1. Figure 2 shows a critical aspect of fluctuations—the bulge of unemployment that occurs during each recession. When GDP falls, the unemploy-

POTENTIAL AND ACTUAL REAL GDP AND EMPLOYMENT, 1960–1999

FIGURE 1



In panel (a), the red line shows full-employment (or potential) real GDP since 1960. It indicates how much output would be produced if the economy were always at full employment. The blue line shows actual real GDP. During recessions (shaded), output declines; during expansions, it rises quickly.

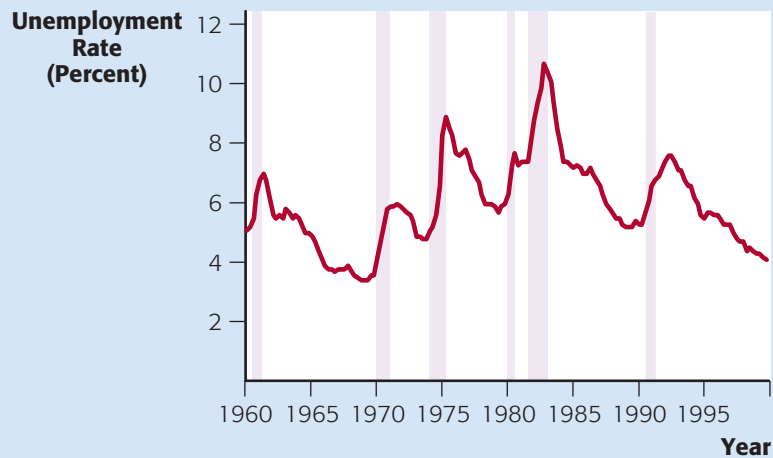
Panel (b) shows how employment fluctuates over the business cycle. During expansions, employment grows rapidly. During recessions, employment declines.

ment rate increases. In the last few decades, the worst bulge in unemployment occurred in 1982, when more than 10 percent of the labor force was looking for work. In expansions, on the other hand, the unemployment rate falls. In our most recent expansion—which is still continuing as this is being written—unemployment dropped to 4 percent. In some expansions, the unemployment rate can drop even lower than the full-employment level. In the sustained expansion of the late 1960s,

FIGURE 2

The unemployment rate—the fraction of the labor force without a job—rises during recessions and falls during expansions.

U.S. UNEMPLOYMENT RATE, 1960–1999



Carl Walsh explores recent economic fluctuations in his "Changes in the Business Cycle," available at <http://www.frbsf.org/econsrch/wklytr/wklytr99/e199-16.html>.

for example, it reached a low of just over 3 percent. At the same time, output exceeded its potential, as you can verify in Figure 1.

Figure 1 also shows something else: Expansions and recessions don't last forever. Indeed, sometimes they are rather brief. The recession of 1990–91, for example, ended within a year.

But if you look carefully at the figure, you'll see that the back-to-back recessions of the early 1980s extended over three full years. And during the Great Depression of the 1930s (not shown), it took more than a decade for the economy to return to full employment. Expansions can last for extended periods, too. The expansion of the 1980s lasted about seven years, from 1983 to 1990. And as this is being written (March 2000), the expansion that began in March 1991 had become the longest expansion in U.S. economic history—already nine years old and still going strong.

If we are to explain economic fluctuations, then, we have three things to explain: (1) *why* they occur in the first place, (2) why they do not last forever, and (3) why they sometimes last so long. Our first step is to see whether the macroeconomic model you've already studied—the classical, long-run model—can explain why economic fluctuations occur.

CAN THE CLASSICAL MODEL EXPLAIN ECONOMIC FLUCTUATIONS?

Can the classical model help us understand the facts of economic fluctuations, as shown in Figures 1 and 2? Or do we need to modify the model to explain them? More specifically, can the classical model explain why GDP and employment typically fall *below* potential during a recession and often rise above it in an expansion? Let's see.

SHIFTS IN LABOR DEMAND

One idea, studied by a number of economists, is that a recession might be caused by a leftward shift of the labor demand curve. This possibility is illustrated in Figure 3, in which a leftward shift in the labor demand curve would move us down and to the

A RECESSION CAUSED BY DECLINING LABOR DEMAND?

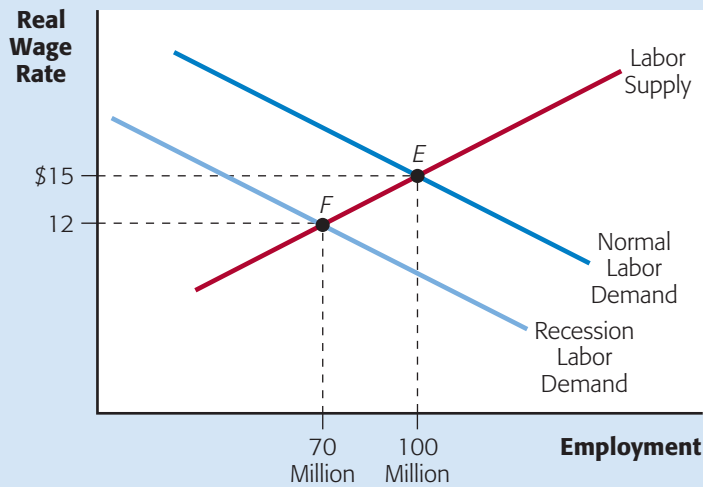


FIGURE 3

If a recession were caused by a leftward shift of the labor demand curve, both employment and the real wage would fall—as in the movement from point *E* to point *F*. In fact, large, sudden shifts in labor demand are an unlikely explanation for real-world fluctuations.

left along the labor supply curve. In the diagram, as the labor market equilibrium moves from point *E* to point *F*, employment falls and so does the real wage rate. Is this a reasonable explanation for recessions? Most economists feel that the answer is no, and for a very good reason.

The labor demand curve tells us the number of workers the nation's firms want to employ at each real wage rate. A leftward shift of this curve would mean that firms want to hire *fewer* workers at any given wage than they wanted to hire before. What could make them come to such a decision? One possibility is that firms are suddenly unable to sell all the output they produce. Therefore, the story would go, they must cut back production and hire fewer workers at any wage.

But as you've learned, total spending is *never* deficient in the classical model. On the contrary, from the classical viewpoint, total spending is automatically equal to whatever level of output firms decide to produce. A decrease in spending by one sector of the economy would cause an equal *increase* in spending by another sector, with no change in total spending. While it is true that a decrease in output and employment could cause total spending to decrease (because Say's law tells us total spending is always equal to total output), the causation cannot go the other way in the classical model. In that model, changes in total spending cannot arise on their own. Therefore, if we want to explain a leftward shift in the labor demand curve using the classical model, we must look for some explanation other than a sudden change in spending.

Another possibility is that the labor demand curve shifts leftward because workers have become less *productive* and therefore less valuable to firms. This might happen if there were a sudden decrease in the capital stock, so that each worker had less equipment to work with. Or it might happen if workers suddenly forgot how to do things—how to operate a computer or use a screwdriver or fix an oil rig. Short of a major war that destroys plant and equipment, or an epidemic of amnesia, it is highly unlikely that workers would become less productive so suddenly. Thus, a leftward shift of the labor demand curve is an unlikely explanation for recessions.

What about booms? Could a *rightward* shift of the labor demand curve (not shown in Figure 3) explain them? Once again, a change in total spending cannot be

the answer. In the classical model, as discussed a few paragraphs ago, changes in spending are caused by changes in employment and output, not the other way around. Nor can we explain a boom by arguing that workers have suddenly become more productive. While it is true that the capital stock grows over time and workers continually gain new skills—and that both of these movements shift the labor demand curve to the right—such shifts take place at a glacial pace. Compared to the amount of machinery already in place, and to the knowledge and skills that the labor force already has, annual increments in physical capital or knowledge are simply too small to have much of an impact on labor demand. Thus, a sudden rightward shift of the labor demand curve is an unlikely explanation for an expansion that pushes us beyond potential output.

Because shifts in the labor demand curve are not very large from year to year, the classical model cannot explain real-world economic fluctuations through shifts in labor demand.

SHIFTS IN LABOR SUPPLY

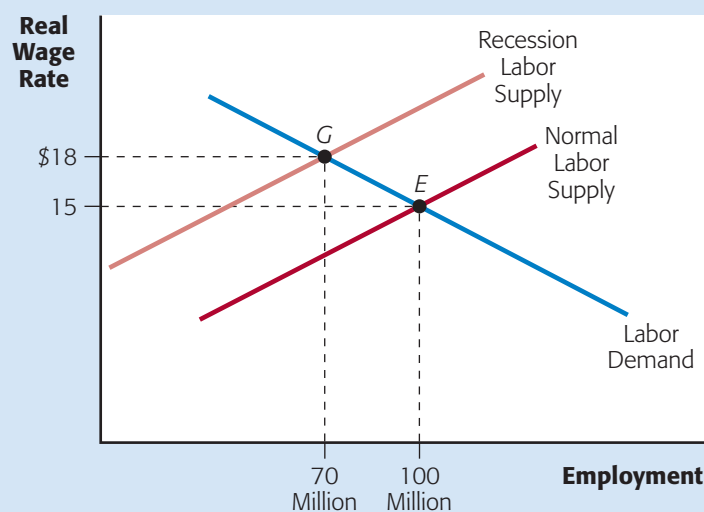
A second way the classical model might explain a recession is through a shift in the labor supply curve. Figure 4 shows how this would work. If the labor supply curve shifts to the left, the equilibrium moves up and to the left along the labor demand curve, from point *E* to point *G*. The level of employment falls, and output falls with it.

This explanation of recessions has almost no support among economists. First, remember that the labor supply schedule tells us, at each real wage rate, the number of people who would like to work. This number reflects millions of families' preferences about working in the market rather than pursuing other activities, such as taking care of children, going to school, or enjoying leisure time. A leftward shift in labor supply would mean that fewer people want to work at any given wage—that preferences have changed toward these other, nonwork activities. But in reality, preferences tend to change very slowly, and certainly not rapidly enough to explain recessions.

FIGURE 4

A RECESSION CAUSED BY DECLINING LABOR SUPPLY?

If a recession were caused by a leftward shift of the labor supply curve, employment would fall, but the real wage would rise—as in the movement from point *E* to point *G*. In fact, shifts in labor supply occur very slowly, so they cannot explain economic fluctuations.



Second, even if such a shift in preferences did occur, it could not explain the facts of real-world downturns. Recessions are times when unusually large numbers of people are looking for work (see Figure 2). It would be hard to square that fact with a shift in preferences away from working.

The same arguments could be made about expansions: To explain them with labor supply shifts, we would have to believe that preferences suddenly change *toward* market work and away from other activities—an unlikely occurrence. And, in any case, expansions are periods when the unemployment rate typically falls to unusually low levels; *fewer*—not more—people are seeking work.

Because sudden shifts of the labor supply curve are unlikely to occur, and because they could not accurately describe the facts of the economic cycle, the classical model cannot explain fluctuations through shifts in the supply of labor.

VERDICT: THE CLASSICAL MODEL CANNOT EXPLAIN ECONOMIC FLUCTUATIONS

In earlier chapters, we stressed that the classical model works well in explaining the movements of the economy in the longer run. Now, we see that it does a rather poor job of explaining the economy in the short run. Why is this? Largely because the classical model involves assumptions about the economy that make sense in the longer run, but not in the short run. Chief among these is the assumption that the labor market clears—that is, that the labor market operates at the point of intersection of the labor supply and labor demand curves. As long as this assumption holds, a boom or recession would have to arise from a sudden, significant *movement* in that intersection point, caused by a sudden and significant *shift* in either the labor demand curve or the labor supply curve.

But now, we've seen that such sudden shifts are very unlikely. Moreover, even if they did occur, they could not explain the changes in job-seeking activity that we observe in real-world recessions. And this, in a nutshell, is why we must reject the classical model when we turn our attention to the short run.

We cannot explain the facts of short-run economic fluctuations with a model in which the labor market always clears. This is why the classical model, which assumes that the market always clears, does a poor job of explaining the economy in the short run.

ECONOMIC FLUCTUATIONS: A MORE REALISTIC VIEW

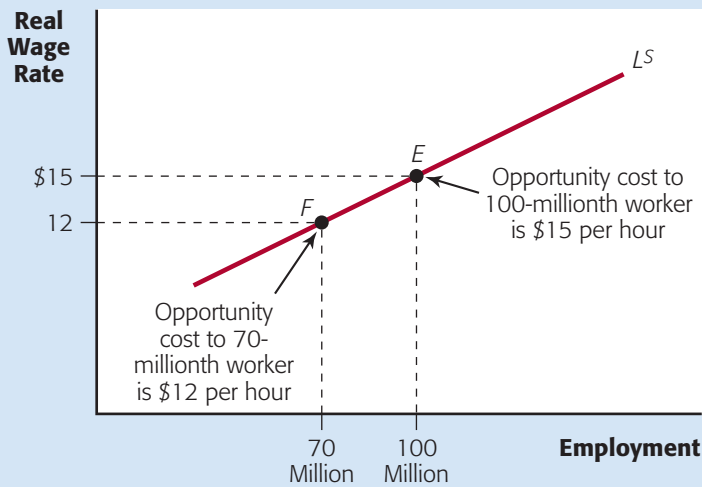
Booms and recessions are two of the most interesting and persistent facets of the economy. Earlier in this book, you learned that recessions can be very costly to society, and in future chapters you'll learn why even booms—the periods during which output exceeds its potential—present serious problems. Yet in spite of determined and often heroic efforts, no economy has been able to eradicate economic fluctuations. In universities and government agencies, economists are conducting research to better understand economic fluctuations. And while there is not yet complete agreement on every feature of them, there is a growing consensus on many aspects. In this and the next several chapters, we will present some of the key ideas behind that consensus. These ideas are based on the concept of **disequilibrium**—the term used to describe a market that does not clear. In particular, we will focus on

Disequilibrium A situation in which a market does not clear—quantity supplied is not equal to quantity demanded.

FIGURE 5

The labor supply curve tells us the wage that must be offered to attract any given number of workers. For example, point *E* indicates that, in order to attract 100 million workers, the real wage must be at least \$15 per hour. This is because the opportunity cost of working for the 100-millionth worker is \$15 per hour.

OPPORTUNITY COST AND LABOR SUPPLY



disequilibrium in the labor market—situations in which the level of employment is *not* where the supply and demand curves intersect.

We begin by taking a closer look at the labor supply and labor demand curves themselves.

OPPORTUNITY COST AND LABOR SUPPLY

So far in this book, we've viewed the labor supply curve as telling us the number of people who want to work at any given real wage. For example, the labor supply curve in Figure 5 tells us that if the wage is \$15 per hour, 100 million people would wish to have jobs (point *E*). But we can also interpret the curve in another way: It tells us the wage that must be offered to attract any given number of workers into the labor market. For example, point *E* shows us that, in order to attract 100 million workers, the real wage must be at least \$15 per hour.

How can we interpret that wage? Each of the 100 million individuals who would work at \$15 would be deciding that it is better to work in the market than to spend time at home or in school. For those workers, \$15 exceeds the *opportunity cost* of working—the value of the other activities sacrificed by going to work.

Now consider the 100-millionth worker—the *last* worker to be attracted into the labor force when the wage is \$15. For this person, the opportunity cost of working must be *exactly* \$15—at any wage greater than \$15, he will choose to work; at any lower wage, his choice will be to stay home or go to school. More generally,

at every point along the labor supply curve, the wage rate tells us the opportunity cost of working for the last worker to enter the labor force.

FIRMS' BENEFITS FROM HIRING: THE LABOR DEMAND CURVE

Now let's take a look at the labor demand curve. We've been viewing this curve as telling us the number of workers firms want to hire at any real wage. But it also tells us the highest wage firms would be willing to pay to hire any given number of

LABOR DEMAND AND THE VALUE OF LABOR

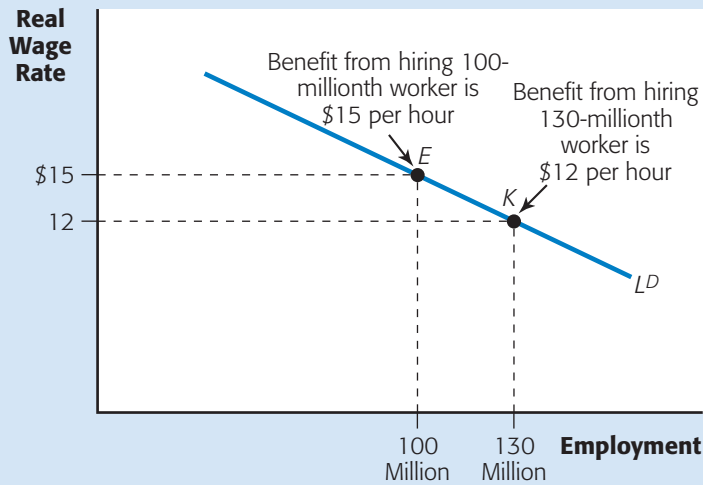


FIGURE 6

The labor demand curve tells us the highest wage firms would be willing to pay to hire any given number of workers. For example, point *E* indicates that, for firms to be willing to hire 100 million workers, the real wage can be no higher than \$15 per hour. This is because the 100-millionth worker benefits the firm by \$15 per hour.

workers. For example, look at the labor demand curve in Figure 6. Point *E* shows that, in order for firms to employ 100 million workers, the wage can be no greater than \$15 per hour. Each of those 100 million workers must benefit firms by \$15 or more per hour, or else he or she would not be hired. And the 100-millionth worker—the one that would be hired at a wage of \$15, but not at any greater wage—must benefit some firm by exactly \$15. In general,

at every point along the labor demand curve, the wage rate tells us the benefit obtained by some firm from the last worker hired.

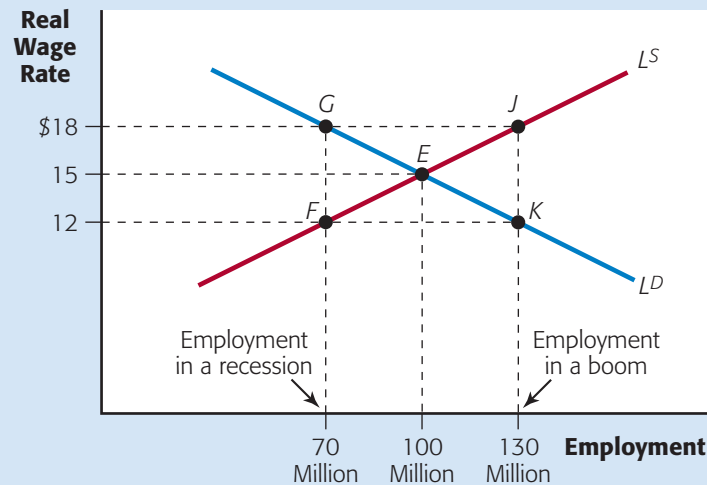
THE MEANING OF LABOR MARKET EQUILIBRIUM

Now look at Figure 7, and for the moment, focus on point *E*, the equilibrium. The idea that the labor market must be at this equilibrium point is essential to the classical model. Point *E* is on both the labor supply and the labor demand curves. This tells us that the opportunity cost for the last worker hired—the 100-millionth—is just equal to the benefit some firm receives from hiring that worker. In ordinary circumstances, as you've learned, the labor market will settle at point *E*, with 100 million people working and a real wage equal to \$15 per hour. Moreover, at this equilibrium point, workers and firms are exploiting all available opportunities for mutual gain. How do we know this?

We can reason as follows: At any employment level less than 100 million, there are workers whose opportunity cost of working is less than the benefit firms would get from hiring them. For example, if employment were 70 million, the opportunity cost of the *next* worker—who would *not* be working—would be just a tiny bit more than \$12 per hour, as shown by point *F*. But some firm could enjoy benefits from that person's work of just under \$18 per hour, as shown by point *G*. At any wage between \$12 and \$18, both parties would gain if that worker were hired. For example, if the firm hired the worker for \$15 per hour, the worker would gain: Her opportunity cost of working is only about \$12, but she would actually get \$15. The

FIGURE 7

LABOR MARKET EQUILIBRIUM



The labor supply and labor demand curves intersect at point E to determine an equilibrium employment of 100 million and an equilibrium real wage of \$15 per hour. At any lower level of employment, such as 70 million, the benefit some firm would enjoy from hiring an additional worker exceeds the opportunity cost to that worker. For example, the 70-millionth worker would benefit some firm by \$18 per hour, but her opportunity cost of working is only \$12 per hour. Mutually beneficial gains are possible for both worker and firm. Only at equilibrium (E) are there no further gains to be exploited.

If employment *exceeds* the equilibrium at point E , firms would be hiring workers whose opportunity cost exceeded firms' benefit from hiring them. For example, the 130-millionth worker would benefit some firm by only \$12 per hour, but her cost of working is \$18 per hour. Employment has increased beyond the level of mutual gain, and there are incentives to reduce it.

firm would also gain, since the benefit of hiring the worker is about \$18, but the firm would pay only \$15. Similar gains would be possible for any increase in hiring, until total employment reached 100 million.

Now suppose that employment has already reached 100 million. What would happen if the *next* worker (the 100,000,001st) were hired? This person would have an opportunity cost a tiny bit more than \$15 per hour, but the benefit from hiring him would be a bit less than \$15 per hour. There is no wage at which this worker could be hired for mutual gain. The same would be true of all other workers beyond 100 million. For example, if employment were to rise all the way to 130 million, the opportunity cost for the 130-millionth worker would be \$18 (point J), but a firm could benefit by only \$12 (point K) from hiring him.

In sum, there is only one level of employment that exhausts all of the mutually beneficial opportunities for trade among workers and firms, and this is where the labor supply and demand curves intersect:

At the equilibrium level of employment, all opportunities for mutually beneficial trade in the labor market have been exploited.

THE LABOR MARKET WHEN OUTPUT IS BELOW POTENTIAL

Now we can see what happens in the labor market during a recession. When employment falls below the classical, full-employment level at point E , both firms and workers could gain if employment increased. But employment *doesn't* increase.

Something in the overall economic system isn't working right, and the opportunities for mutual gain are not exploited as they should be. In Figure 7, for example, employment might fall to 70 million, where the opportunity cost for the next worker would be just \$12, while the benefit to some firm from hiring this worker would be \$18. A mutually beneficial deal between them is certainly possible . . . but it doesn't happen. The labor market is in *disequilibrium*:

During a recession, the labor market is in disequilibrium, and the benefit from hiring another worker exceeds the opportunity cost to that worker.

What *causes* disequilibrium in the labor market? We'll discuss that a bit later in this chapter. But our analysis so far helps us understand why recessions—once they occur—do not last forever. There are strong incentives for the labor market to return to equilibrium, namely, the benefits workers and firms would enjoy from an increase in employment. Until employment returns to the level at which the labor demand and supply curves intersect, these opportunities for mutual gain are not being fully exploited.

In recessions, there are incentives to increase the level of employment because the benefit to firms from additional employment exceeds the opportunity cost to workers. These incentives help explain why recessions do not last forever.

THE LABOR MARKET WHEN OUTPUT IS ABOVE POTENTIAL

What about booms? They are just as temporary as are recessions and recoveries. Once again, Figure 7 shows why. Suppose the economy is experiencing a boom in which 130 million people are working. Then there are workers whose opportunity cost of working exceeds the benefit of their work to firms. The 130-millionth worker, for example, has an opportunity cost of \$18, but his firm benefits by only \$12. No matter what wage we choose, one side of the deal—either the worker or the firm—loses out if that worker is hired. Suppose the 130-millionth worker is paid \$18 in order to convince him to take a job. Then the firm has an incentive to let that worker go. The same is true for every level of employment beyond 100 million workers: If workers are paid their opportunity cost, firms will have an incentive to reduce employment.

In booms, there are incentives to decrease the level of employment because the benefit to firms from some who have been hired is smaller than the opportunity cost to those workers. These incentives help explain why booms do not last forever.

Now you can understand one of the observations about expansions and recessions that we set out to explain: that they do not last forever. But why do they occur in the first place?

WHAT TRIGGERS ECONOMIC FLUCTUATIONS?

Recessions that bring output below potential and expansions that drive output above potential are periods during which the economy is going a bit haywire: Opportunities for mutual gain are not being exploited. But why? In particular, why

does the labor market move away from its equilibrium in the short run? Let's start to answer this question by looking at a world that is much simpler than our own.

A VERY SIMPLE ECONOMY

Imagine an economy with just two people: Yasmin and Pepe. Yasmin is especially good at making popcorn, but she eats only yogurt. Pepe, by contrast, is very good at making yogurt, but eats only popcorn. If things are going well, Yasmin and Pepe will make suitable amounts of popcorn and yogurt and trade with each other. Because of the gains from specialization, their trade will make them both better off than if they tried to function without trading. And under ordinary circumstances, Yasmin and Pepe will take advantage of all mutually beneficial opportunities for trading. Our two-person economy will thus operate at full employment, since both individuals will be fully engaged in making products for the other. You can think of their trading equilibrium as being like the labor market equilibrium in the classical model (such as point *E* in Figure 7), where workers and firms are taking advantage of all mutually beneficial opportunities for hiring and producing.

Now, suppose there is a breakdown in communication. For example, Yasmin may get the impression that Pepe is not going to want as much popcorn as before. She would then decide to make less popcorn for Pepe. At their next trading session, Pepe will be offered less popcorn, so he will decide to produce less yogurt. The result: Total production in the economy declines, and our two traders will lose some of the benefits of trading. This corresponds to a recession.

In reading the previous paragraph, you might be thinking, "Wait a minute. If either Yasmin or Pepe got the impression that the other might want less of the other's product, wouldn't a simple conversation between them straighten things out?" If these are your thoughts, you are absolutely right. A breakdown in communication and a drop in production would be extremely unlikely . . . *in a simple economy with just two people*. And therein lies the problem: The real-world economy is much more complex than the world of Yasmin and Pepe.

THE REAL-WORLD ECONOMY

Think about the U.S. economy, with its millions of businesses producing goods and services for hundreds of millions of people. In many cases, production must be planned long before goods are actually sold. For example, from inception to final production, it takes nearly a year to build a house and two years to develop a new automobile model or produce a Hollywood film. If one firm—say, General Motors—believes that consumers will buy fewer of its cars next year, it cannot simply call a meeting of all potential customers and find out whether its fears are justified. Nor can it convince people, as Yasmin can convince Pepe, that their own jobs depend on their buying a GM car. Most potential car buyers do *not* work for General Motors and don't perceive any connection between buying a car and keeping their own job. Under the circumstances, it may be entirely logical for General Motors to plan for a lower production level and lay off some of its workers.

Of course, this would not be the end of the story. By decreasing its workforce, GM would create further problems for the economy. The workers it has laid off, who will earn less income or none at all, will cut back on *their* spending for a variety of consumer goods—restaurant meals, movies, vacation travel—and they will certainly postpone any large purchases they'd been planning, such as a new large-screen television or that family trip to Disney World. This will cause other firms—the firms producing these consumer goods and services—to cut back on *their* production, laying off *their* workers, and so on. In other words, what began as a

perceived decrease in spending in one sector of the economy can work its way through other sectors, causing a full-blown recession.

This example illustrates a theme that we will revisit in the next chapter: The interdependence between production and income. When people spend their incomes, they give firms the revenue they need to hire workers . . . and pay the workers' income! If any link in this chain is broken, output and income may both decline. In our example, the link was broken because of incorrect expectations by firms in one sector of the economy. But there are other causes of recessions as well, also centering on the interdependence between production and income, and a failure to coordinate the decisions of millions of firms and households.

The classical model, however, waves these potential problems aside. It assumes that workers and firms, with the aid of markets, can work things out—like Yasmin and Pepe—and enjoy the benefits of producing and trading. And the classical model is right: People *will* work things out . . . eventually. But in the short run, we need to look carefully at the problems of coordinating production, trade, and consumption in an economy with hundreds of millions of people and tens of millions of businesses.

A boom can arise in much the same way as a recession. It might start because of an increase in production in one sector of the economy—say, the housing sector. With more production and more workers earning higher incomes, spending increases in other sectors as well, until output rises above the classical, full-employment level.

SHOCKS THAT PUSH THE ECONOMY AWAY FROM EQUILIBRIUM

In our discussion above, General Motors decided to cut back on its production of cars because its managers believed, rightly or wrongly, that the demand for GM cars had decreased. Often, many firms will face a real or predicted drop in spending at the same time. We call this a **spending shock** to the economy—a change in spending that initially affects one or more sectors and ultimately works its way through the entire economy.

In the real world, the economy is constantly buffeted by shocks, and they often cause full-fledged macroeconomic fluctuations. Table 1 lists some of the recessions and expansions of the last 50 years, along with the events and spending shocks that are thought to have caused them, or at least contributed heavily. You can see that each of these shocks first affected spending and output in one or more sectors of the economy. For example, several recessions have been set off by increases in oil prices, which caused a decrease in spending on products that depend on oil and energy, such as automobiles, trucks, and new factory buildings. Other recessions were precipitated by military cutbacks. Still others came about when the Federal Reserve caused sudden increases in interest rates that led to decreased spending on new homes and other goods. (You'll learn about the Federal Reserve and its policies a few chapters from now.) Expansions, on the other hand, have been caused by military buildups, and by falling oil prices that stimulated spending on energy-related products. The expansion of the mid- and late-1990s began when the development of the Internet, and improvements in computers more generally, led to an increase in investment spending. Once the economy began expanding, it was further spurred by other factors, such as a rise in stock prices and consumer optimism, both of which led to an increase in consumption spending.

In addition to these identifiable spending shocks, the economy is buffeted by other shocks whose origins are harder to spot. For example, consumption was higher than expected in the late 1980s, contributing to the rapid expansion that occurred in those years. In the early 1990s, consumption fell back to normal, helping

Spending shock A change in spending that ultimately affects the entire economy.



Katherine Bradbury's "Job Creation and Destruction in Massachusetts" <http://www.bos.frb.org/economic/pdf/neer599c.pdf> provides a case study of how the labor market adjusts.

TABLE 1

EXPANSIONS, RECESSIONS,
AND SHOCKS THAT CAUSED
THEM

Period		Event	Spending Shock
Early 1950s	Expansion	Korean War	Defense Spending ↑
1953	Recession	End of Korean War	Defense Spending ↓
Late 1960s	Expansion	Vietnam War	Defense Spending ↑
1970	Recession	Change in Federal Reserve Policy	Spending on New Homes ↓
1974	Recession	Dramatic Increase in Oil Prices	Spending on Cars and Other Energy-using Products ↓
1980	Recession	Dramatic Increase in Oil Prices	Spending on Cars and Other Energy-using Products ↓
1981–82	Recession	Change in Federal Reserve Policy	Spending on New Homes, Cars and Business Investment ↓
Early 1980s	Expansion	Military Buildup	Defense Spending ↑
Late 1980s	Expansion	Huge Decline in Oil Prices	Spending on Energy-using Products ↑
1990	Recession	Large Increase in Oil Prices; Collapse of Soviet Union	Spending on Cars and Other Energy-using Products ↓ ; Defense Spending ↓
1991–2000	Expansion	Technological Advances in Com- puters; Development of the Internet; High Wealth Creation	Spending on Capital Equipment ↑ ; Consumption ↑



Half of our recessions since the early 1950s have been caused, at least in part, by rapid rises in oil prices.

to cause the recession of that period. There was no obvious event that caused these changes in consumption.

As you can see in Table 1, the economy barely has time to adjust to one shock before it is hit by another. But we can usually see the beginnings of the adjustment process, and sometimes we can follow it through to its end. In the case of an adverse shock, large numbers of workers lose their jobs. The shock puts the labor market into the situation like that depicted a few pages earlier in Figure 7, with employment at 70 million workers. At recession levels of employment, the benefit from working exceeds the opportunity cost of working, providing an incentive for firms to increase their hiring. This incentive guides the economy through a long and gradual period of recovery, during which output and employment rise to their equilibrium levels. Unemployed workers are gradually reabsorbed into the economy until full employment is restored.

But notice the word *gradually*. The process of adjustment back to equilibrium in the labor market can take surprisingly long. This is in sharp contrast to what happens in other markets. In most microeconomic markets, like the one for maple syrup, or macroeconomic markets, like the stock market, there are strong incentives to return to equilibrium, and the response to these incentives is rapid. If quantity supplied does not equal quantity demanded, equilibrium will be restored within hours, days, or weeks. In the labor market, the *incentives* to get to equilibrium are similar to those in other markets, but the process of getting there takes much longer. It can take—and has taken—years for the economy to return to full employment after a recession, as we saw in Figures 1 and 2. For example, the unemployment rate exceeded 10 percent in 1982 and did not fall below 6 percent until 1986.

A positive shock triggers an expansion, and may push the economy into a boom. This puts the labor market in the situation like the one in Figure 7 in which employment rises to 130 million. Again, there is an incentive to return to normal conditions. In this case, cutting the workforce releases workers whose opportunity cost of working is greater than the benefits firms get from their work. As firms re-

spond to these incentives, employment and output will gradually fall back to their full-employment levels. But once again, the process of adjustment back to equilibrium can take years.

Why does it take so long for employment and output to return to normal after a shock?

THE ECONOMICS OF SLOW ADJUSTMENT

To see why the economy does not adjust immediately and fully to a shock, let's take a close look at a representative firm—say, a hotel. Imagine that you manage a hotel with 100 rooms. You would learn, as do most hotel operators, that you do *not* want to fill all 100 rooms night after night. Instead, you do better with some excess capacity—enough vacant rooms to enable some early arrivals to move in when they first show up, to permit some flexibility in case of problems like broken telephones or leaky faucets, and to accommodate the occasional surge in demand without turning away your regular customers and losing their business to another hotel. We'll suppose that you try to manage your operation so that, on an average night, you will fill 70 of the 100 rooms. (In fact, the hotel industry, like the airline industry, tends to operate at around 70 percent of capacity.)

Of course, if you are aiming to fill 70 percent of your rooms, on average, then you will hire the appropriate number of workers to clean the rooms, provide room service, wash dishes and towels, and so on. This is your normal employment level.

ADJUSTMENT IN A BOOM

Now suppose the economy experiences a boom. Output is above its potential, income is high, and so there is an increase in the number of travelers who want to stay at your hotel. As a result, you begin to find that all 100 rooms are filled. What will you do? Eventually, you will take steps to restore normal utilization, such as reducing the amount of advertising or changing your directory listings to show higher prices. But these changes take time, and it would not make sense to make them until you were sure they were necessary. After all, the jump in utilization may not last more than a few weeks. Reversing any changes you make would be costly, and you might regret having made them in haste. For a while, therefore, you will likely hold off making changes. That is, in the short run, you would probably accept unusually high utilization of your hotel.

But what about the additional work that must be done with higher occupancy? For a day or two, you might get your employees to work longer hours and work harder on the job, but you cannot expect them to do so for very long. Soon you will have to hire more workers, even if just temporarily. As we saw earlier, it will take time to bring utilization down to normal levels. In the meantime, your best choice is to increase employment above its normal level. Thus, in the short run, the increase in demand for rooms will lead to higher-than-normal employment.

What is true for your hotel will also be true of other firms in the economy. As they experience the immediate effects of a positive spending shock, they will temporarily operate their factories, stores, or offices at above-normal rates of utilization. As a consequence, they will increase employment to higher-than-normal levels. At these employment levels, the benefits firms get from hiring the additional workers will be smaller than the opportunity cost of their work, but—when all options are considered—this is the sensible thing for firms to do.

When a positive shock causes a boom, firms operate—temporarily—at above-normal rates of utilization. As a consequence, employment rises above its normal, full-employment level.

Now let's go back to your hotel. Suppose that the increase in demand turns out to be long lasting: Month after month, you find yourself filling all 100 rooms. Eventually, you will decide to start making the changes that will restore your normal rate of utilization. You might raise prices, cut back on your advertising, offer fewer frills, or take some combination of steps to get you back to your normal 70-percent occupancy rate.

As your occupancy rate falls back to normal, you will lay off those additional employees you hired, so your level of employment, too, will fall back to normal. Of course, you are not the only firm in the economy behaving this way. Other firms, too, are laying off workers as they bring their businesses back to normal operating ranges. When these adjustments are completed, employment in the nation as a whole will be back at its normal, full-employment level:

Over time, firms that have experienced an increase in demand will return to normal utilization rates, and employment will fall back to its normal, full-employment level.

ADJUSTMENT IN A RECESSION

Now consider a quite different situation. The economy enters a recession, and you begin to find that only 30 of your rooms are rented. Do you take action on the spot to get to your normal 70 guests? Probably not, for two reasons. First, you cannot immediately bring your utilization rate back to normal: Most of the steps you could take to make your hotel more attractive (offer lower prices, more frills, and so on) will benefit the 30 guests who are renting your rooms already, but it takes time for the word to get out and attract *additional* guests. Second, you don't want to change your policies in haste, only to make costly reversals in a few weeks. You will probably wait a while, operating at below-normal capacity for several weeks or even months, meanwhile laying off some of your workers because they are no longer needed. As a result, you—and managers at thousands of other firms—will find yourself laying off some workers whose benefits to you are ordinarily greater than their opportunity cost of working. Yet, considering all of your options, it's a sensible thing to do.

When an adverse shock causes a recession, firms operate—temporarily—at below-normal rates of utilization. As a consequence, employment drops below its normal, full-employment level.

But what if the decrease in demand turns out to be long lasting? After several months, you—and other firms—will realize that it is time to make the changes necessary to bring rates of utilization back up. This might mean lowering prices, offering better amenities, stepping up advertising, and more. As you take these steps, and your occupancy rate rises back to normal, you will hire additional employees, since the benefits of hiring them exceed the opportunity cost of their work. Your employment level will rise back to normal. As other firms behave the same way, employment in the nation will rise back to its normal, full-employment level.

Over time, firms that have experienced a decrease in demand will return to normal utilization rates, and employment will rise back to its normal, full-employment level.

THE SPEED OF ADJUSTMENT

The way we have told our story, it seems that the labor market should adjust fully to a shock—and return to full employment—in a few weeks or months, not the years it often actually takes in the real world. What accounts for the slow adjustment of employment? There is some controversy about this issue, but one of the likely explanations has to do with a realistic view of how jobs are destroyed and created.

Think about what happens in a recession: Workers are laid off, and employment decreases until firms decide to return to normal capacity. But unemployed workers don't necessarily wait around for their original employers to rehire them. Instead, many will look for other jobs, and some will find them. In fact, the rate of new hiring remains high in a recession, suggesting that many of those whose jobs are lost in contracting sectors find jobs in other sectors, even during a recession. This means that when you, as hotel manager, decide to return to normal employment levels, many of those you laid off will have found jobs elsewhere. You will have to search once again for people suitable for hotel work, and you will have to train them. This searching and training is both costly *and time consuming*. We shouldn't be surprised, then, that it can take considerable time—even a few years—for employment to recover fully from a recession.

Job-searching behavior by firms and workers is just one explanation for the slow pace of adjustment back to full employment. In later chapters, we'll carefully examine other explanations that involve the behavior of wages and prices.

WHERE DO WE GO FROM HERE?

The classical model that you've learned in previous chapters is certainly useful: It helps us understand economic growth over time, and how economic events and economic policies affect the economy over the long run. But in trying to understand expansions and recessions—where they come from, and why they last for one or more years—we've had to depart from the strict framework of the classical model. In particular, you've seen that *the labor market will not always clear in the short run*, and you've learned why. As we saw with our hotel example, in order to maintain normal employment at every moment in time—which would add up nationally to the classical, market-clearing employment level—firms would have to adjust more quickly than it makes sense for them to do.

You've also seen how a shock to the economy can affect spending and production in one sector and spread to other sectors, causing a recession or a boom. And you've seen why it can take a year or more to return to full employment after a shock.

One theme of our discussion has been the central role of spending in understanding economic fluctuations. In the classical model, spending could be safely ignored. First, Say's law assured us that total spending would always be sufficient to buy the output produced at full employment. Second, a change in spending—for example, a decrease in military spending by the government—causes other categories of spending to rise by just the right amount to use the resources being freed up by the government. In the long run, we can have faith in the classical perspective on spending.

But in the short run, we've seen that spending shocks to the economy affect production—usually in one specific sector. When employment changes in that sector, the spending of workers *there* will change as well, affecting demand in still other sectors. Clearly, if we want to understand fluctuations, we need to take a close look at spending. This is what we will do in the next chapter, when we study the *short-run macro model*.

S U M M A R Y

The classical model does not always do a good job of describing the economy over short time periods. Over periods of a few years, national economies experience economic fluctuations in which output rises above or falls below its long-term growth path. Periods of rapidly rising output are referred to as expansions, while periods of falling output are called recessions. When real GDP fluctuates, it causes the level of employment and the unemployment rate to fluctuate as well.

The classical model cannot explain economic fluctuations because it assumes that the labor market always clears—that is, it always operates at the point where the labor supply and demand curves intersect. Evidence suggests that this market-clearing assumption is not always valid over short time periods. Instead, the labor market is sometimes characterized by *disequilibrium*, in which employment is above or below the level at which the supply and demand curves intersect.

Whenever the labor market—or any market—is out of equilibrium, there are forces that tend to drive it back to equilibrium. If employment is below equilibrium, then there are opportunities for mutually beneficial deals between employers and unemployed workers. If these deals go through, then employment—and output—will increase. But sometimes it

takes time for these mutually beneficial agreements to be discovered and negotiated. During that time period, the economy can continue to operate below potential. When employment is above equilibrium, firms have incentives to cut back employment, and eventually they will do so. But in the meantime, the economy will experience a boom.

Deviations from the full-employment level of output are often caused by *spending shocks*—changes in spending that initially affect one sector, and then work their way through the entire economy. Negative shocks can cause recessions, while positive shocks can cause expansions that lead to booms. Eventually, output will return to its long-run equilibrium level, but it does not do so immediately. The return to full employment takes time because of the costs of adjusting back to normal output levels, and also because of time-consuming job search by workers and firms. Workers laid off in a recession, for example, will seek work elsewhere—a process that takes time. Similarly, it takes time for employers to find new employees to replace those laid off. The origins of economic fluctuations can be understood more fully with the short-run macro model, which we will study in the next chapter.

K E Y T E R M S

boom

disequilibrium

spending shock

R E V I E W Q U E S T I O N S

1. How does a *recession* differ from an *expansion*? Describe the typical behavior of GDP and the unemployment rate during each of these periods.
2. Why can't a recession be explained in terms of a reduction in labor demand? In terms of a reduction in labor supply?
3. In an economy with just two people, economic fluctuations would be unlikely to occur. Why? What is the key difference in the real-world economy that makes economic fluctuations more likely?
4. "During the last half-century economic fluctuations in the United States have been caused entirely by changes in military spending." True or false? Explain.
5. Suppose the economy is disturbed by a negative spending shock. Describe a typical pattern of adjustment to that shock. What will happen to real GDP and the unemployment rate over time?
6. In what sense are mutual opportunities for gain not being exploited during a recession?

P R O B L E M S A N D E X E R C I S E S

1. Use the following data to construct a labor demand and supply diagram.

Wage Rate	Quantity of Labor Demanded	Quantity of Labor Supplied
\$ 9	95 million	65 million
10	90	70
11	85	75
12	80	80
13	75	85
14	70	90

- a. What are the equilibrium wage rate and level of employment?
- b. Explain the opportunity for mutually beneficial trade that exists if employment is 70 million.
2. Suppose you run a photocopy shop and for one month you experience a surge in business above normal levels. What steps would you take during the month? What additional steps would you take if the surge lasted for two years? How do your answers help explain why booms occur and why they are temporary?



CHAPTER

23

THE SHORT-RUN MACRO MODEL

CHAPTER OUTLINE

Consumption Spending

Consumption and Disposable Income
Consumption and Income
Shifts in the Consumption–Income Line

Getting to Total Spending

Investment Spending
Government Purchases
Net Exports
Summing Up: Aggregate Expenditure
Income and Aggregate Expenditure

Finding Equilibrium GDP

Inventories and Equilibrium GDP
Finding Equilibrium GDP with a Graph
Equilibrium GDP and Employment

What Happens When Things Change?

A Change in Investment Spending
The Expenditure Multiplier
The Multiplier in Reverse
Other Spending Shocks
A Graphical View of the Multiplier
An Important Proviso About the Multiplier

Comparing Models: Long Run and Short Run

The Role of Saving
The Effect of Fiscal Policy

Using the Theory: The Recession of 1990–1991

Appendix 1: Finding Equilibrium GDP Algebraically

Appendix 2: The Special Case of the Tax Multiplier

Every December, newspapers and television news broadcasts focus their attention on spending. You might see a reporter standing in front of a Toys-“R”-Us outlet, warning that unless holiday shoppers loosen their wallets and spend big on toys, computers, vacation trips, dishwashers, and new cars, the economy is in for trouble.

Of course, spending matters during the rest of the year, too. But holiday spending attracts our attention because the normal forces at work during the rest of the year become more concentrated in late November and December. Factories churn out merchandise and stores stock up at higher than normal rates. If consumers are in Scrooge-like moods, unsold goods will pile up in stores. In the months that follow, these stores will cut back on their orders for new goods. As a result, factories will decrease production and lay off workers.

And the story will not end there. The laid-off workers—even those who collect some unemployment benefits—will see their incomes decline. As a consequence, they will spend less on a variety of consumer goods. This will cause other firms—the ones that produce those consumer goods—to cut back on *their* production.

This hypothetical example reinforces a conclusion we reached in the last chapter: Spending is very important in the short run. And it points out an interesting circularity: The more income households have, the more they will spend. That is, *spending depends on income*. But the more households spend, the more output firms will produce—and the more income they will pay to their workers. Thus, *income depends on spending*.

In the short run, spending depends on income, and income depends on spending.

In this chapter, we will explore this circular connection between spending and income. We will do so with a very simple macroeconomic model, which we’ll call the *short-run macro model*. Many of the ideas behind the model were originally developed by the British economist John Maynard Keynes in the 1930s. The **short-run macro model** focuses on the role of spending in explaining economic fluctuations. It explains how shocks that initially affect one sector of the economy quickly influence other sectors, causing changes in total output and employment.

To keep the model as simple as possible, we will—for the time being—ignore all influences on production *besides* spending. As a result, the short-run model may appear strange to you at first, like a drive along an unfamiliar highway. You may wonder: Where is all the scenery you are used to seeing along the classical road? Where are the labor market, the production function, the loanable funds market, and the market-clearing assumption? Rest assured that many of these concepts are still with us, lurking in the background and waiting to be exposed, and we will come back to them in later chapters. But in this chapter, we assume that spending—and *only* spending—determines how much output the economy will produce.

Thinking About Spending. Before we begin our analysis of spending, we have some choices to make.

First, spending on *what*? People spend on food, clothing, furniture, and vacations. They also spend to buy stocks and bonds, to buy homes, to buy used goods, and to buy things produced in foreign countries. In order to know what kind of spending we are going to discuss, we need to use Key Step #1 of our four-step procedure. That is, we need to decide which market's spending to analyze. How should we choose?

Remember our main purpose in building the short-run macro model: to explain fluctuations in real GDP that the long-run, classical model cannot explain. Accordingly, we will ignore spending on things that are *not* part of our GDP, like stocks and bonds and real estate and goods produced abroad. Instead,

in the short-run macro model, we focus on spending in markets for currently produced U.S. goods and services—that is, spending on things that are included in U.S. GDP.

Next, to fully characterize our market, we must also identify the *participants* in that market. We know who the sellers are: U.S. firms. But there are so many different types of buyers of U.S. goods and services: city dwellers and suburbanites; government agencies like the Department of Defense and the local school board; businesses of all types, ranging from the corner convenience store to a huge corporation such as AT&T; and foreigners from nearby Canada and distant Fiji. How should we organize our thinking about all of these different types of buyers?

Macroeconomists have found that the most useful approach is to divide them into four broad categories:

- Households, whose spending is called consumption spending (C)
- Business firms, whose spending is called investment spending (I^P)
- Government agencies, whose spending on goods and services is called government purchases (G)
- Foreigners, whose spending we measure as net exports (NX)

These categories should seem familiar to you. They were the same ones we used to break down GDP in the expenditure approach. In the first part of this chapter, we'll take another look at each of these types of buyers. Then, we'll add their purchases together to explore the behavior of *total* spending in the economy.

Finally, one more choice: Should we look at *nominal* or *real* spending? (Recall that a nominal variable is measured in current dollars, while a real variable is measured in the constant dollars of some base year.) Ultimately, we care more about real variables, such as real output and real income, because they are the more closely related to our economic well-being. For example, a rise in *nominal* output might

Short-run macro model A macroeconomic model that explains how changes in spending can affect real GDP in the short run.



Characterize the Market

mean that we are producing more goods and services, or it might just mean that prices have risen and production has remained the same or fallen. But a rise in *real* output always means that production has increased. For this reason, we will think about real variables right from the beginning. When we discuss “consumption spending,” we mean “real consumption spending,” “investment spending” means “real investment spending,” and so on.

CONSUMPTION SPENDING

Identify Goals and Constraints



A natural place for us to begin our look at spending is with its largest component: *consumption spending*. In all, household spending on consumer goods—groceries, restaurant meals, rent, car repairs, movies, telephone calls, and furniture—is about two-thirds of total spending in the economy. Total consumption spending in the economy is the sum of spending by over a hundred million U.S. households. Each household is trying to achieve the highest level of economic well-being attainable, given the constraints that they face. Because we are interested in the macroeconomy, we don’t concern ourselves with the differences between one consumer good and another. Instead, we want to know: What determines the *total* amount of consumption spending?

The answer is, many different things. Think about yourself: What determines how much you spend in a given year? The most obvious determinant is your income, or—more precisely—your **disposable income**, the part of your income left over after you pay taxes:¹

Disposable income The part of household income that remains after paying taxes.

$$\text{Disposable Income} = \text{Income} - \text{Taxes.}$$

Each of us—in trying to pursue our goal of economic satisfaction—is faced with a constraint: We only have so much disposable income. And when that constraint is relaxed or tightened—when we find ourselves with more or less disposable income—our consumption spending changes. All else equal, you’d certainly spend more on consumer goods with a disposable income of \$50,000 per year than with a disposable income of \$20,000 per year. (Here, as elsewhere, we are speaking about *real* variables: *real* consumption and *real* disposable income.)

But other factors besides your disposable income influence how much you spend. For example, suppose your disposable income is \$50,000 per year. How much of that sum will you spend, and how much will you save? Since the *interest rate* determines your reward for saving, you would probably save more at a higher interest rate like 10 percent than at a lower interest rate like 2 percent. But since you’d be saving more, you’d be spending less. So we can expect consumption spending to be smaller at higher interest rates, and larger at lower interest rates.

Another determinant of consumption is *wealth*—the total value of your assets (home, stocks, bonds, bank accounts, and the like) minus your outstanding liabilities (mortgage loans, credit card debt, student loans, and so on). Even if your disposable income stayed the same, an increase in your wealth—say, because your stocks or bonds became more valuable—would probably induce you to spend more.

¹ Strictly speaking, we deduct *net* taxes from income to obtain disposable income. Net taxes are the taxes households pay *minus* the transfer payments households receive from the government.

Expectations about your future would affect your spending as well. If you become more optimistic about your job security or expect a big raise, you might spend more of your income now. Similarly, increased pessimism, such as greater worries about losing your job, would lead you to decrease spending now.

We could list many other variables that would influence your consumption spending—how long you expect to live, inheritances you expect to receive over your lifetime, and more.

What do these personal observations tell us about *aggregate* consumption spending? Just as your own consumption spending would be influenced by a variety of variables in the economy, so, too, would the consumption spending of other households. Each of the variables we've discussed will therefore influence aggregate consumption spending in predictable ways. We would expect a rise in aggregate disposable income—the total of every household's disposable income in the economy—to cause a rise in aggregate consumption spending. Similarly, a rise in the overall level of interest rates should cause a decrease in aggregate consumption spending.

Figure 1 summarizes some of the important variables that influence consumption spending, and the direction of their effects. A plus sign indicates that consumption spending moves in the same direction as the variable; for example, a rise in disposable income will cause a rise in consumption. A minus sign indicates that the variables are negatively related—a rise in the interest rate will cause consumption spending to fall.

CONSUMPTION AND DISPOSABLE INCOME

Of all the factors that might influence consumption spending, the most important is disposable income. Figure 2 shows the relationship between real consumption spending and real disposable income in the United States from 1960 to 1999. Each point in the diagram represents a different year. For example, the point labeled “1982” represents a disposable income in that year of \$3,773 billion and consumption spending of \$3,260 billion. Notice that as disposable income rises, consumption spending rises as well. Indeed, almost all of the variation in consumption spending from year to year can be explained by variations in disposable income. Although the other factors in Figure 1 do affect consumption spending, their impact appears to be relatively minor.

DETERMINANTS OF CONSUMPTION SPENDING

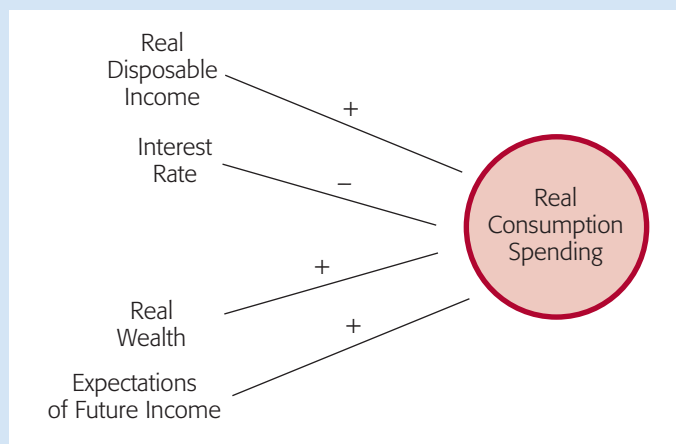
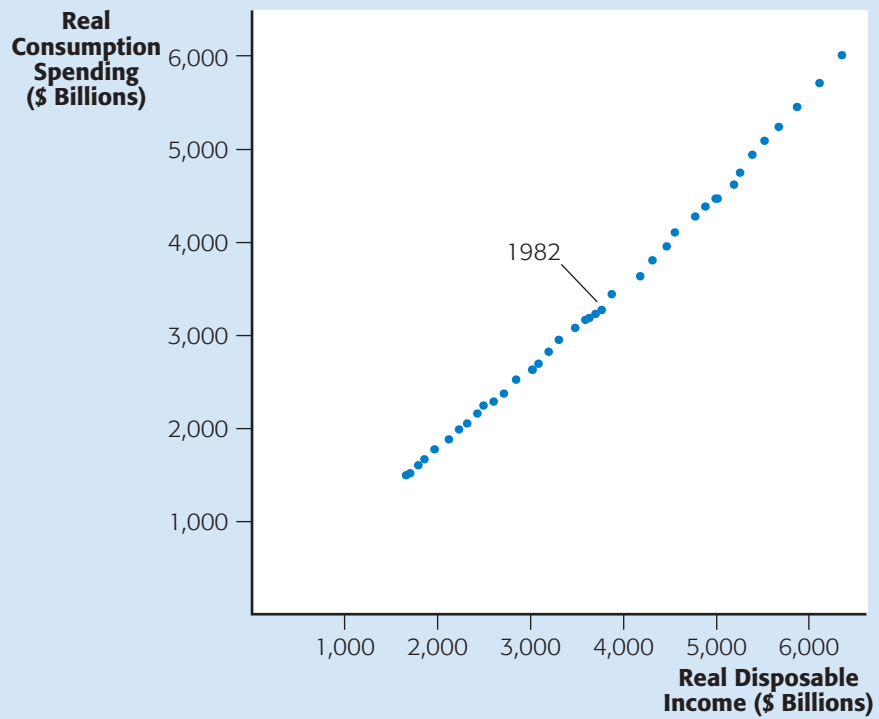


FIGURE 1

FIGURE 2

When real consumption expenditure is plotted against real disposable income, the resulting relationship is almost perfectly linear: As real disposable income rises, so does real consumption spending.

U.S. CONSUMPTION AND DISPOSABLE INCOME, 1960–1999



There is something even more interesting about Figure 2: The relationship between consumption and disposable income is almost perfectly *linear*—the points lie remarkably close to a straight line. This almost-linear relationship between consumption and disposable income has been observed in a wide variety of historical periods and a wide variety of nations. This is why, when we represent the relationship between disposable income and consumption with a diagram or an equation, we use a straight line.

Our discussion will be clearer if we move from the actual data in Figure 2 to the hypothetical example in Table 1. Each row in the table represents a combination of real disposable income and consumption we might observe in an economy. For example, the table shows us that if disposable income were equal to \$7,000 billion in some year, consumption spending would equal \$6,200 billion in that year. When we plot this data on a graph, we obtain the straight line in Figure 3. This line is called the **consumption function**, because it illustrates the functional relationship between consumption and disposable income.

Like every straight line, the consumption function in Figure 3 has two main features: a vertical intercept and a slope. Mathematically, the intercept—in this case, \$2,000 billion—tells us how much consumption spending there would be in the economy if disposable income were zero. However, the real purpose of the vertical intercept is not to identify what would actually happen at zero disposable income, but rather to help us determine which particular line represents consumption spending in the diagram. After all, there are many lines we could draw that have the same slope as the one in the figure. But only one of them has a vertical intercept of \$2,000.

Consumption function A positively sloped relationship between real consumption spending and real disposable income.

TABLE 1

Real Disposable Income (Billions of Dollars per Year)	Real Consumption Spending (Billions of Dollars per Year)
0	2,000
1,000	2,600
2,000	3,200
3,000	3,800
4,000	4,400
5,000	5,000
6,000	5,600
7,000	6,200
8,000	6,800

HYPOTHETICAL DATA ON DISPOSABLE INCOME AND CONSUMPTION

The vertical intercept in the figure also has a name: **autonomous consumption spending**. It represents the combined impact on consumption spending of everything *other than* disposable income. For example, if household wealth were to increase, or the interest rate were to decrease, consumption would be greater at any level of disposable income. The entire consumption function in the figure would shift upward, so its vertical intercept would increase. We would call this *an increase*

Autonomous consumption spending

The part of consumption spending that is independent of income; also, the vertical intercept of the consumption function.

THE CONSUMPTION FUNCTION

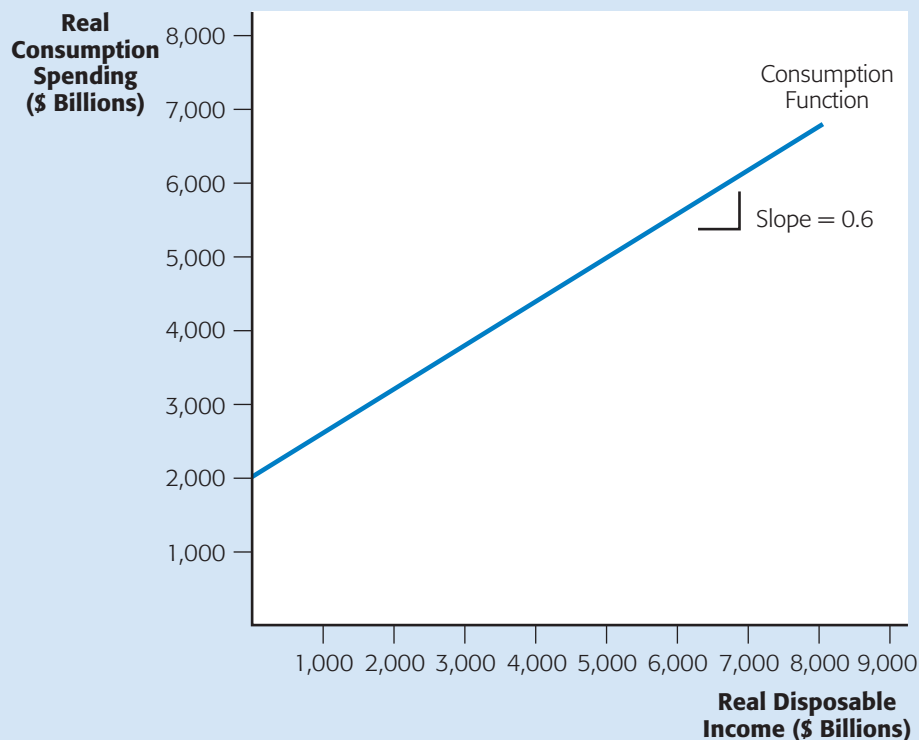


FIGURE 3

Real consumption spending is linearly related to real disposable income. The vertical axis intercept of the line, $a = \$2,000$ billion, shows autonomous consumption expenditure. The slope of the line, $b = 0.6$, is the marginal propensity to consume.

in autonomous consumption spending. Similarly, a decrease in wealth, or a rise in interest rates, would cause a *decrease in autonomous consumption spending*, and shift the consumption function downward.

The second important feature of Figure 3 is the slope, which shows the change along the vertical axis divided by the change along the horizontal axis as we go from one point to another on the line. If we use ΔC to represent the change in real consumption spending, and ΔY_D to represent the change in real disposable income, then the slope of the consumption function is given by

$$\text{slope} = \frac{\Delta C}{\Delta Y_D}.$$

As you can see in the table, each time disposable income rises by \$1,000 billion, consumption spending rises by \$600 billion, so that the slope is $\Delta C/\Delta Y_D = \$600 \text{ billion}/\$1,000 \text{ billion} = 0.6$.

The slope in Figure 3 is an important feature not just of the consumption function itself, but also of the macroeconomic analysis we will build from it. This is why economists have given this slope a special name, the *marginal propensity to consume*, abbreviated *MPC*. In our example, the *MPC* is 0.6.

We can think of the *MPC* in three different ways, but each of them has the same meaning:

Marginal propensity to consume

The amount by which consumption spending rises when disposable income rises by one dollar.

The marginal propensity to consume (MPC) is (1) the slope of the consumption function; (2) the change in consumption divided by the change in disposable income ($\Delta C/\Delta Y_D$); or (3) the amount by which consumption spending rises when disposable income rises by one dollar.

Logic suggests that the *MPC* should be larger than zero (when income rises, consumption spending will rise), but less than 1 (the rise in consumption will be *smaller* than the rise in disposable income). This is certainly true in our example: With an *MPC* of 0.6, each one-dollar rise in disposable income causes spending to rise by 60 cents. It is also observed to be true in economies throughout the world. Accordingly,

we will always assume that $0 < \text{MPC} < 1$.

Representing Consumption with an Equation. Sometimes, we'll want to use an equation to represent the straight-line consumption function. The most general form of the equation is

$$C = a + b Y_D.$$

The term a is the vertical intercept of the consumption function. It represents the theoretical level of consumption spending at $Y_D = 0$, which you've learned is called *autonomous consumption spending*. In the equation, you can see clearly that autonomous consumption (a) is the part of consumption that does *not* depend on disposable income. In our example in Figure 3, a is equal to \$2,000 billion.

The other term, b , is the slope of the consumption function. This is our familiar marginal propensity to consume (*MPC*), telling us how much consumption *increases* each time disposable income rises by a dollar. In our example in Figure 3, b is equal to 0.6.

CONSUMPTION AND INCOME

The consumption function is an important building block of our analysis. Consumption is the largest component of spending, and disposable income is the most important determinant of consumption. But there is one limitation of the line as we've drawn it in Figure 3: It shows us the value of consumption at each level of *disposable* income, whereas we will need to know the value of consumption spending at each level of *income*. Disposable income, you remember, is the income that people have left over after taxes: $Y_D = Y - T$. How can we convert the line in Figure 3 into a relationship between consumption and income?

If the government collected no taxes, total income and disposable income would be equal, so that the relationship between consumption and income on the one hand, and consumption and disposable income on the other hand, would be identical. In that case, the line in Figure 3 would show the relationship between consumption and income. But what about when taxes are not zero?

Table 2 illustrates the consumption–income relationship when households must pay taxes. In the table, we treat taxes as a fixed amount—in this case, \$2,000 billion. Some taxes are, indeed, fixed in this way, such as the taxes assessed on real estate by local governments. Other taxes, like the personal income tax and the sales tax, rise and fall with income in the economy. Treating all taxes as if they are independent of income, as in Table 2, will simplify our discussion without changing our results in any important way.

Notice that the last two columns of the table are identical to the columns in Table 1: In both tables, we assume that the relationship between consumption spending and disposable income is the same. For example, both tables show us that, when disposable income is \$7,000 billion, consumption spending is \$6,200 billion. But in Table 2, we see that a disposable income of \$7,000 is associated with an income of \$9,000. Thus, when income is \$9,000, consumption spending is \$6,200. By comparing the first and last columns of Table 2, we can trace out the relationship between consumption and income. This relationship—which we call the **consumption–income line**—is graphed in Figure 4.

If you compare the consumption–income line in Figure 4 with the line in Figure 3, you will notice that both have the same slope of 0.6, but the consumption–income

Consumption–income line A line showing aggregate consumption spending at each level of income or GDP.

TABLE 2

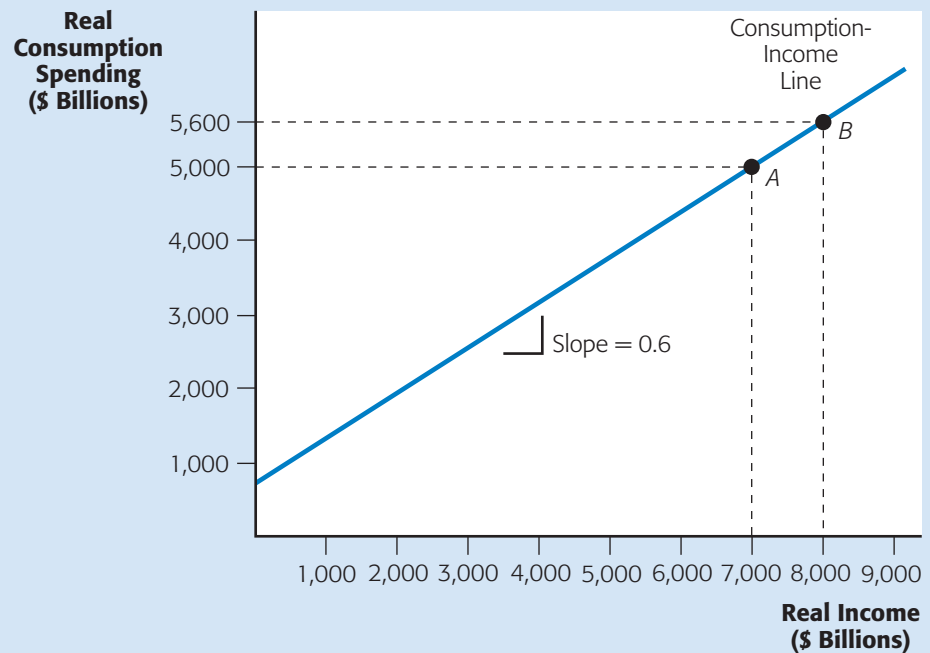
Income or GDP (Billions of Dollars per Year)	Tax Collections (Billions of Dollars per Year)	Disposable Income (Billions of Dollars per Year)	Consumption Spending (Billions of Dollars per Year)
2,000	2,000	0	2,000
3,000	2,000	1,000	2,600
4,000	2,000	2,000	3,200
5,000	2,000	3,000	3,800
6,000	2,000	4,000	4,400
7,000	2,000	5,000	5,000
8,000	2,000	6,000	5,600
9,000	2,000	7,000	6,200
10,000	2,000	8,000	6,800

THE RELATIONSHIP
BETWEEN CONSUMPTION
AND INCOME

FIGURE 4

Real consumption spending is linearly related to real income. The slope of the line, $b = 0.6$, is the marginal propensity to consume.

THE CONSUMPTION–INCOME LINE



line is lower by \$1,200 billion. This raises three important questions. First, why do taxes lower the consumption–income line? Because at any level of income, taxes reduce disposable income and therefore reduce consumption spending.

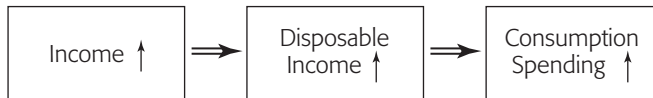
Second, why is the consumption–income line lower by precisely \$1,200 billion? Because any decrease in taxes (T) will cause consumption spending to fall by $MPC \times \Delta T$. In our example, when we impose taxes of \$2,000 billion on the population, disposable income will drop by \$2,000 billion at any level of income. With an MPC of 0.6, consumption at any level of income falls by $0.6 \times \$2,000 \text{ billion} = \$1,200 \text{ billion}$.

Finally, why is the *slope* of the consumption–income line unaffected by taxes? Because when taxes are a fixed amount, disposable income rises dollar-for-dollar with income. With an MPC of 0.6, consumption spending will rise by 60 cents each time income rises by a dollar, just as it would if there were no taxes at all. In other words, while a fixed amount of taxes affects the relationship between the *level* of income and the *level* of consumption spending, it does not affect the relationship between a *change* in income and a *change* in consumption spending. You can verify this in Table 2: Each time income rises by \$1,000 billion, consumption spending rises by \$600 billion, giving a slope of $\Delta C/\Delta Y = \$600 \text{ billion}/\$1,000 \text{ billion} = 0.6$, just as in the case with no taxes. More generally,

when the government collects a fixed amount of taxes from households, the line representing the relationship between consumption and income is shifted downward by the amount of the tax times the marginal propensity to consume (MPC). The slope of this line is unaffected by taxes, and is equal to the MPC.

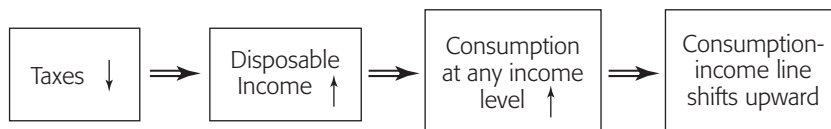
SHIFTS IN THE CONSUMPTION-INCOME LINE

As you've learned, consumption spending depends positively on income: If income increases and taxes remain unchanged, disposable income will rise, and consumption spending will rise along with it. The chain of causation can be represented this way:



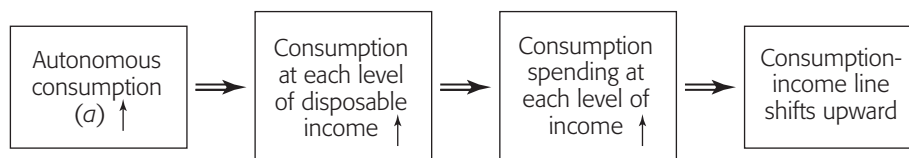
In Figure 4, this change in consumption spending would be represented by a *movement along* the consumption-income line. For example, a rise in income from \$7,000 billion to \$8,000 billion would cause consumption spending to increase from \$5,000 billion to \$5,600 billion, moving us from point *A* to point *B* along the consumption-income line.

But consumption spending can also change for reasons other than a change in income, causing the consumption-income line itself to shift. For example, a decrease in taxes will increase disposable income at each level of income. Consumption spending will then increase at any income level, shifting the entire line upward. The mechanism works like this:



In Figure 5, a decrease in taxes from \$2,000 billion to \$500 billion increases disposable income at each income level by \$1,500 billion, and causes consumption at each income level to increase by $0.6 \times \$1,500 \text{ billion} = \900 billion . This means that the consumption line shifts upward, to the upper line in the figure.

Other changes besides increases or decreases in taxes can shift the consumption-income line as well. All of these other changes work by changing *autonomous consumption*—the vertical intercept of the consumption function in Figure 3. By shifting the relationship between consumption and disposable income, we shift the relationship between consumption and income as well. For example, an increase in household wealth would increase autonomous consumption, and shift the consumption-income line upward, as in Figure 5. Increases in autonomous consumption could also occur if the interest rate decreased, if households developed a taste for spending more of their disposable incomes, or if they became more optimistic about the future. In general, increases in autonomous consumption work this way:



We can summarize our discussion of changes in consumption spending as follows:

When a change in income causes consumption spending to change, we move along the consumption-income line. When a change in anything else besides income causes consumption spending to change, the line will shift.

FIGURE 5

A change in any non-income determinant of consumption spending causes the consumption–income line to shift. A decrease in taxes, for example, increases disposable income and leads to increased consumption spending at any level of income. This is reflected in the upward shift of the consumption–income line. In addition to a tax cut, an increase in autonomous consumption—due to higher wealth, greater optimism, or a lower interest rate—would also lead to an upward shift of the line.

A SHIFT IN THE CONSUMPTION–INCOME LINE

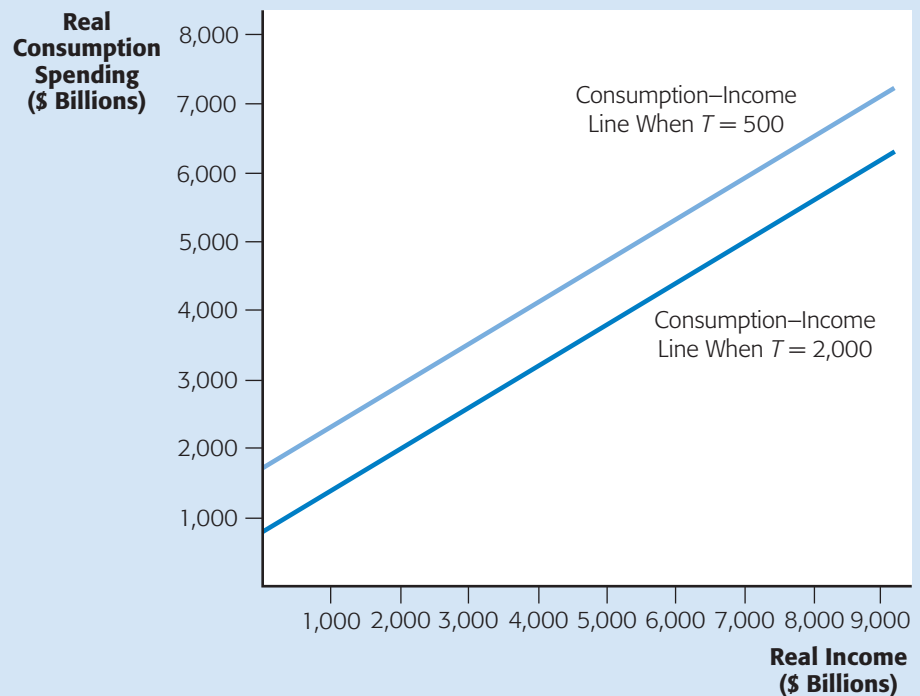


Table 3 provides a more specific summary of how different types of changes in consumption spending are represented with the consumption–income line. Remember that all of the changes that *shift* the line—other than a change in taxes—work by increasing or decreasing autonomous consumption (a).

GETTING TO TOTAL SPENDING

In addition to household consumption spending, there are three other types of spending on goods and services produced by American firms: investment, government purchases, and purchases by foreigners. Let's consider each of these types of spending in turn.

TABLE 3

CHANGES IN CONSUMPTION SPENDING AND THE CONSUMPTION–INCOME LINE

Rightward Movement Along the Line	Leftward Movement Along the Line	Entire Line Shifts Upward	Entire Line Shifts Downward
When	When	When	When
Income \uparrow	Income \downarrow	Taxes \downarrow Household wealth \uparrow Interest rate \downarrow Greater optimism	Taxes \uparrow Household wealth \downarrow Interest rate \uparrow Greater pessimism

INVESTMENT SPENDING

Remember that in the definition of GDP, *investment* (I) consists of three components: (1) business spending on plant and equipment; (2) purchases of new homes; and (3) accumulation of unsold inventories. In this chapter, as we did when we studied the classical model, we focus not on actual investment, but on *planned investment* or *investment spending* (we'll use these two terms interchangeably). Planned investment (I^p) is business purchases of plant and equipment, and construction of new homes.

Why do we focus on planned investment and leave out inventory accumulation? When we look at how spending influences the economy, we are interested in the purchases households, firms, and the government *want* to make. But some inventory changes, as you learned a few chapters ago, are an *unplanned* and *undesired* occurrence that firms try to avoid. While firms want to have *some* inventories on hand, sudden *changes* in inventories are typically not desirable. To keep the model simple, we treat *all* inventory changes as temporary, unplanned occurrences for the firm, and we exclude them when we measure spending in the economy. But even though they are excluded from spending, inventory changes will play an important part in our analysis, as you will see below.

In the short-run macro model, we define investment spending as plant and equipment purchases by business firms, and new home construction. Inventory investment is treated as unintentional and undesired, and is therefore excluded from our definition of investment spending.

What determines the level of investment spending in a given year? In this chapter, we will regard investment spending as a *fixed value*, determined by forces outside of our analysis. This may seem surprising. After all, aren't there variables that affect investment spending in predictable ways? Indeed, there are.

For example, in the classical model, you learned that planned investment is likely to be affected by the interest rate. Indeed, in the real world, the investment–interest rate relationship is quite strong. Investment is also influenced by the general level of optimism or pessimism about the economy and by new technological developments. But if we introduced all of these other variables into our analysis, we would find ourselves working with a very complex framework, and much too soon. In future chapters, we'll explore some of the determinants of investment spending, but in this chapter, to keep things simple, we assume that investment spending is some given amount. We'll explore what happens when that amount changes, but we will not, in this chapter, try to explain what *causes* investment spending to change.

For now, we regard investment spending as a given value, determined by forces outside of our model.

GOVERNMENT PURCHASES

Government purchases include all of the goods and services that government agencies—federal, state, and local—buy during the year. We treat government purchases in the same way as investment spending: as a given value, determined by forces outside of our analysis. Why?

The relationship between government purchases and other macroeconomic variables—particularly income—is rather weak. In recent decades, the biggest changes in government purchases have involved military spending. These changes have been based on world politics, rather than macroeconomic conditions. So when we

assume that government spending is a given value, independent of the other variables in our model, our assumption is actually realistic.

In the short-run macro model, government purchases are treated as a given value, determined by forces outside of the model.

As with investment spending, we'll be exploring what happens when the "given value" of government purchases changes. But we will not try to explain what causes it to change.

NET EXPORTS

If we want to measure total spending on U.S. output, we must also consider the international sector. About 11 percent of U.S.-produced goods are sold to *foreign* consumers, *foreign* businesses, and *foreign* governments. These are U.S. *exports*, and they are as much a part of total spending on U.S. output as the other types of spending we've discussed so far. Thus, exports must be included in our measure of total spending.

But international trade in goods and services also requires us to make an adjustment to the other components of spending. A portion (about 14 percent) of the output bought by *American* consumers, firms, and government agencies was produced abroad. From the U.S. point of view, these are *imports*—spending on foreign, rather than U.S., output. These imports are included in our measures of consumption, investment, and government spending, giving us an exaggerated measure of spending on *American* output. But we can easily correct for this overcount by simply deducting imported consumption goods from our measure of consumption, deducting imported investment goods from our measure of investment, and deducting imported government purchases from our measure of government purchases. Of course, this means we will be deducting total imports from our measure of total spending.

In sum, to incorporate the international sector into our measure of total spending, we must add U.S. exports, and subtract U.S. imports. These two adjustments can be made together by simply including *net exports* (*NX*) as the foreign sector's contribution to total spending.

$$\text{Net Exports} = \text{Total Exports} - \text{Total Imports.}$$

By including net exports, we simultaneously ensure that we have included U.S. output that is sold to foreigners, and excluded consumption, investment, and government spending on output produced abroad.

Net exports can change for a variety of reasons: changes in tastes toward or away from a particular country's goods, changes in the price of foreign currency on world foreign exchange markets, and more. In the final chapter of this book, we'll discuss in more detail how and why net exports change. But in this chapter, to keep things simple, we assume that net exports—like investment spending and government purchases—are some given amount. We'll explore what happens when that amount changes, but we will not, in this chapter, try to explain what causes net exports to change.

For now, we regard net exports as a given value, determined by forces outside of our analysis.

It's important to remember that net exports can be *negative*, and—in the United States—they have been negative since 1982. Negative net exports means that our imports are greater than our exports. Or, equivalently, Americans are buying more foreign goods and services than foreigners are buying of ours. In that case, net exports contribute *negatively* to total spending on U.S. output.

SUMMING UP: AGGREGATE EXPENDITURE

Now that we've discussed all of the components of spending in the economy, we can be more precise about measuring total spending. First, we'll use the phrase *aggregate expenditure* to mean total spending on U.S. output over some period of time. More formally,

Aggregate expenditure is the sum of spending by households, businesses, the government, and the foreign sector on final goods and services produced in the United States.

Remembering that C stands for household consumption spending, I^p for investment spending, G for government purchases, and NX for net exports, we have

$$\text{Aggregate expenditure} = C + I^p + G + NX.$$

Aggregate expenditure spending plays a key role in explaining economic fluctuations. Why? Because over several quarters or even a few years, business firms tend to respond to changes in aggregate expenditure by changing their level of output. That is, a rise in aggregate expenditure leads firms throughout the economy to raise their output level, while a drop in aggregate expenditure causes a decrease in output throughout the economy. While these changes are temporary, they persist long enough to create the kinds of economic fluctuations that you saw in the previous chapter's Figures 1 and 2. In the next section, we'll explore just how changes in spending create these economic fluctuations.

INCOME AND AGGREGATE EXPENDITURE

As we discussed earlier, the relationship between income and spending is circular: Spending depends on income, and income depends on spending. In Table 4, we take up the first part of that circle: how total spending depends on income. In the table, column 1 lists some possible income levels, and column 2 shows the level of consumption spending we can expect at each income level. These two columns are just the consumption–income relationship we introduced earlier, in Table 2.

Column 3 shows that business firms in this economy buy \$700 billion per year in plant and equipment, regardless of the level of income. Government purchases are also fixed in value, as shown by column 4: At every level of income, the government buys \$500 billion in goods and services. And net



The definition of aggregate expenditure looks very similar to the definition of GDP presented in the chapter entitled “Production, Income, and Employment.” Does this mean that aggregate expenditure and total output are always the same number? Not at all. There is a slight—but important—difference in the definitions. GDP is defined as $C + I + G + NX$. Aggregate expenditure, by contrast, is defined as $C + I^p + G + NX$. The difference is that GDP adds actual investment (I), which includes business firms' inventory investment. Aggregate expenditure adds just planned investment (I^p), which *excludes* inventory investment. The two numbers will not be equal unless inventory investment is zero. (And we'll use this fact to help us find the equilibrium GDP in the next section.)

Aggregate expenditure (AE) The sum of spending by households, business firms, the government, and foreigners on final goods and services produced in the United States.

TABLE 4

THE RELATIONSHIP BETWEEN INCOME AND AGGREGATE EXPENDITURE

(1) Income or GDP (Billions of Dollars per Year)	(2) Consumption Spending (Billions of Dollars per Year)	(3) Investment Spending (Billions of Dollars per Year)	(4) Government Purchases (Billions of Dollars per Year)	(5) Net Exports (Billions of Dollars per Year)	(6) Aggregate Expenditure (AE) (Billions of Dollars per Year)	(7) Change in Inventories (Billions of Dollars per Year)
2,000	2,000	700	500	400	3,600	-1,600
3,000	2,600	700	500	400	4,200	-1,200
4,000	3,200	700	500	400	4,800	-800
5,000	3,800	700	500	400	5,400	-400
6,000	4,400	700	500	400	6,000	0
7,000	5,000	700	500	400	6,600	400
8,000	5,600	700	500	400	7,200	800
9,000	6,200	700	500	400	7,800	1,200
10,000	6,800	700	500	400	8,400	1,600

exports, in column 5, are assumed to be \$400 billion at each level of income. Finally, if we add together the entries in columns 2, 3, and 4, we get $C + I^p + G + NX$, or aggregate expenditure, shown in column 6. (For now, ignore column 7.)

Notice that aggregate expenditure increases as income rises. But notice also that the rise in aggregate expenditure is *smaller* than the rise in income. For example, you can see that when income rises from \$5,000 billion to \$6,000 billion (column 1), aggregate expenditure rises from \$5,400 billion to \$6,000 billion (column 6). Thus, a \$1,000 billion increase in income is associated with a \$600 billion increase in aggregate expenditure. This is because, in our analysis, consumption is the only component of spending that depends on income, and consumption spending always increases according to the marginal propensity to consume, here equal to 0.6. More generally,

when income increases, aggregate expenditure (AE) will rise by the MPC times the change in income: $\Delta AE = MPC \times \Delta Y$.

Find the Equilibrium



FINDING EQUILIBRIUM GDP

Table 4 shows how spending depends on income. In this section, you will see how income depends on spending—that is, how the spending behavior of households, firms, and government agencies determines the economy’s *equilibrium income* or *equilibrium GDP*—a level of GDP that represents, at least in the short run, a point of rest for the economy. That is, we are about to use Key Step #3 of our four-step procedure. As always, the equilibrium will be a point of rest of the economy: a value for GDP that remains the same until something we’ve been assuming constant begins to change. That part of Key Step #3 will be familiar to you.

However, be forewarned: Our method of *finding* equilibrium in the short run is very different from anything you’ve seen before in this text.

Our starting point in finding the economy’s short-run equilibrium is to ask ourselves what would happen, hypothetically, if the economy were operating at differ-

ent levels of output. Let's start with a GDP of \$9,000 billion. Could this be the equilibrium GDP we seek? That is, if firms were producing this level of output, would they keep doing so? Let's see.

Table 4 tells us that when GDP is equal to \$9,000 billion, aggregate expenditure is equal to \$7,800 billion. Business firms are *producing* \$1,200 billion more than they

are *selling*. Since firms will certainly not be willing to continue producing output they cannot sell, we can infer that, in future periods, they will slow their production. Thus, if the economy finds itself at a GDP of \$9,000 billion, it will not stay there. In other words, \$9,000 billion is *not* where the economy will settle in the short run, so it is *not* our equilibrium GDP. More generally,

when aggregate expenditure is less than GDP, output will decline in the future. Thus, any level of output at which aggregate expenditure is less than GDP cannot be the equilibrium GDP.

Now let's consider the opposite case: a level of GDP of \$3,000 billion. At this level of output, the table shows aggregate expenditure of \$4,200 billion—spending is actually *greater* than output by \$1,200 billion. What will business firms do in response? Since they are selling more output than they are currently producing, we can expect them to *increase* their production in future months. Thus, if GDP is \$3,000 billion, it will tend to rise in the future. So \$3,000 is *not* our equilibrium GDP.

When aggregate expenditure is greater than GDP, output will rise in the future. Thus, any level of output at which aggregate expenditure exceeds GDP cannot be the equilibrium GDP.

Now consider a GDP of \$6,000 billion. At this level of output, our table shows that aggregate expenditure is precisely equal to \$6,000 billion: Output and aggregate expenditure are equal. Since firms, on the whole, are selling just what they produce—no more and no less—they should be content to produce that same amount in the future. We have found our equilibrium GDP:

In the short run, equilibrium GDP is the level of output at which output and aggregate expenditure are equal.

Equilibrium GDP In the short run, the level of output at which output and aggregate expenditure are equal.

INVENTORIES AND EQUILIBRIUM GDP

When firms *produce* more goods than they sell, what happens to the unsold output? It is added to their inventory stocks. When firms *sell* more goods than they produce, where do the additional goods come from? They come from firms' inventory stocks. You can see that the gap between output and spending determines what will happen to inventories during the year.

More specifically,



You may be wondering why, in the short-run macro model, a firm that produces more output than it sells wouldn't just lower the price of its goods.

That way, it could sell more of them, and not have to lower its output as much. Similarly, a firm whose sales exceeded its production could take advantage of the opportunity to raise its prices, which would result in lower sales.

To some extent, firms *do* change prices—even in the short run. But they change their output levels, too. To remain as simple as possible, the short-run macro model assumes that firms adjust *only* their output to match aggregate expenditure. That is, *in the short-run macro model, prices don't change at all*. In a later chapter, we'll make the more realistic assumption that firms adjust both prices and output.

the change in inventories during any period will always equal output minus aggregate expenditure.

For example, Table 4 tells us that if GDP is equal to \$9,000 billion, aggregate expenditure is equal to \$7,800 billion. In this case, we can find that the change in inventories is

$$\begin{aligned}\Delta\text{Inventories} &= \text{GDP} - \text{AE} \\ &= \$9,000 \text{ billion} - \$7,800 \text{ billion} = \$1,200 \text{ billion.}\end{aligned}$$

When GDP is equal to \$3,000 billion, aggregate expenditure is equal to \$4,200 billion, so that the change in inventories is

$$\begin{aligned}\Delta\text{Inventories} &= \text{GDP} - \text{AE} \\ &= \$3,000 \text{ billion} - \$4,800 \text{ billion} = -\$1,200 \text{ billion.}\end{aligned}$$

Notice the negative sign in front of the \$1,200 billion; if output is \$3,000 billion, then inventory stocks will *shrink* by \$1,200 billion.

Only when output and total sales are equal—that is, when GDP is at its equilibrium value—will the change in inventories be zero. In our example, when GDP is at its equilibrium value of \$6,000 billion, so that aggregate expenditure is also \$6,000 billion, the change in inventories is equal to zero. At this output level, we have

$$\begin{aligned}\Delta\text{Inventories} &= \text{GDP} - \text{AE} \\ &= \$6,000 \text{ billion} - \$6,000 \text{ billion} = \$0.\end{aligned}$$

What you have just learned about inventories suggests another way to find the equilibrium GDP in the economy: Find the output level at which the change in inventories is equal to zero. Firms cannot allow their inventories of unsold goods to keep growing for very long (they would go out of business), nor can they continue to sell goods out of inventory for very long (they would run out of goods). Instead, they will desire to keep their production in line with their sales, so that their inventories do not change.

To recap,

$$\text{AE} < \text{GDP} \implies \Delta\text{Inventories} > 0 \implies \text{GDP} \downarrow \text{ in future periods.}$$

$$\text{AE} > \text{GDP} \implies \Delta\text{Inventories} < 0 \implies \text{GDP} \uparrow \text{ in future periods.}$$

$$\text{AE} = \text{GDP} \implies \Delta\text{Inventories} = 0 \implies \text{No change in GDP.}$$

Now look at the last column in Table 4, which lists the change in inventories at different levels of output. This column is obtained by subtracting column 6 from column 1. The equilibrium output level is the one at which the change in inventories equals zero, which, as we've already found, is \$6,000 billion.

FINDING EQUILIBRIUM GDP WITH A GRAPH

To get an even clearer picture of how equilibrium GDP is determined, we'll illustrate it with a graph, although it will take us a few steps to get there. Figure 6 begins the process by showing how we can construct a graph of aggregate expenditure. The lowest line in the figure, labeled C, is our familiar consumption-income line, obtained from the data in the first two columns of Table 4.

DERIVING THE AGGREGATE EXPENDITURE LINE

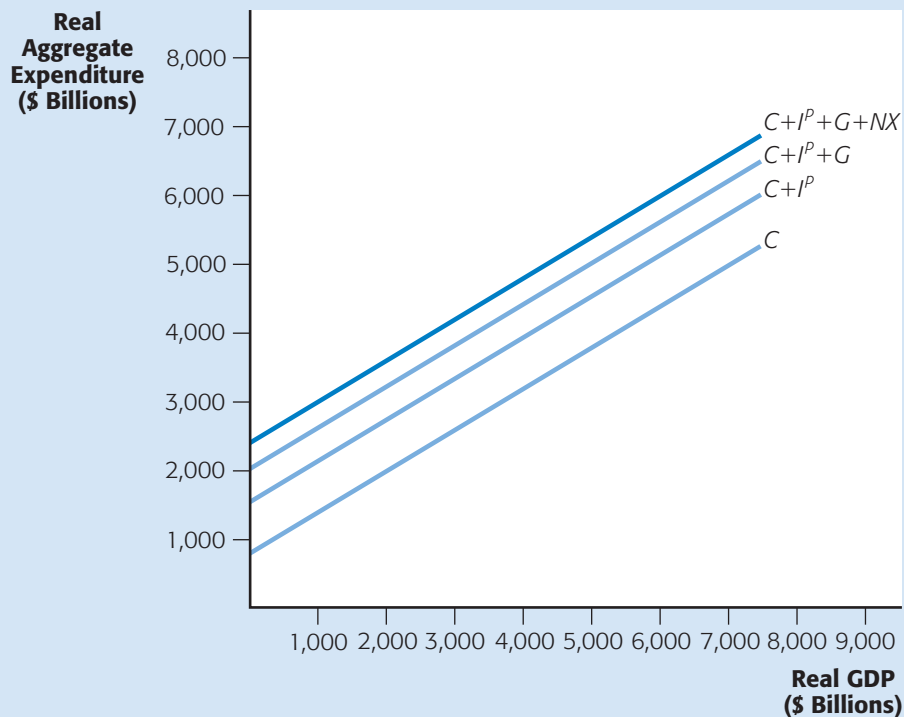


FIGURE 6

Aggregate expenditure is the total of consumption, investment, government purchases, and net exports at a given level of real income. The aggregate expenditure line is derived by adding fixed amounts of investment, government purchases, and net exports to consumption, as determined by the consumption–income line. The slope of the aggregate expenditure line is the marginal propensity to consume.

The next line, labeled $C + I^p$, shows the *sum* of consumption and investment spending at each income level. Notice that this line is parallel to the C line, which means that the vertical distance between them—\$700 billion—is the same at any income level. This vertical difference is investment spending, which remains the same at all income levels.

The next line adds government purchases to consumption and investment spending, giving us $C + I^p + G$. The $C + I^p + G$ line is parallel to the $C + I^p$ line. The vertical distance between them—\$500 billion—represents government purchases. Like investment, government purchases are the same at all income levels.

Finally, the top line adds net exports, giving us $C + I^p + G + NX$, or aggregate expenditure. The distance between the $C + I^p + G + NX$ line and the $C + I^p + G$ line—\$400 billion—represents net exports, which are assumed to be the same at any level of income.

Now look just at the aggregate expenditure line—the top line—in Figure 6. Notice that it slopes upward, telling us that as income increases, so does aggregate expenditure. And the slope of the aggregate expenditure line is less than 1: When income increases, the rise in aggregate expenditure is *smaller* than the rise in income. In fact, the slope of the aggregate expenditure line is equal to the *MPC*, or 0.6 in this example. This tells us that a one-dollar rise in income causes a 60-cent increase in aggregate expenditure. (Question: Which of the four components of aggregate expenditure rises when income rises? Which remain the same?)

Now we're almost ready to use a graph like the one in Figure 6 to locate equilibrium GDP, but first we must develop a little geometric trick.

FIGURE 7

When both axes are measured in the same units, the 45° line can be used to show all points at which the value measured on the horizontal axis equals the value measured on the vertical axis. In the figure, the distances OC , OB , and BA are all equal.

THE 45° LINE

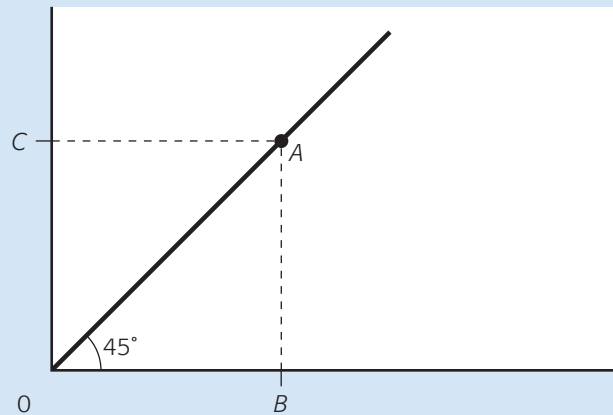


Figure 7 shows a graph in which the horizontal and vertical axes are both measured in the same units, such as dollars. It also shows a line drawn at a 45° angle that begins at the origin. This 45° line has a useful property: Any point along it represents the same value along the vertical axis as it does along the horizontal axis. For example, look at point A on the line. Point A corresponds to the horizontal distance OB , and it also corresponds to the vertical distance OC . But because the line is a 45° line, we know that these two distances are equal: $OB = OC$. Moreover, a glance at the figure shows that that OB and BA are equal as well. Now we have two choices for measuring the distance OB : We can measure it horizontally, or we can measure it as the vertical distance BA . In fact, *any* horizontal distance can also be read vertically, merely by going from the horizontal value (point B in our example) up to the 45° line.

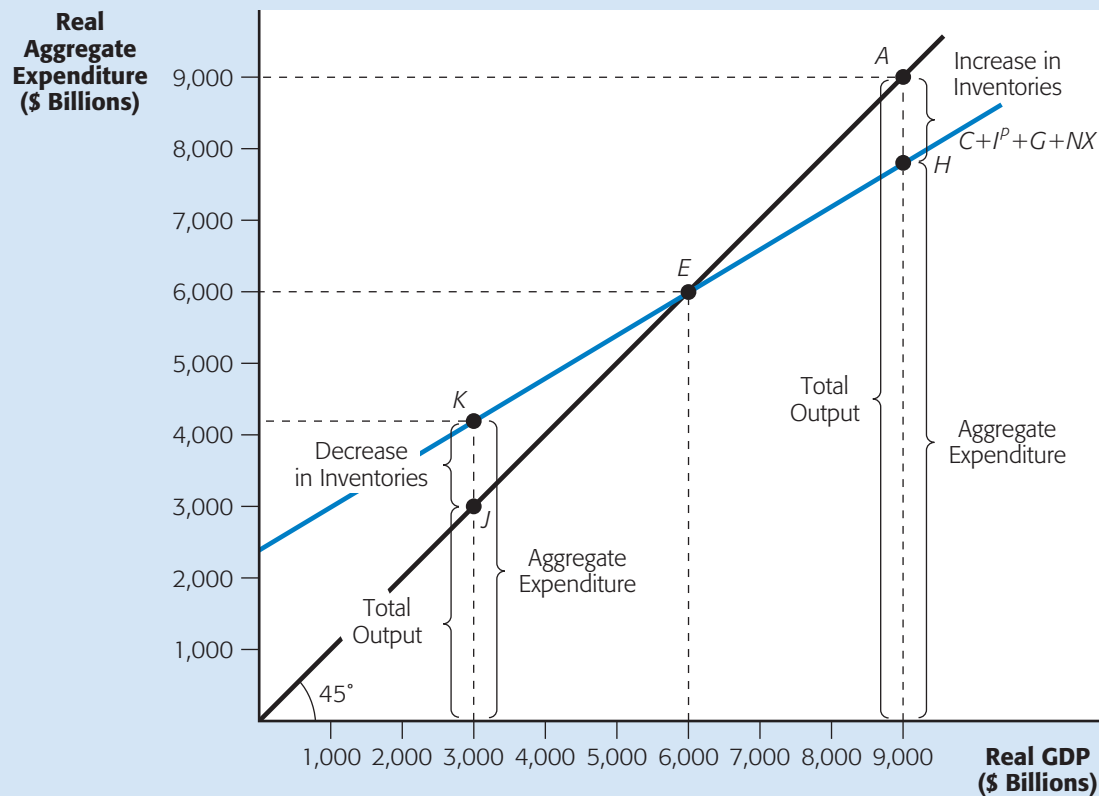
A 45° line is a translator line: It allows us to measure any horizontal distance as a vertical distance instead.

Now we can apply this geometric trick to help us find the equilibrium GDP. In our aggregate expenditure diagram, we want to compare output with aggregate expenditure. But output is measured horizontally, while aggregate expenditure is measured vertically. Our 45° line, however, enables us to measure output vertically as well as horizontally, and thus permits us to compare two vertical distances.

Figure 8 shows how this is done. The solid line is the aggregate expenditure line ($C + I^p + G + NX$) from Figure 6. We've dispensed with the other three lines that were drawn in Figure 7 because we no longer need them. The black line is our 45° translator line. Now, let's search for the equilibrium GDP by considering a number of possibilities. For example, could the output level \$9,000 billion be our sought-after equilibrium? Let's see. We can measure the output level \$9,000 billion as the vertical distance from the horizontal axis up to point A on the 45° line. But when output is \$9,000 billion, aggregate expenditure is the vertical distance from the horizontal axis to point H on the aggregate expenditure line. Notice that, since point H lies below point A , aggregate expenditure is less than output. If firms *did* produce \$9,000 billion worth of output, they would accumulate inventories equal to the vertical distance HA (the excess of output over spending). We conclude graphically (as

DETERMINING EQUILIBRIUM REAL GDP

FIGURE 8



At point E , where the aggregate expenditure line crosses the 45° line, the economy is in short-run equilibrium. With real GDP equal to \$6,000 billion, aggregate expenditure equals real GDP. At higher levels of real GDP—such as \$9,000 billion—total production exceeds aggregate expenditures. At point A , firms will be unable to sell all they produce. Unplanned inventory increases equal to HA will lead them to reduce production. At lower levels of real GDP—such as \$3,000 billion—aggregate expenditure exceeds total production. Firms find their inventories falling, and they will respond by increasing production.

we did earlier, using our table) that if output is \$9,000 billion, firms will accumulate inventories of unsold goods and reduce output in the future. Thus, \$9,000 billion is not our equilibrium. In general,

at any output level at which the aggregate expenditure line lies below the 45° line, aggregate expenditure is less than GDP. If firms produce any of these output levels, their inventories will grow, and they will reduce output in the future.

Now let's see if an output of \$3,000 billion could be our equilibrium. First, we read this output level as the vertical distance up to point J on the 45° line. Next, we note that when output is \$3,000 billion, aggregate expenditure is the vertical distance up to point K on the aggregate expenditure line. Point K lies *above* point J , so aggregate expenditure is greater than output. If firms *did* produce \$3,000 billion in output, inventories would *decrease* by the vertical distance JK . With declining inventories, firms would want to increase their output in the future, so \$3,000 billion is not our equilibrium. More generally,

at any output level at which the aggregate expenditure line lies above the 45° line, aggregate expenditure exceeds GDP. If firms produce any of these output levels, their inventories will decline, and they will increase their output in the future.

Finally, consider an output of \$6,000 billion. At this output level, the aggregate expenditure line and the 45° line cross. As a result, the vertical distance up to point E on the 45° line (representing output) is the same as the vertical distance up to point E on the aggregate expenditure line. If firms produce an output level of \$6,000 billion, aggregate expenditure and output will be precisely equal, inventories will remain unchanged, and firms will have no incentive to increase or decrease output in the future. We have thus found our equilibrium on the graph: \$6,000 billion.

Equilibrium GDP is the output level at which the aggregate expenditure line intersects the 45° line. If firms produce this output level, their inventories will not change, and they will be content to continue producing the same level of output in the future.



During the Great Depression of the 1930s, the economy's short-run equilibrium output fell far below potential, and at least a quarter of the labor force became unemployed.

EQUILIBRIUM GDP AND EMPLOYMENT

Now that you've learned how to find the economy's equilibrium GDP in the short run, a question may have occurred to you: When the economy operates at equilibrium, will it also be operating at full employment? The answer is: *not necessarily*. Let's see why.

If you look back over the two methods we've employed to find equilibrium GDP—using columns of numbers as in Table 4, and using a graph as in Figure 8—you will see that in both cases we've asked only one question: How much will households, businesses, the government, and foreigners *spend* on goods produced in the United States? We did not ask any questions about the number of people who want to work. Therefore, it would be quite a coincidence if our equilibrium GDP happened to be the output level at which the entire labor force were employed.

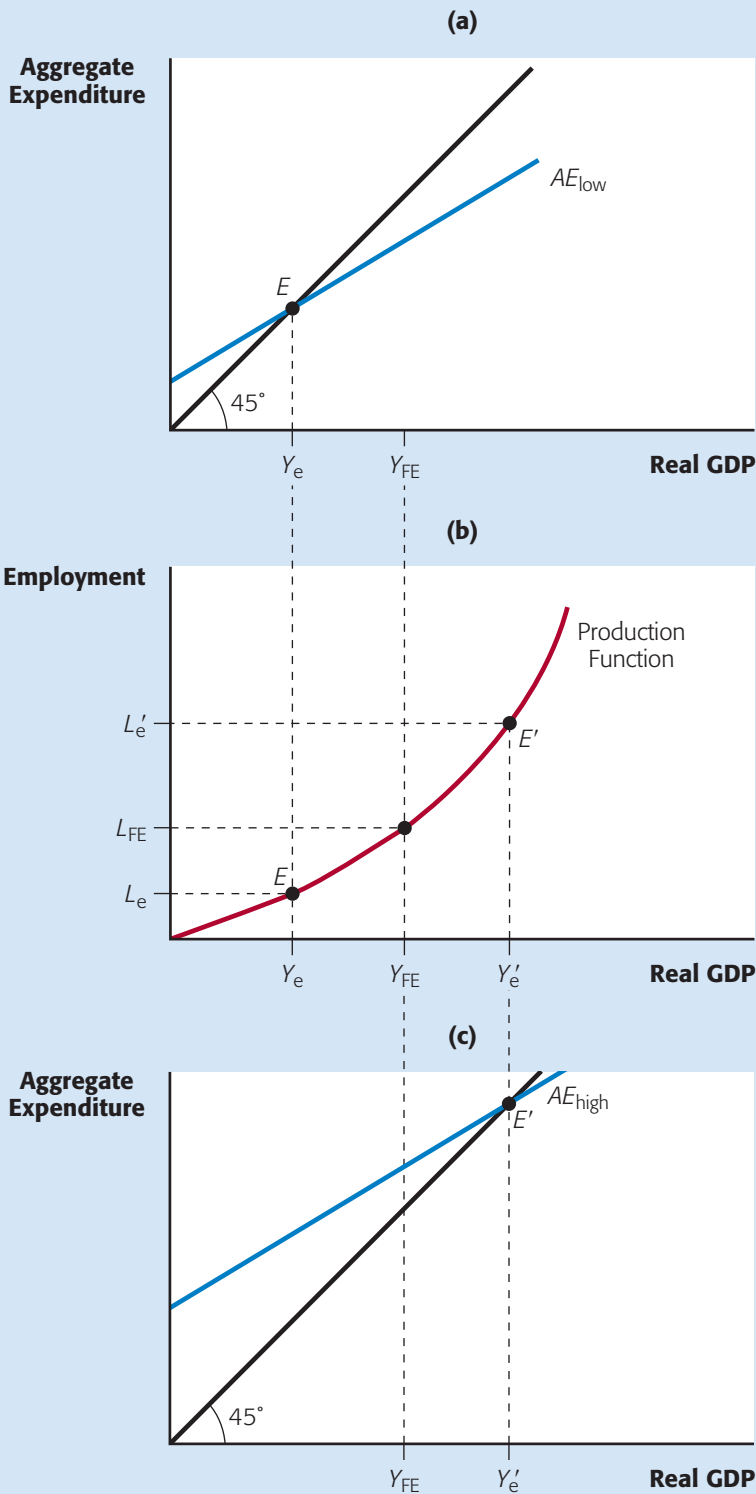
Figure 9 shows how we can find total employment in the economy. In panel (b), we show the economy's production function—the relationship between employment and output for a given capital stock and technology. This production function is similar to the one we used a few chapters ago in the classical model. But there is one important difference: The axes are reversed. Instead of measuring labor on the horizontal axis and output on the vertical axis, the production function in Figure 9 is turned on its side, with labor measured vertically and output measured horizontally. On the vertical axis, L_{FE} is the number of people who *would* be working if the economy were operating at full employment. The production function tells us that, at full employment, GDP would be Y_{FE} (potential output). This is the long-run equilibrium from the classical model.

But will Y_{FE} be the equilibrium in the short run? Not necessarily. One possible outcome is shown in panel (a). Here, the aggregate expenditure line and the 45° line intersect at point E . Equilibrium GDP in the short run is Y_e . But—according to the production function in panel (b)—to produce an output of Y_e requires employment of only L_e . Since L_e is less than L_{FE} , we will have abnormally low employment. Or, looked at another way, the level of *unemployment* will be higher than normal.

But why? What prevents firms from hiring the extra people who want jobs? After all, with more people working, producing more output, wouldn't there be more income in the economy and therefore more spending? Indeed, there would be. But

EQUILIBRIUM GDP DOES NOT NECESSARILY EQUAL FULL-EMPLOYMENT GDP

FIGURE 9



Panel (a) shows that, in the short run, equilibrium GDP can fall short of full employment GDP. This is illustrated by point E , where the aggregate expenditure line crosses the 45° line to determine an equilibrium GDP of Y_e . This is below full-employment output, Y_{FE} . The production function in panel (b) shows that at Y_e employment is L_e which lies below the full-employment level, L_{FE} .

Panel (c) shows the opposite case, in which equilibrium GDP exceeds its full-employment level. At point E' , the aggregate expenditure line crosses the 45° line to determine an equilibrium GDP of Y'_e , which exceeds full-employment output, Y_{FE} . The production function in panel (b) shows that with output at Y'_e employment is L'_e which lies above the full-employment level, L_{FE} .

not *enough* additional spending to justify the additional employment. To prove this, just look at what would happen if firms *did* hire L_{FE} workers. Output would rise to Y_{FE} , but at this output level, the aggregate expenditure line would lie below the 45° line, so *firms would be unable to sell all their output*. Unsold goods would pile up in inventories, and firms would cut back on production until output reached Y_e again, with employment back at L_e .

Panel (a) of Figure 9 shows that we can be in short-run equilibrium and yet have abnormally high unemployment. The reason: The aggregate expenditure line is *too low* to create an intersection at full-employment output.

In the short-run macro model, cyclical unemployment is caused by insufficient spending. As long as spending remains low, production will remain low, and unemployment will remain high.

What about the opposite possibility? In the short run, is it possible for spending to be *too high*, causing unemployment to be *too low*? Absolutely. Panel (c) of Figure 9 illustrates such a case. Here, the aggregate expenditure line and the 45° line intersect at point E' , giving us a short-run equilibrium GDP at Y'_e . According to the production function, producing an output of Y'_e requires employment of L'_e . Since L'_e is greater than the economy's normal employment L_{FE} , we will have abnormally high employment, and abnormally low unemployment.

In the short-run macro model, the economy can overheat because spending is too high. As long as spending remains high, production will exceed potential output, and unemployment will be unusually low.

In the previous chapter, we concluded that the classical model could not explain economic fluctuations. The short-run macro model, on the other hand, does provide an explanation: The aggregate expenditure line may be low, meaning that in the short run, equilibrium GDP is below full employment. Or aggregate expenditure may be high, meaning that in the short run, equilibrium GDP is above the full-employment level. (Of course, this is just a first step in explaining economic fluctuations. In later chapters, we'll add more realism to the model.)

What Happens When
Things Change?



WHAT HAPPENS WHEN THINGS CHANGE?

So far, you've seen how the economy's equilibrium level of output is determined in the short run, and the important role played by spending in determining that equilibrium. But now it's time to use Key Step #4 (What Happens When Things Change?) and explore how a spending shock—a sudden change in spending—affects equilibrium output.

A CHANGE IN INVESTMENT SPENDING

Suppose the equilibrium GDP in an economy is \$6,000 billion, and then business firms increase their investment spending on plant and equipment. This might happen because business managers feel more optimistic about the economy's future, or because there is a new “must have” technology (such as Internet connections in the late 1990s), or because government policy has changed and increased the incentive for firms to buy new plant and equipment. Whatever the cause, firms decide to increase yearly planned investment purchases by \$1,000 billion above the original level. What will happen?

THE EFFECT OF A CHANGE IN INVESTMENT SPENDING

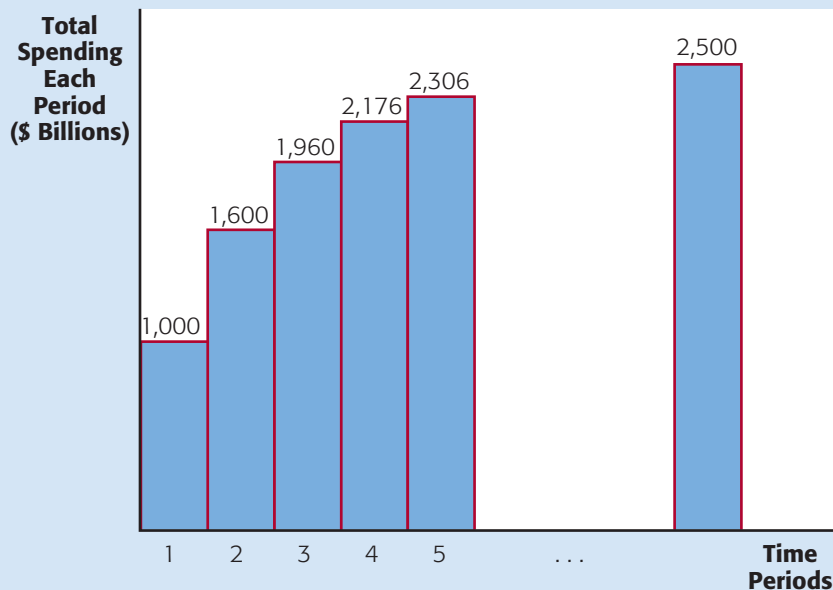


FIGURE 10

An increase in investment spending sets off a chain reaction, leading to successive rounds of increased spending and income. As shown here, a \$1,000 billion increase in investment first causes real GDP to increase by \$1,000 billion. Then, with higher incomes, households increase consumption spending by the *MPC* times the change in disposable income. In round 2, spending and GDP increase by another \$600 billion. In succeeding rounds, increases in income lead to further changes in spending, but in each round the increases in income and spending are smaller than in the preceding round.

First, sales revenue at firms that manufacture investment goods—firms like IBM, Bethlehem Steel, Caterpillar, and Westinghouse—will increase by \$1,000 billion. But remember, each time a dollar in output is produced, a dollar of income (factor payments) is created. Thus, the \$1,000 billion in additional sales revenue will become \$1,000 billion in additional income. This income will be paid out as wages, rent, interest, and profit to the households who own the resources these firms have purchased.²

What will households—as consumers—do with their \$1,000 billion in additional income? Remember that taxes are fixed, so that households are free to spend or save their additional income as they desire. What they will do depends crucially on the *marginal propensity to consume (MPC) in the economy*. If the *MPC* is 0.6, then consumption spending will rise by $0.6 \times \$1,000 \text{ billion} = \600 billion . Households will save the remaining \$400 billion.

But that is not the end of the story. When households spend an additional \$600 billion, firms that produce consumption goods and services—firms such as McDonald's, Coca-Cola, American Airlines, and Disney—will receive an additional \$600 billion in sales revenue, which, in turn, will become income for the households that supply resources to these firms. And when *these* households see *their* incomes rise by \$600 billion, they will spend part of it as well. With an *MPC* of 0.6, consumption spending will rise by $0.6 \times \$600 \text{ billion} = \360 billion , creating still more sales revenue for firms, and so on and so on. . . .

As you can see, an increase in investment spending will set off a chain reaction, leading to successive rounds of increased spending and income. The process is illustrated in Figure 10. After the \$1,000 billion increase in investment spending, there is a

² Some of the sales revenue will also go to pay for intermediate goods, such as raw materials, electricity, and supplies. But the intermediate-goods suppliers will also pay wages, rent, interest, and profit for *their* resources, so that household income will still rise by the full \$1,000 billion.

TABLE 5

**CUMULATIVE INCREASES
IN SPENDING WHEN
INVESTMENT INCREASES
BY \$1,000 BILLION**

Round	Additional Spending in This Round (Billions of Dollars)	Additional Spending in All Rounds (Billions of Dollars)
Initial Increase in Investment	1,000	1,000
Round 2	600	1,600
Round 3	360	1,960
Round 4	216	2,176
Round 5	130	2,306
Round 6	78	2,384
Round 7	47	2,431
Round 8	28	2,459
Round 9	17	2,476
Round 10	10	2,486
.	.	.
.	.	.
.	.	.
All other rounds	Very close to 14	Very close to 2,500

\$600 billion increase in consumption, then a \$360 billion increase in consumption, and on and on. Each successive round of additional spending is 60 percent of the round before. Total spending rises from \$1,000 billion to \$1,600 billion to \$1,960 billion and so on. And each time spending increases, output rises to match it. These successive increases in spending and output occur quickly—the process is largely completed within a year. At the end of the process, when the economy has reached its new equilibrium, spending and output will have increased considerably. But by how much?

Table 5 gives us the answer. The second column shows us the additional spending in each round, while the third column shows the cumulative rise in spending. As you can see, the cumulative increase gets larger and larger with each successive round, but it grows by less and less each time. Eventually, the additional spending in a given round is so small that we can safely ignore it. At this point, the cumulative increase in spending and output will be very close to \$2,500 billion—so close that we can ignore any difference.

THE EXPENDITURE MULTIPLIER

Let's go back and summarize what happened in our example: Business firms increased their investment spending by \$1,000 billion, and as a result, spending and output rose by \$2,500 billion. Equilibrium GDP increased by *more than* the initial increase in investment spending. In our example, the increase in equilibrium GDP (\$2,500 billion) was two-and-a-half times the initial increase in investment spending (\$1,000 billion). As you can verify, if investment spending had increased by half as much (\$500 billion), GDP would have increased by 2.5 times *that* amount (\$1,250 billion). In fact, *whatever* the rise in investment spending, equilibrium GDP would increase by a factor of 2.5, so we can write

$$\Delta GDP = 2.5 \times \Delta I^p.$$

In our example, the change in investment spending was *multiplied* by the number 2.5 in order to get the change in GDP that it causes. For this reason, 2.5 is called the *expenditure multiplier* in this example.

The expenditure multiplier is the number by which the change in investment spending must be multiplied to get the change in equilibrium GDP.

The value of the expenditure multiplier depends on the value of the *MPC* in the economy. If you look back at Table 5, you will see that each round of additional spending would have been larger if the *MPC* had been larger. For example, with an *MPC* of 0.9 instead of 0.6, spending in round 2 would have risen by \$900 billion, in round 3 by \$810 billion, and so on. The result would have been a larger cumulative change in GDP, and a larger multiplier.

There is a very simple formula we can use to determine the multiplier for *any* value of the *MPC*. To obtain it, let's start with our numerical example in which the *MPC* is 0.6. When investment spending rises by \$1,000 billion, the change in equilibrium GDP can be written as follows:

$$\Delta GDP = \$1,000 \text{ billion} + \$600 \text{ billion} + \$360 \text{ billion} + \$216 \text{ billion} + \dots$$

Factoring out the \$1,000 billion change in planned investment, this becomes

$$\begin{aligned} \Delta GDP &= \$1,000 \text{ billion} [1 + 0.6 + 0.36 + 0.216 + \dots] \\ &= \$1,000 \text{ billion} [1 + 0.6 + 0.6^2 + 0.6^3 + \dots] \end{aligned}$$

In this equation, \$1,000 billion is the change in investment (ΔI^p), and 0.6 is the *MPC*. To find the change in GDP that applies to *any* ΔI^p and *any* *MPC*, we can write

$$\Delta GDP = \Delta I^p \times [1 + (MPC) + (MPC)^2 + (MPC)^3 + \dots]$$

Now we can see that the term in brackets—the infinite sum $1 + MPC + (MPC)^2 + (MPC)^3 + \dots$ —is our multiplier. But what is its value?

We can borrow a rule from the mathematics of sums just like this one. The rule tells us that for any variable *H* that has a value between zero and 1, the infinite sum

$$1 + H + H^2 + H^3 + \dots$$

always has the value $1/(1 - H)$. So we can replace *H* with the *MPC*, since the *MPC* is always between zero and 1. This gives us a value for the multiplier of $1/(1 - MPC)$.

For any value of the MPC, the formula for the expenditure multiplier is $1/(1 - MPC)$.

In our example, the *MPC* was equal to 0.6, so the expenditure multiplier had the value $1/(1 - 0.6) = 1/0.4 = 2.5$. If the *MPC* had been 0.9 instead, the expenditure multiplier would have been equal to $1/(1 - 0.9) = 1/0.1 = 10$. The formula $1/(1 - MPC)$ can be used to find the multiplier for any value of the *MPC* between zero and one.

Using the general formula for the expenditure multiplier, we can restate what happens when investment spending increases:

$$\Delta GDP = \left[\frac{1}{(1 - MPC)} \right] \times \Delta I^p.$$

The multiplier effect is a rather surprising phenomenon. It tells us that an increase in investment spending ultimately affects GDP by *more* than the initial

Expenditure multiplier The amount by which equilibrium real GDP changes as a result of a one-dollar change in autonomous consumption, investment, or government purchases.

increase in investment. Moreover, the multiplier can work in the other direction, as you are about to see.

THE MULTIPLIER IN REVERSE

Suppose that, in Table 5, investment spending had *decreased* instead of increased. Then the initial change in spending would be $-\$1,000$ billion ($\Delta I^p = -\$1,000$ billion). This would cause a $\$1,000$ billion decrease in revenue for firms that produce investment goods, and they, in turn, would pay out $\$1,000$ billion less in factor payments. In the next round, households—with $\$1,000$ billion less in income—would spend $\$600$ billion less on consumption goods, and so on. The final result would be a $\$2,500$ billion *decrease* in equilibrium GDP.

Just as increases in investment spending cause equilibrium GDP to rise by a multiple of the change in spending, decreases in investment spending cause equilibrium GDP to fall by a multiple of the change in spending.

The multiplier formula we've already established will work whether the initial change in spending is positive or negative.

OTHER SPENDING SHOCKS

Shocks to the economy can come from other sources besides investment spending. In fact, when *any* sector's spending behavior changes, it will set off a chain of events similar to that in our investment example. Let's see how an increase in government spending could set off the same chain of events as an increase in investment spending.

Suppose that government agencies increased their purchases above previous levels. For example, the Department of Defense might raise its spending on new bombers, or state highway departments might hire more road-repair crews, or cities and towns might hire more teachers. If total government purchases rise by $\$1,000$ billion, then, once again, household income will rise by $\$1,000$ billion. As before, households will spend 60 percent of this increase, causing consumption—in the next round—to rise by $\$600$ billion, and so on and so on. The chain of events is exactly like that of Table 5, with one exception: The first line in column 1 would read, "Initial Increase in Government Purchases" instead of "Initial Increase in Investment." Once again, output would increase by $\$2,500$ billion.

Besides planned investment and government purchases, there are two other components of spending that can set off the same process. One is an increase in net exports. This can come about either from an increase in the economy's exports to foreigners, or a *decrease* in imports *from* foreigners. For example, either an increase in exports of $\$1,000$ billion, or a decrease in imports of $\$1,000$ billion, would increase net exports by $\$1,000$ billion and set off the same multiplier process described above.

Finally, a change in *autonomous consumption* can set off the process. For example, after a $\$1,000$ billion increase in autonomous consumption spending we would see further increases in consumption spending of $\$600$ billion, then $\$360$ billion, and so on. This time, the first line in column 1 of Table 5 would read, "Initial Increase in Autonomous Consumption," but every entry in the table would be the same.

Changes in planned investment, government purchases, net exports, or autonomous consumption lead to a multiplier effect on GDP. The expenditure multiplier— $1/(1 - MPC)$ —is what we multiply the initial change in spending by in order to get the change in equilibrium GDP.

The following four equations summarize how we use the expenditure multiplier to determine the effects of different spending shocks in the short-run macro model. Keep in mind that these formulas work whether the initial change in spending is positive or negative.

$$\Delta GDP = \left[\frac{1}{(1 - MPC)} \right] \times \Delta I^P$$

$$\Delta GDP = \left[\frac{1}{(1 - MPC)} \right] \times \Delta G$$

$$\Delta GDP = \left[\frac{1}{(1 - MPC)} \right] \times \Delta NX$$

$$\Delta GDP = \left[\frac{1}{(1 - MPC)} \right] \times \Delta a$$

A GRAPHICAL VIEW OF THE MULTIPLIER

Figure 11 illustrates the multiplier using our aggregate expenditure diagram. The darker line is the aggregate expenditure line from Figure 8. The aggregate expenditure line intersects the 45° line at point *E*, giving us an equilibrium GDP of \$6,000 billion.

Now, suppose that either autonomous consumption, investment spending, net exports, or government purchases rises by \$1,000 billion. Regardless of which of these types of spending increases, the effect on our aggregate expenditure line is the same: It will *shift upward* by \$1,000 billion, to the higher line in the figure. The new aggregate expenditure line intersects the 45° line at point *F*, showing that our new equilibrium GDP is equal to \$8,500 billion.

What has happened? An initial spending increase of \$1,000 billion has caused equilibrium GDP to increase from \$6,000 billion to \$8,500 billion—an increase of \$2,500 billion. This is just what our multiplier of 2.5 tells us. In general,

$$\Delta GDP = \left[\frac{1}{(1 - MPC)} \right] \times \Delta \text{Spending}$$

and in this case,

$$\text{\$2,500 billion} = 2.5 \times \text{\$1,000 billion.}$$

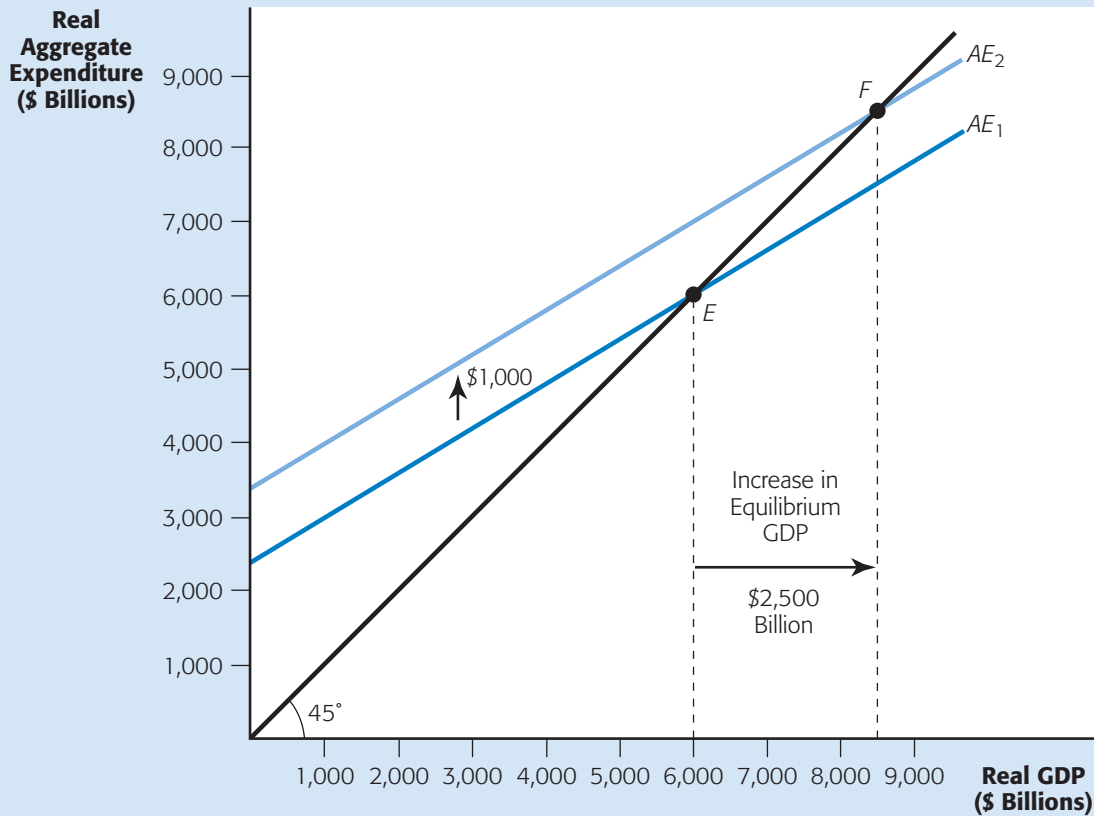
An increase in autonomous consumption spending, investment spending, government purchases, or net exports will shift the aggregate expenditure line upward by the increase in spending, causing equilibrium GDP to rise. The increase in GDP will equal the initial increase in spending times the expenditure multiplier.

AN IMPORTANT PROVISIO ABOUT THE MULTIPLIER

In this chapter, we've presented a model to help us focus on the central relationship between spending and output. To keep the model as simple as possible, we've ignored many real-world factors that interfere with, and reduce the size of, the

FIGURE 11

A GRAPHICAL VIEW OF THE MULTIPLIER



The economy starts off at point *E* with equilibrium real GDP of \$6,000 billion. A \$1,000 billion increase in spending shifts the aggregate expenditure line upward by \$1,000 billion, triggering the multiplier process. Eventually, the economy will reach a new equilibrium at point *F*, where the new, higher aggregate expenditure line crosses the 45° line. At *F*, real GDP is \$8,500 billion—an increase of \$2,500 billion.

Automatic stabilizers Forces that reduce the size of the expenditure multiplier and diminish the impact of spending shocks.

multiplier effect. These forces are called **automatic stabilizers** because, with a smaller multiplier, spending shocks will cause a much smaller change in GDP. As a result, economic fluctuations will be milder.

Automatic stabilizers reduce the size of the multiplier and therefore reduce the impact of spending shocks on the economy. With milder fluctuations, the economy is more stable.

How do automatic stabilizers work? They shrink the additional spending that occurs in each round of the multiplier, and thereby reduce the final multiplier effect on equilibrium GDP. In Table 5, automatic stabilizers would reduce each of the numerical entries after the first \$1,000 billion, and lead to a final change in GDP smaller than \$2,500 billion.

Here are some of the real-world automatic stabilizers we've ignored in the simple, short-run macro model of this chapter:

Taxes. We've been assuming that taxes remain constant, so that a rise in income causes an equal rise in disposable income. But some taxes (like the personal income tax) rise with income. As a result, in each round of the multiplier, the increase in disposable income will be smaller than the increase in income. With a smaller rise in disposable income, there will be a smaller rise in consumption spending as well.

Transfer Payments. Some government transfer payments fall as income rises. For example, many laid-off workers receive unemployment benefits, which help support them for several months while they are unemployed. But when income and output rise, employment also rises, and newly hired workers must give up their unemployment benefits. As a result, a rise in income will cause a smaller rise in *disposable* income. Consumption will then rise by less in each round of the multiplier.

Interest Rates. In a later chapter, you'll learn that an increase in output often leads to rising interest rates as well. This will crowd out some investment spending, making the increase in aggregate expenditure smaller than our simple story suggests.

Prices. In a later chapter, you'll learn that the price level tends to rise as spending and production increase. This, in turn, tends to counteract any increase in spending.

Imports. Some additional spending is on goods and services imported from abroad. That is, instead of remaining constant, imports often rise as income rises, and net exports therefore fall as income rises. This helps to counteract any increase in spending caused by a rise in income.

Forward-looking Behavior. Consumers may be *forward looking*. If they realize that the fluctuations in the economy are temporary, their consumption spending may be less sensitive to changes in their current income. Therefore, any change in income will cause a smaller change in consumption spending, and lead to a smaller multiplier effect.

Remember that each of these automatic stabilizers reduces the size of the multiplier, making it smaller than the simple formulas given in this chapter. For example, the simple formula for the expenditure multiplier is $1/(1 - MPC)$. With an *MPC* of about 0.9—which is in the ballpark for the United States and many other countries—we would expect the multiplier to be about 10 . . . *if the simple formula were accurate*. In that case, a \$1,000 billion increase in government spending would cause output to rise by \$10,000 billion—quite a large multiplier effect.



It's easy to become confused about the relationship between consumption spending and the expenditure multiplier. Does a change in consumption spending *cause* a multiplier effect? Or does the multiplier effect create an increase in consumption spending? Actually, the causation runs in both directions. The key is to recognize that there are *two* kinds of changes in consumption spending.

One kind of change is a change in autonomous consumption spending (the term *a* in the consumption function). This change will *shift* the aggregate expenditure line up or down, telling us that total spending will be greater or smaller at *any* level of income. It is the kind of change that *causes* a multiplier effect.

But consumption also changes when something other than autonomous consumption sets off a multiplier effect. This is because consumption depends on income, and income always increases during the successive rounds of the multiplier effect. Such a change in consumption is represented by a movement *along* the aggregate expenditure line, rather than a shift.

Whenever you discuss a change in consumption spending, make sure you know whether it is a change in autonomous consumption (a shift of the curve) or a change in consumption caused by a change in income (movement along the curve).

But after we take account of all of the automatic stabilizers, the multiplier is considerably smaller. How much smaller? Most of the forecasting models used by economists in business and government predict that the multiplier effect takes about nine months to a year to work its way through the economy. At the end of the process, the multiplier has a value of about 1.5. This means that a \$1,000 billion increase in, say, government spending should cause GDP to increase by only about \$1,500 billion in a year. This is much less than the \$10,000 billion increase predicted by the simple formula $1/(1 - MPC)$ when the MPC is equal to 0.9.

In the real world, due to automatic stabilizers, spending shocks have much weaker impacts on the economy than our simple multiplier formulas would suggest.

Finally, there is one more automatic stabilizer you should know about, perhaps the most important of all: the *passage of time*. Why is this an automatic stabilizer? Because, as you've learned, the impact of spending shocks on the economy are *temporary*. As time passes, the classical model—lurking in the background—stands ready to take over. A few months after a shock, the corrective mechanisms we discussed in the previous chapter begin to operate, and the economy begins to return to full employment. As time passes, the impact of the spending shock gradually disappears. And if we wait long enough—a few years or so—the effects of the shock will be gone entirely. That is, after a shock pulls us away from full-employment GDP, the economy will eventually return to full-employment GDP—right where it started. We thus conclude that

in the long run, our multipliers have a value of zero: No matter what the change in spending or taxes, output will return to full employment, so the change in equilibrium GDP will be zero.

Of course, the year or two we must wait can seem like an eternity to those who are jobless when the economy is operating below its potential. The short run is not to be overlooked. This is why, in the next several chapters, we will continue with our exploration of the short run, building on the macro model you've learned in this chapter. However, we'll make the analysis more complete and more realistic by bringing in some of the real-world features that were not fully considered here.

COMPARING MODELS: LONG RUN AND SHORT RUN

Before leaving this chapter, it's important to note some startling differences between the long-run classical model you learned about a few chapters ago, and the short-run macro model of this chapter. We've already discussed one of these differences: In the classical model, the economy operates *automatically* at full-employment, or potential, output. In the short-run macro model, by contrast, the economy can operate above its potential or below its potential. The reason for the difference is that, in the short run, spending affects output: A negative spending shock can cause a recession that pushes output below potential GDP; a positive spending shock can cause a rapid expansion that pushes the economy above potential GDP.

There are two other important contrasts between the predictions of the two models. One concerns the role of saving in the economy, and the other concerns the effectiveness of fiscal policy. Let's explore each of these issues in turn.

THE ROLE OF SAVING

In the long run, saving has positive effects on the economy. This was demonstrated two chapters ago, when—using the classical model—we discussed economic growth. Suppose, for example, that households decide to save more at any level of income. In the long run, the extra saving will flow into the loanable funds market, where it will be borrowed by business firms to purchase new plant and equipment. Thus, an increase in saving automatically leads to an increase in planned investment, faster growth in the capital stock, and a faster rise in living standards. Indeed, we can expect an increase in saving to have precisely these effects . . . in the long run.

But in the short run, the automatic mechanisms of the classical model do not keep the economy operating at its potential. On the contrary, *spending* influences output in the short run. If households decide to save more at each income level, they also—by definition—*spend less* at each income level. Or, putting it another way, an increase in saving is the same as a *decrease* in autonomous consumption spending, a . As you’ve learned in this chapter, a decrease in autonomous consumption spending causes a decrease in output through the multiplier process. If the economy is initially operating at full employment, the increase in saving will push output *below* its potential.

In the long run, an increase in the desire to save leads to faster economic growth and rising living standards. In the short run, however, it can cause a recession that pushes output below its potential.

You can see that there are two sides to the “savings coin.” The impact of increased saving is positive in the long run and potentially dangerous in the short run. Are you wondering how we get from the potentially harmful short-run effect of higher saving to the beneficial long-run effect? We’ll address this question a few chapters later, when we examine how the economy adjusts from its short-run equilibrium to its long-run equilibrium.

THE EFFECT OF FISCAL POLICY

In the classical model, you learned that fiscal policy—changes in government spending or taxes designed to change equilibrium GDP—is completely ineffective. More specifically, an increase in government purchases *crowds out* an equal amount of household and business spending: The rise in G is exactly matched by the decrease in C and I . . . in the long run.

But in the short run, once again, we cannot rely on the mechanisms of the classical model that are so effective in the long run. In the short run, *an increase in government purchases causes a multiplied increase in equilibrium GDP*. Therefore, in the short run, fiscal policy can actually change equilibrium GDP!

This important observation suggests that fiscal policy could, in principle, play a role in altering the path of the economy. If output begins to dip below potential, couldn’t we use fiscal policy to pull us out of it or even prevent the recession entirely? For example, if investment spending decreases by \$100 billion, setting off a negative multiplier effect, couldn’t we just increase government purchases by \$100 billion to set off an equal, positive multiplier effect? Why wait the many months or years it would take for the classical model to “kick in” and bring the economy back to full employment when we have such a powerful tool—fiscal policy—at our disposal?

Indeed, in the 1960s and early 1970s, this was the thinking of many economists. At that time, the view that fiscal policy could effectively smooth out economic

fluctuations—perhaps even eliminate them entirely—was very popular. But very few economists believe this today. Why? In part, because of practical difficulties in executing the right fiscal policy at the right time. But more importantly, the rules of economic policy making have changed: The Federal Reserve now attempts to neutralize fiscal policy changes long before they can affect spending and output in the economy. In later chapters, we'll discuss the practical difficulties of executing fiscal policy and how the Federal Reserve has changed the “rules of the game.”

THE RECESSION OF 1990–1991

Using the THEORY



Our most recent recession began in the second half of 1990 and continued into 1991. Table 6 tells the story. The second column shows real GDP in 1996 dollars in each of several quarters. For example, “1990:2” denotes the second quarter of 1990, and during that three-month period, GDP was \$6,705 billion at an annual rate. (That is, if we had continued producing that quarter’s GDP for an entire year, we *would* have produced a total of \$6,705 billion worth of goods and services in the year 1990.)

As you can see, real GDP began to fall in the third quarter, and it continued to drop until the second quarter of 1991. In all, GDP fell for three consecutive quarters. During this time, real output fell by \$100 billion, a drop of about 1.5 percent. At the same time, the unemployment rate rose, from 5.1 percent in June of 1990 to 7.7 percent in June of 1992. The economy had not completely recovered by the presidential election of November 1992, and many observers believe that the recession and slow recovery were the deciding factors in George Bush’s loss to Bill Clinton.

Can our short-run model help us understand what caused this recession? Very much so. In retrospect, we can see that there were two separate spending shocks to the economy in early 1990.

First, for a variety of reasons, a financial crisis had developed, in which some banks and savings and loan associations were near bankruptcy. Many banks, playing it safe, responded by cutting back on loans for new home purchases, as well as for business expansion. The media began to speak of a “credit crunch,” in which homebuyers and businesses were forced to pay very high interest rates on loans, or were unable to borrow at all. The consequence was a sizable decrease in the demand for new housing and for plant and equipment—an investment spending

TABLE 6

THE RECESSION OF 1990–1991

Quarter	Real GDP (Billions of 1996 Dollars)	Change in Real GDP from Previous Quarter (Billions of 1996 Dollars)	Real Investment Spending (Billions of 1996 Dollars)	Consumer Confidence Index
1990:2	6,705		933	105
1990:3	6,695	–10	913	90
1990:4	6,644	–51	850	61
1991:1	6,616	–28	815	65
1991:2	6,658	+42	809	77

shock. (Remember that investment spending includes new housing construction as well as plant and equipment.)

The second shock resulted from global politics. In the summer of 1990, Iraqi troops invaded and occupied Kuwait. The United States responded by sending troops to Kuwait and, in early 1991, launched an attack on Iraqi troops. Americans began to fear a prolonged and costly war in the Middle East, one that would, among other things, cause a large increase in the price of oil. They remembered that in the early 1970s, the last time that oil prices had risen substantially, the U.S. economy plunged into recession. As a result, American households became less confident about the economy.

The fifth column of Table 6 shows the rapid decline in the *consumer confidence index* that was occurring at the time. The index is based on a survey of about 5,000 households. Each month, these households respond to questions about their job and career prospects in the months ahead, their expected income, their spending plans, and so forth. A drop in consumer confidence makes households spend less at *any* income level. Or, put another way, households wanted to *save more* at any income level. Viewed either way, the drop in consumer confidence caused a decrease in autonomous consumption, a . This was the second spending shock to the economy.

In sum, in early 1990, there were two spending shocks to the economy: a decline in planned investment and a decline in autonomous consumption. Each of these shocks had a multiplier effect on the economy, causing income and spending to decline in successive rounds for almost a year. Beginning in 1992, the credit crunch began to subside, increasing investment spending, and the Gulf War ended, increasing consumer confidence. At the same time, the long-run corrective forces of the classical model were beginning to work. Together, all of these factors helped the economy to recover in 1992 and on into 1993.



Jennifer Gardner has explored what happened in the labor market during 1990 and 1991. You can find her analysis at <http://www.bls.gov/opub/mlr/1994/06/art1full.pdf>.

S U M M A R Y

In the short run, spending depends on income and income depends on spending. The short-run macro model was developed to explore this circular connection between spending and income.

Total spending is the sum of four aggregates—consumption spending by households, investment spending by firms, government purchases of goods and services, and net exports. Consumption spending (C) depends primarily on disposable income—what households have left over after paying taxes. The consumption function is a linear relationship between disposable income and consumption spending. The marginal propensity to consume—a number between zero and 1—indicates the fraction of each additional dollar of disposable income that is consumed. For a given level of income, consumption spending can change as a result of changes in the interest rate, wealth, or expectations about the future. Each of these changes will shift the consumption function.

Investment spending (I^p), government purchases (G), and net exports (NX) are taken as given values, determined by forces outside our analysis. Aggregate expenditure (AE) is the sum $C + I^p + G + NX$; it varies with income because consumption spending varies with income.

Equilibrium GDP is the level of output at which aggregate expenditure is just equal to GDP (Y). If AE exceeds Y , then firms will experience unplanned decreases in inventories. They will respond by increasing production. If Y exceeds AE , firms will find their inventories increasing and will respond by reducing production. Only when $AE = Y$ will there be no unplanned inventory changes and no reason for firms to change production. Graphically, this occurs at the point where the aggregate expenditure line intersects the 45° line.

Spending shocks will change the economy's short-run equilibrium. An increase in investment spending, for example, shifts the aggregate expenditure line upward and triggers the multiplier process. The initial increase in investment spending causes income to increase. That, in turn, leads to an increase in consumption spending, a further increase in income, more consumption spending, and so on. The economy eventually reaches a new equilibrium with a change in GDP that is a multiple of the original increase in spending. Other spending shocks would have similar effects. The size of the *expenditure multiplier* is determined by the marginal propensity to consume.

There are several important differences between the short-run macro model and the long-run classical model. In

the long run, the economy operates at potential output; in the short run, GDP can be above or below potential. In the long run, saving contributes to economic growth by making funds available for firms to invest in new capital. In the short run,

increased saving means reduced spending and a lower level of output. Finally, fiscal policy is completely ineffective in the long run, but can have important effects on total demand and output in the short run.

KEY TERMS

short-run macro model
disposable income
consumption
function

autonomous consumption
spending
marginal propensity to
consume

consumption–income line
aggregate expenditure
equilibrium GDP

expenditure multiplier
automatic stabilizers

REVIEW QUESTIONS

- Briefly describe the four main categories of spending.
- There are three different ways to interpret the marginal propensity to consume. What are they?
- List, and briefly explain, the main determinants of consumption spending. Indicate whether a change in each determinant causes a movement along, or a shift of, the consumption–income line.
- What are the main components of *planned investment* or *investment spending*? How does the definition of actual investment differ from planned investment?
- What conditions must be satisfied in order for GDP to be at its equilibrium value? Is this equilibrium GDP the same as the economy's potential GDP? Why or why not?
- Suppose that an increase in government purchases disturbs the economy's short-run equilibrium. Describe what happens as the economy adjusts to the change in spending.
- What is the expenditure multiplier? How is it calculated, and how is it used?
- What is an automatic stabilizer? List some automatic stabilizers for the U.S. economy. Which of these stabilizers do you think have gotten stronger, and which weaker, over the past several decades? Why?
- Compare the macroeconomic role of saving in the short run and in the long run.

PROBLEMS AND EXERCISES

- | 1. | <i>Y</i> | <i>C</i> | <i>I</i> | <i>G</i> | <i>NX</i> |
|----|----------|----------|----------|----------|-----------|
| | 3,000 | 2,500 | 300 | 500 | 200 |
| | 4,000 | 3,250 | 300 | 500 | 200 |
| | 5,000 | 4,000 | 300 | 500 | 200 |
| | 6,000 | 4,750 | 300 | 500 | 200 |
| | 7,000 | 5,500 | 300 | 500 | 200 |
| | 8,000 | 6,250 | 300 | 500 | 200 |
- What is the marginal propensity to consume implicit in this data?
 - Plot a 45° line, and then use the data to draw an aggregate expenditure line.
 - What is the equilibrium level of real GDP?
- | 2. | <i>Y</i> | <i>C</i> | <i>I</i> | <i>G</i> | <i>NX</i> |
|----|----------|----------|----------|----------|-----------|
| | 7,000 | 6,100 | 400 | 1,000 | 500 |
| | 8,000 | 6,900 | 400 | 1,000 | 500 |
| | 9,000 | 7,700 | 400 | 1,000 | 500 |
| | 10,000 | 8,500 | 400 | 1,000 | 500 |
| | 11,000 | 9,300 | 400 | 1,000 | 500 |
| | 12,000 | 10,100 | 400 | 1,000 | 500 |
| | 13,000 | 10,900 | 400 | 1,000 | 500 |
- What is the marginal propensity to consume implicit in this data?
 - What is the numerical value of the multiplier for this economy?
 - What is the equilibrium level of real GDP?

- d. Suppose that government purchases (G) decreased from 1,000 to 400 at each level of income. What would happen to the equilibrium level of real GDP?
3. Use an aggregate expenditure diagram to show the effect of each of the following changes:
- an increase in autonomous consumption spending due, say, to optimism on the part of consumers
 - an increase in U.S. exports
 - a decreases in taxes
 - an increase in U.S. imports
- In each case, be sure to label the initial equilibrium and the new equilibrium.
4. What would be the effect on real GDP and total employment of each of the following changes?
- As a result of restrictions on imports into the United States, net exports (NX) increase.
 - The federal government launches a new program to improve highways, bridges, and airports.
 - Banks are offering such high interest rates that consumers decide to save a larger proportion of their incomes.
- d. The growth of Internet retailing leads business firms to purchase more computer hardware and software.
5. Using the data given in Problem 2, construct a table similar to Table 5 in this chapter.
- Show what would happen in the first five rounds following an increase in investment spending from 400 to 800.
 - What would be the ultimate effect of that increase in investment spending?
 - How much would households spend on consumption goods in the new equilibrium?
6. Suppose that households become thrifter— that is, they now wish to save a larger proportion of their disposable income and spend a smaller proportion.
- In the table in Problem 2, which column of data would be affected? How is it affected?
 - Draw an aggregate expenditure diagram and show how an increase in saving can be measured in that diagram.
 - Use your aggregate expenditure diagram to show how an economy that is initially in short-run equilibrium will respond to an increase in thriftiness.

CHALLENGE QUESTIONS

- Read Appendix 1 (if you have not already done so). Then, suppose that $a = 600$, $b = 0.75$, $T = 400$, $I^p = 600$, $G = 700$, and $NX = 200$. Calculate the equilibrium level of real GDP. Then check that the equilibrium value equals the sum $C + I^p + G + NX$.
- The short-run equilibrium condition that $Y = C + I^p + G + NX$ can be reinterpreted as follows. First, subtract C from both sides to get $Y - C = I^p + G + NX$. Then

note that all income not spent on consumption goods is either taxed or saved, so that $Y - C = S + T$. Now combine the two equations to obtain $S + T = I^p + G + NX$.

Construct a diagram with real GDP measured on the horizontal axis. Draw two lines—one for $S + T$ and the other for $I^p + G + NX$. How would you interpret the point where the two lines cross? What would happen if investment spending increased?

EXPERIENTIAL EXERCISES

- Read Jane Katz's "When the economy goes south: What happens during a recession?" available from the Federal Reserve Bank of Boston at http://www.bos.frb.org/economic/nerr/katz99_3.htm. Also, re-read the "Using the Theory" section in this chapter. Then use a graph to show your interpretation of the 1990–91 recession, using the ideas you've learned in this chapter.



- Business investment spending is an important component of aggregate expenditure. Review the "Business Bulletin" column in the Thursday *Wall Street Journal*. What are some recent trends in investment spending? Are these trends likely to cause an increase or a decrease in aggregate expenditure? (*Note:* Purchases of stocks and bonds are *not* investment in the sense described in this chapter!)

APPENDIX 1

FINDING EQUILIBRIUM GDP ALGEBRAICALLY

The chapter showed how we can find equilibrium GDP using tables and graphs. This appendix demonstrates an algebraic way of finding the equilibrium GDP.

Our starting point is the relationship between consumption and disposable income given in the chapter,

$$C = a + bY_D$$

where a represents autonomous consumption spending, and b represents the marginal propensity to consume. Remember that disposable income (Y_D) is the income that the household sector has left after taxes. Letting T represent taxes, and Y represent total income or GDP, we have

$$Y_D = Y - T$$

If we now substitute $Y_D = Y - T$ into $C = a + bY_D$, we get an equation showing consumption at each level of income:

$$C = a + b(Y - T).$$

We can rearrange this equation algebraically to read

$$C = (a - bT) + bY.$$

This is the general equation for the consumption-income line. When graphed, the term in parentheses ($a - bT$) is the vertical intercept, and b is the slope. (Figure 5 shows a specific example of this line in which $a = \$2,000$, $b = 0.6$, and $T = \$2,000$.)

As you've learned, total spending or aggregate expenditure (AE) is the sum of consumption spending (C), investment spending (I^p), government spending (G) and net exports (NX):

$$AE = C + I^p + G + NX.$$

If we substitute for C the equation $C = (a - bT) + bY$, we get

$$AE = a - bT + bY + I^p + G + NX.$$

Now we can use this equation to find the equilibrium GDP. Equilibrium occurs when output (Y) and aggregate expenditure (AE) are the same. That is,

$$Y = AE$$

or, substituting the equation for AE ,

$$Y = a - bT + bY + I^p + G + NX.$$

This last equation will hold true only when Y is at its equilibrium value. We can solve for equilibrium Y by first bringing all terms involving Y to the left-hand side:

$$Y - bY = a - bT + I^p + G + NX.$$

Next, factoring out Y , we get

$$Y(1 - b) = a - bT + I^p + G + NX.$$

Finally, dividing both sides of this equation by $(1 - b)$ yields

$$Y = \frac{a - bT + I^p + G + NX}{1 - b}.$$

This last equation shows how equilibrium GDP depends on a (autonomous consumption), b (the MPC), T (taxes), I^p (investment spending), G (government purchases), and NX (net exports). These variables are all determined "outside our model." That is, they are given values that we use to determine equilibrium output, but they are not themselves affected by the level of output. Whenever we use actual numbers for these given variables in the equation, we find the same equilibrium GDP we would find using a table or a graph.

In the example we used throughout the chapter, the given values (found in Tables 1, 2, and 4) are, in billions of dollars, $a = 2,000$; $b = 0.6$; $T = 2,000$; $I^p = 700$; $G = 500$; and $NX = 400$. Plugging these values into the equation for equilibrium GDP, we get

$$\begin{aligned}
 Y &= \frac{2,000 - (0.6 \times 2,000) + 700 + 500 + 400}{1 - 0.6} \\
 &= \frac{2,400}{0.4} \\
 &= 6,000.
 \end{aligned}$$

This is the same value we found in Table 4 and Figure 8.

APPENDIX 2

THE SPECIAL CASE OF THE TAX MULTIPLIER

You learned in this chapter how changes in autonomous consumption, investment, and government purchases affect aggregate expenditure and equilibrium GDP. But there is another type of change that can influence equilibrium GDP: a change in taxes. For this type of change, the formula for the multiplier is slightly different from the one presented in the chapter.

Let's suppose that household taxes (T) decrease by \$1,000 billion. The immediate impact is to increase households' *disposable income* (Y_D) by \$1,000 billion at the current level of income. As a result, consumption spending will increase. But by how much?

The answer is, *less* than \$1,000 billion. When households get a tax cut, they increase their spending *not* by the full amount of the cut, but only by a *part* of it. The amount by which spending initially increases depends on the *MPC*. In our example, in which the *MPC* is 0.6, and disposable income rises by \$1,000 billion, the initial change in consumption spending is just \$600 billion. *This is the first change in spending that occurs after the tax cut.* Of course, once consumption spending rises, every subsequent round of the multiplier will work just as in Table 5: In the next round, consumption spending will rise by \$360 billion, and then \$216 billion, and so on.

Now let's compare what happens when taxes are cut by \$1,000 billion with what happens when spending rises by \$1,000 billion. As you can see from Table 5, when investment rises by \$1,000 billion, the initial change in spending is, by definition, \$1,000 billion. But when taxes are cut by \$1,000 billion, the initial change in spending is *not* \$1,000 billion, but \$600 billion.

Thus, the first line of the table is missing in the case of a \$1,000 billion tax cut. All subsequent rounds of the multiplier are the same, however. Therefore, we would expect the \$1,000 billion tax cut to cause a \$1,500 billion increase in equilibrium GDP—not the \$2,500 billion increase listed in the table.

Another way to say this is: For each dollar that taxes are cut, equilibrium GDP will increase by \$1.50 rather than \$2.50—the increase is one dollar less in the case of the tax cut. This observation tells us that the tax multiplier must have a numerical value *1.0 less than* the spending multiplier of the chapter.

Finally, there is one more difference between the spending multiplier of the chapter and the tax multiplier: While the spending multiplier is a positive number (because an increase in spending causes an increase in equilibrium GDP), the tax multiplier is a negative number, since a tax cut (a negative change in taxes) must be multiplied by a *negative* number to give us a *positive* change in GDP. Putting all this together, we conclude that

the tax multiplier is 1.0 less than the spending multiplier, and negative in sign.

Thus, if the *MPC* is 0.6 (as in the chapter), so that the spending multiplier is 2.5, then the tax multiplier will have a value of $-(2.5 - 1) = -1.5$.

More generally, since the tax multiplier is 1.0 less than the spending multiplier and is also negative, we can write

$$\text{Tax multiplier} = -(\text{spending multiplier} - 1).$$

Because the spending multiplier is $1/(1 - MPC)$, we can substitute to get

$$\begin{aligned} \text{Tax multiplier} &= -\left[\frac{1}{1 - MPC} - 1\right] \\ &= -\frac{1 - (1 - MPC)}{1 - MPC} \\ &= \frac{-MPC}{1 - MPC}. \end{aligned}$$

Hence,

the general formula for the tax multiplier is

$$\frac{-MPC}{(1 - MPC)}.$$

For any change in taxes, we can use the formula to find the change in equilibrium GDP as follows:

$$\Delta GDP = \frac{-MPC}{1 - MPC} \times \Delta T.$$

In our example, in which taxes were cut by \$1,000 billion, we have $\Delta T = -\$1,000$ billion and $MPC = 0.6$. Plugging these values into the formula, we obtain

$$\begin{aligned} \Delta GDP &= \left[\frac{-0.6}{1 - 0.6}\right] \times -\$1,000 \text{ billion} \\ &= \$1,500 \text{ billion.} \end{aligned}$$

THE BANKING SYSTEM AND THE MONEY SUPPLY

Everyone knows that money doesn't grow on trees. But where does it actually come from? You might think that the answer is simple: The government just prints it. Right?

Sort of. It is true that much of our money supply is, indeed, paper currency, provided by our national monetary authority. But most of our money is *not* paper currency at all. Moreover, the monetary authority in the United States—the Federal Reserve System—is technically not a part of the executive, legislative, or judicial branches of government. Rather, it is a quasi-independent agency that operates *alongside* of the government.

In future chapters, we'll make our short-run macro model more realistic by bringing in money and its effects on the economy. This will deepen your understanding of economic fluctuations, and help you understand our policy choices in dealing with them. But in this chapter, we focus on money itself, and the institutions that help create it. We will begin, in the next section, by taking a close look at what money is and how it is measured.

WHAT COUNTS AS MONEY

Money, loosely defined, is the means of payment in the economy. And as you will learn in the next chapter, the amount of money in circulation can affect the macroeconomy. This is why governments around the world like to know how much money is available to their citizens.

In practice, the standard definition of money is *currency*, *checking account balances*, and *travelers checks*. What do these have in common and why are they included in the definition of money when other means of payment—such as credit cards—are not included?

First, only *assets*—things of value that people own—are regarded as money. Paper currency, travelers checks, and funds held in checking accounts are all examples of assets. But *the right to borrow* is not considered an asset, so it is not part of the money supply. This is why the credit limit on your credit card, or your ability to go into a bank and borrow money, is not considered part of the money supply.

CHAPTER OUTLINE

What Counts as Money

Measuring the Money Stock

Assets and Their Liquidity
M1 and M2

The Banking System

Financial Intermediaries
Commercial Banks
A Bank's Balance Sheet

The Federal Reserve System

The Structure of the Fed
The Federal Open Market
Committee
The Functions of the Federal
Reserve

The Fed and the Money Supply

How the Fed Increases the
Money Supply
The Demand Deposit Multiplier
The Fed's Effect on the Banking
System as a Whole
How the Fed Decreases the
Money Supply
Some Important Provisos About
the Demand Deposit
Multiplier
Other Tools for Controlling the
Money Supply

Using the Theory: Bank Failures and Banking Panics

Second, only things that are widely *acceptable* as a means of payment are regarded as money. Currency, travelers checks, and personal checks can all be used to buy things or pay bills. Other assets—such as the funds in your savings account—cannot generally be used to pay for goods and services, and so they fail the acceptability test.

Finally, only highly *liquid* assets are regarded as money.

Liquidity The property of being easily converted into cash.

*An asset is considered **liquid** if it can be converted to cash quickly and at little cost. An illiquid asset, by contrast, can be converted to cash only after a delay, or at considerable cost.*

Checking account balances are highly liquid because you can convert them to cash at the ATM or by cashing a check. Travelers checks are also highly liquid. But stocks and bonds are *not* as liquid as checking accounts or travelers checks. Stock- and bondholders must go to some trouble and pay brokers' fees to convert these assets into cash.

MEASURING THE MONEY STOCK

In practice, governments have several alternative definitions of the money stock. These definitions include a selection of *assets* that are (1) generally acceptable as a means of payment and (2) relatively liquid.

Notice the phrase “*relatively* liquid.” This does not sound like a hard and fast rule for measuring the money supply, and indeed it is not. This is why there are different measures of the money supply: Each interprets the phrase “relatively liquid” in a different way. To understand this better, let’s look at the different kinds of liquid assets that people can hold.

ASSETS AND THEIR LIQUIDITY

Figure 1 lists a spectrum of assets, ranked according to their liquidity, along with the amounts of each asset in the U.S. public’s hands on January 31, 2000. The most liquid asset of all is **cash in the hands of the public**. It takes no time and zero expense to convert this asset into cash, since it’s *already* cash. At the beginning of 2000, the public—including residents of other countries—held about \$521 billion in cash.

Next in line are three asset categories of about equal liquidity. **Demand deposits** are the checking accounts held by households and business firms at commercial banks, including huge ones like the Bank of America or Citibank, and smaller ones like Simmons National Bank in Arkansas. These checking accounts are called “demand” deposits because when you write a check to someone, that person can go into a bank and, on demand, be paid in cash. This is one reason that demand deposits are considered very liquid: The person who has your check can convert it into cash quickly and easily. Another reason is that you can withdraw cash from your own checking account very easily—24 hours a day with an ATM card, or during banking hours if you want to speak to a teller. As you can see in the figure, the U.S. public held \$345 billion in demand deposits in early 2000.

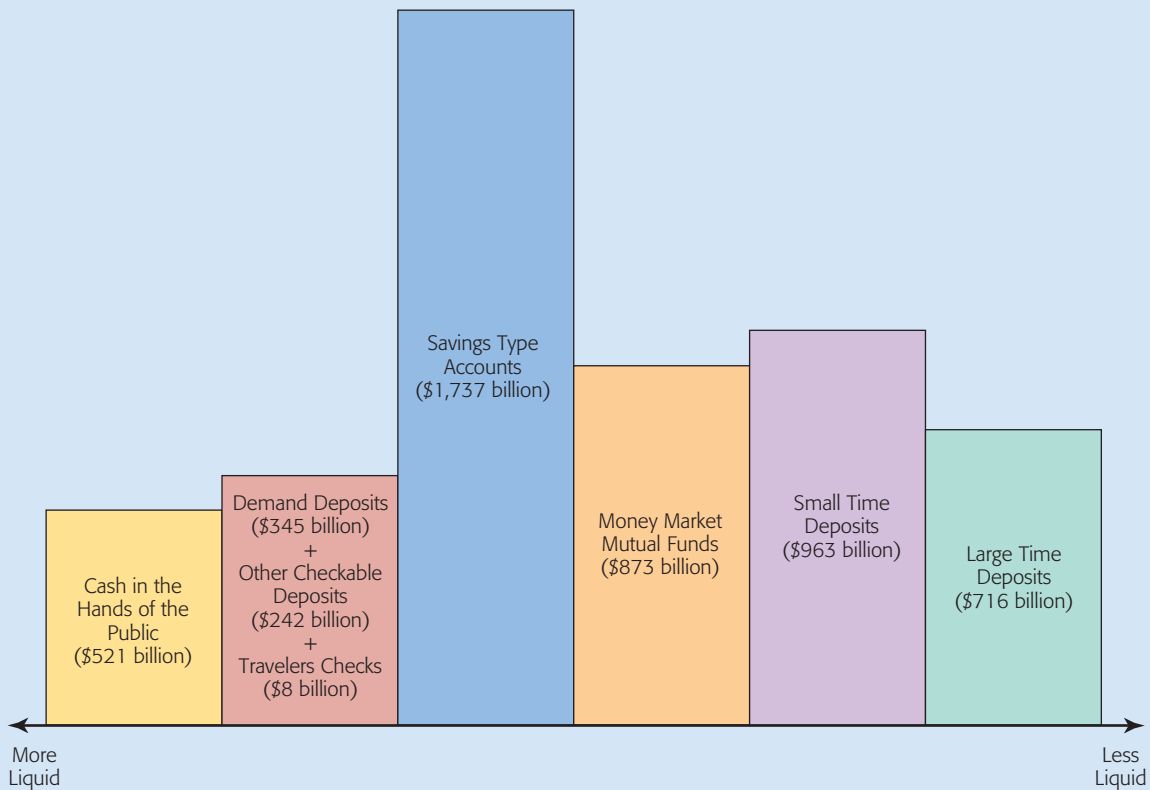
Other checkable deposits is a catchall category for several types of checking accounts that work very much like demand deposits. This includes *automatic transfers from savings accounts*, which are interest-paying savings accounts that automatically transfer funds into checking accounts when needed. On January 31, 2000, the U.S. public held \$242 billion of these types of checkable deposits.

Cash in the hands of the public
Currency and coins held outside of banks.

Demand deposits Checking accounts that do not pay interest.

MONETARY ASSETS AND THEIR LIQUIDITY (JANUARY 31, 2000)

FIGURE 1



Assets vary according to their liquidity—the ease with which they can be converted into cash. Assets toward the left side of this figure are more liquid than those toward the right side.

Travelers checks are specially printed checks that you can buy from banks or other private companies, like American Express. Travelers checks can be easily spent at almost any hotel or store. You can often cash them at a bank. You need only show an I.D. and countersign the check. In early 2000, the public held about \$8 billion in travelers checks.

Savings-type accounts at banks and other financial institutions (such as *savings and loan* institutions) amounted to \$1,737 billion in early 2000. These are less liquid than checking-type accounts, since they do not allow you to write checks. While it is easy to transfer funds from your savings account to your checking account, you must make the transfer yourself.

Next on the list are deposits in *retail money market mutual funds (MMMFs)*, which use customer deposits to buy a variety of financial assets. Depositors can withdraw their money by writing checks. In early 2000, the general public held about \$873 billion in such MMMFs.

Time deposits (sometimes called *certificates of deposit*, or *CDs*) require you to keep your money in the bank for a specified period of time (usually six months or longer), and impose an interest penalty if you withdraw early. In January 2000, the public held \$963 billion in *small time deposits* (in amounts under \$100,000) and \$716 billion in *large time deposits* (in amounts over \$100,000).

Now let's see how these assets have been used to define "money" in different ways.

M1 A standard measure of the money supply, including cash in the hands of the public, checking account deposits, and travelers checks.

M1 AND M2

The standard measure of the money stock is called **M1**. It is the sum of the first four assets in our list: cash in the hands of the public, demand deposits, other checkable deposits, and travelers checks. These are also the four most liquid assets in our list.

$$\text{M1} = \text{cash in the hands of the public} + \text{demand deposits} + \text{other checking account deposits} + \text{travelers checks.}$$

On January 31, 2000, this amounted to

$$\begin{aligned} \text{M1} &= \$521 \text{ billion} + \$345 \text{ billion} + \$242 \text{ billion} + \$8 \text{ billion} \\ &= \$1,116 \text{ billion.} \end{aligned}$$

When economists or government officials speak about "the money supply," they usually mean M1.

But what about the assets left out of M1? While savings accounts are not as liquid as any of the components of M1, for most of us there is hardly a difference. All it takes is an ATM card and, *presto*, funds in your savings account become cash. Money market funds held by households and businesses are fairly liquid, even though there are sometimes restrictions or special risks involved in converting them into cash. And even time deposits—if they are not too large—can be cashed in early with only a small interest penalty. When you think of how much "means of payment" you have, you are very likely to include the amounts you have in these types of accounts. This is why another common measure of the money supply, **M2**, adds these and some other types of assets to M1:

M2 M1 plus savings account balances, noninstitutional money market mutual fund balances, and small time deposits.

$$\text{M2} = \text{M1} + \text{savings-type accounts} + \text{retail MMMF balances} + \text{small denomination time deposits.}$$

Using the numbers for January 31, 2000 in the United States:

$$\begin{aligned} \text{M2} &= \$1,116 \text{ billion} + \$1,737 \text{ billion} + \$873 \text{ billion} + \$963 \text{ billion} \\ &= \$4,689 \text{ billion.} \end{aligned}$$

There are other official measures of the money supply besides M1 and M2 that add in assets that are less liquid than those in M2. But M1 and M2 have been the most popular, and most commonly watched, definitions.

It is important to understand that the M1 and M2 money stock measures exclude many things that people use regularly as a means of payment. Although M1 and M2 give us important information about the activities of the Fed and of banks, they do not measure all the different ways that people hold their wealth or pay for things. Credit cards, for example, are not included in any of the official measures. But for most of us, unused credit is a means of payment, to be lumped together with our cash and our checking accounts. As credit cards were issued to more and more Americans over the last several decades, the available means of payment increased considerably, much more than the increase in M1 and M2 suggests.

Technological advances—now and in the future—will continue the trend toward new and more varied ways to make payments. For example, at the 1996



In "The Changing Nature of the Payments System," (<http://www.phil.frb.org/files/br/brma00lm.pdf>) Loretta Mester explores the effects of technological change on the means of payment.

Olympics, people used electronic cash to make small transactions—smaller than would make sense with credit cards. You could buy a card worth \$5, \$10, or \$20 and use it in place of cash or checks. In 1999, Citibank began testing similar electronic cash cards in the Upper West Side of Manhattan. Electronic cash is clearly a means of payment, even though it is not yet included in any measure of the money supply. If electronic cash becomes important in the economy, it will probably be included in M1.



In our definitions of money—whether M1, M2, or some other measure—we include cash (coin and paper currency) only if it is *in the hands of the public*. The italicized words are important. Some of the nation's cash is stored in banks' vaults, and is released only when the public withdraws cash from their accounts. Other cash is in the hands of the Federal Reserve, which stores it for future release. But until this cash is released from bank vaults or the Fed, it is *not* part of the money supply. Only the cash possessed by households, businesses, or government agencies (other than the Fed) is considered part of the money supply.

Fortunately, the details and complexities of measuring money are not important for understanding the monetary system and monetary policy. For the rest of our discussion, we will make a simplifying assumption:

We will assume the money supply consists of just two components: cash in the hands of the public and demand deposits.

Money supply = cash in the hands of public + demand deposits.

As you will see later, our definition of the money supply corresponds closely to the liquid assets that our national monetary authority—the Federal Reserve—can control. While there is not much that the Federal Reserve can do directly about the amount of funds in savings accounts, MMMFs, or time deposits, or about the development of electronic cash or the ability to borrow on credit cards, it can tightly control the sum of cash in the hands of the public and demand deposits.¹

We will spend the rest of this chapter analyzing how money is created and what makes the money supply change. Our first step is to introduce a key player in the creation of money: the banking system.

THE BANKING SYSTEM

Think about the last time you used the services of a bank. Perhaps you deposited a paycheck in the bank's ATM, or withdrew cash to take care of your shopping needs for the week. We make these kinds of transactions dozens of times every year without ever thinking about what a bank really is, or how our own actions at the bank—and the actions of millions of other bank customers—might contribute to a change in the money supply.

FINANCIAL INTERMEDIARIES

Let's begin at the beginning: What are banks? They are important examples of **financial intermediaries**—business firms that specialize in assembling loanable funds from households and firms whose revenues exceed their expenditures, and channeling those funds to households and firms (and sometimes the government)

Financial intermediary A business firm that specializes in brokering between savers and borrowers.

¹ The Fed can also control some other types of checkable deposits. To keep our analysis as simple as possible, we consider only demand deposits.

whose expenditures exceed revenues. Financial intermediaries make the economy work much more efficiently than would be possible without them.

To understand this more clearly, imagine that Boeing, the U.S. aircraft maker, wants to borrow a billion dollars for three years. If there were no financial intermediaries, Boeing would have to make individual arrangements to borrow small amounts of money from thousands—perhaps millions—of households, each of which wants to lend money for, say, three months at a time. Every three months, Boeing would have to renegotiate the loans, and it would find borrowing money in this way to be quite cumbersome. Lenders, too, would find this arrangement troublesome. All of their funds would be lent to one firm. If that firm encountered difficulties, the funds might not be returned at the end of three months.

An intermediary helps to solve these problems by combining a large number of small savers' funds into custom-designed packages and then lending them to larger borrowers. The intermediary can do this because it can predict—from experience—the pattern of inflows of funds. While some deposited funds may be withdrawn, the overall total available for lending tends to be quite stable. The intermediary can also reduce the risk to depositors by spreading its loans among a number of different borrowers. If one borrower fails to repay its loan, that will have only a small effect on the intermediary and its depositors.

Of course, intermediaries must earn a profit for providing brokering services. They do so by charging a higher interest rate on the funds they lend than the rate they pay to depositors. But they are so efficient at brokering that both lenders and borrowers benefit. Lenders earn higher interest rates, with lower risk and greater liquidity, than if they had to deal directly with the ultimate users of funds. And borrowers end up paying lower interest rates on loans that are specially designed for their specific purposes.

The United States boasts a wide variety of financial intermediaries, including commercial banks, savings and loan associations, mutual savings banks, credit unions, insurance companies, and some government agencies. Some of these intermediaries—called *depository institutions*—accept deposits from the general public and lend the deposits to borrowers. There are four types of depository institutions:

1. *Savings and loan associations (S&Ls)* obtain funds through their customers' time, savings, and checkable deposits and use them primarily to make mortgage loans.
2. *Mutual savings banks* accept deposits (called *shares*) and use them primarily to make mortgage loans. They differ from S&Ls because they are owned by their depositors, rather than outside investors.
3. *Credit unions* specialize in working with particular groups of people, such as members of a labor union or employees in a specific field of business. They acquire funds through their members' deposits and make consumer and mortgage loans to other members.
4. *Commercial banks* are the largest group of depository institutions. They obtain funds mainly by issuing checkable deposits, savings deposits, and time deposits and use the funds to make business, mortgage, and consumer loans.

Since commercial banks will play a central role in the rest of this chapter, let's take a closer look at how they operate.

COMMERCIAL BANKS

A commercial bank (or just “bank” for short) is a private corporation, owned by its stockholders, that provides services to the public. For our purposes, the most im-

portant service is to provide checking accounts, which enable the bank's customers to pay bills and make purchases without holding large amounts of cash that could be lost or stolen. Checks are one of the most important means of payment in the economy. Every year, U.S. households and businesses write trillions of dollars' worth of checks to pay their bills, and many wage and salary earners have their pay deposited directly into their checking accounts. And as you saw in Figure 1, the public holds about as much money in the form of demand deposits and other checking-type accounts as it holds in cash.

Banks provide checking account services in order to earn a profit. Where does a bank's profit come from? Mostly from lending out the funds that people deposit and charging interest on the loans, but also by charging for some services directly, such as check-printing fees or that annoying dollar or so sometimes charged for using an ATM.

A BANK'S BALANCE SHEET

We can understand more clearly how a bank works by looking at its *balance sheet*, a tool used by accountants. A **balance sheet** is a two-column list that provides information about the financial condition of a bank at a particular point in time. In one column, the bank's *assets* are listed—everything of value that it *owns*. On the other side, the bank's *liabilities* are listed—the amounts that the bank *owes*.

Table 1 shows a simplified version of a commercial bank's balance sheet.

Why does the bank have these assets and liabilities? Let's start with the assets side. The first item, \$20 million, is the value of the bank's real estate—the buildings and the land underneath them. This is the easiest to explain, because a bank must have one or more branch offices in order to do business with the public.

Next, comes \$25 million in *bonds*, and \$65 million in *loans*. **Bonds** are IOUs issued by a corporation or a government agency when it borrows money. A bond promises to pay back the loan either gradually (e.g., each month), or all at once at some future date. **Loans** are IOUs signed by households or noncorporate businesses. Examples are auto loans, student loans, small business loans, and home mortgages (where the funds lent out are used to buy a home). Both bonds and loans generate interest income for the bank.

Next come two categories that might seem curious: \$2 million in “vault cash,” and \$8 million in “accounts with the Federal Reserve.” Vault cash, just like it sounds, is the coin and currency that the bank has stored in its vault. In addition, banks maintain their own accounts with the Federal Reserve, and they add and

Balance sheet A financial statement showing assets, liabilities, and net worth at a point in time.

Bond An IOU issued by a corporation or government agency when it borrows funds.

Loan An IOU issued by a household or noncorporate business when it borrows funds.

TABLE 1

A TYPICAL COMMERCIAL BANK'S BALANCE SHEET

Assets		Liabilities and Net Worth	
Property and buildings	\$ 20 million	Demand deposit liabilities	\$100 million
Government and corporate bonds	\$ 25 million	Net worth	\$ 20 million
Loans	\$ 65 million		
Cash in vault	\$ 2 million		
In accounts with the Federal Reserve	\$ 8 million		
Total Assets	\$120 million	Total Liabilities plus Net Worth	\$120 million

subtract to these accounts when they make transactions with other banks. Neither vault cash nor accounts with the Federal Reserve pay interest. Why, then, does the bank hold them? After all, a profit-seeking bank should want to hold as much of its assets as possible in income-earning form—bonds and loans.

There are two explanations for vault cash and accounts with the Federal Reserve. First, on any given day, some of the bank's customers might want to withdraw more cash than other customers are depositing. The bank must always be prepared to honor its obligations for withdrawals, so it must have some cash on hand to meet these requirements. This explains why it holds vault cash.

Second, banks are required by law to hold **reserves**, which are defined as *the sum of cash in the vault and accounts with the Federal Reserve*. The amount of reserves a bank must hold is called **required reserves**. The more funds its customers hold in their checking accounts, the greater the amount of required reserves. The **required reserve ratio**, set by the Federal Reserve, tells banks the fraction of their checking accounts that they must hold as required reserves.

For example, the bank in Table 1 has \$100 million in demand deposits. If the required reserve ratio is 0.1, this bank's required reserves are $0.1 \times \$100 \text{ million} = \10 million in reserves. The bank must hold *at least* this amount of its assets as reserves. Since our bank has \$2 million in vault cash, and \$8 million in its *reserve account* with the Federal Reserve, it has a total of \$10 million in reserves, the minimum required amount.

Now skip to the right side of the balance sheet. This bank's only liability is its demand deposits. Why are demand deposits a *liability*? Because the bank's customers have the right to withdraw funds from their checking accounts. Until they do, the bank *owes* them these funds.

Finally, the last entry. When we total up both sides of the bank's balance sheet, we find that it has \$120 million in assets, and only \$100 million in liabilities. If the bank were to go out of business, selling all of its assets and using the proceeds to pay off all of its liabilities (its demand deposits), it would have \$20 million left over. Who would get this \$20 million? The bank's owners—its stockholders. The \$20 million is called the bank's **net worth**. More generally,

$$\text{Net worth} = \text{Total assets} - \text{Total liabilities.}$$

We include net worth on the liabilities side of the balance sheet because it is, in a sense, what the bank would owe to its owners if it went out of business. Notice that, because of the way net worth is defined, both sides of a balance sheet must always have the same total: *A balance sheet always balances.*

Private banks are just one of the players that help determine the money supply. Now we turn our attention to the other key player—the Federal Reserve System.

THE FEDERAL RESERVE SYSTEM

Every large nation controls its banking system with a **central bank**. Most of the developed countries established their central banks long ago. For example, England's central bank—the Bank of England—was created in 1694. France was one of the latest in Europe, waiting until 1800 to establish the Banque de France. But the United States was even later. Although we experimented with central banks at various times in our history, we did not get serious about a central bank until 1913, when Congress established the *Federal Reserve System*.

Reserves Vault cash plus balances held at the Fed.

Required reserves The minimum amount of reserves a bank must hold, depending on the amount of its deposit liabilities.

Required reserve ratio The minimum fraction of checking account balances that banks must hold as reserves.

Net worth The difference between assets and liabilities.

Central bank A nation's principal monetary authority.

THE GEOGRAPHY OF THE FEDERAL RESERVE SYSTEM

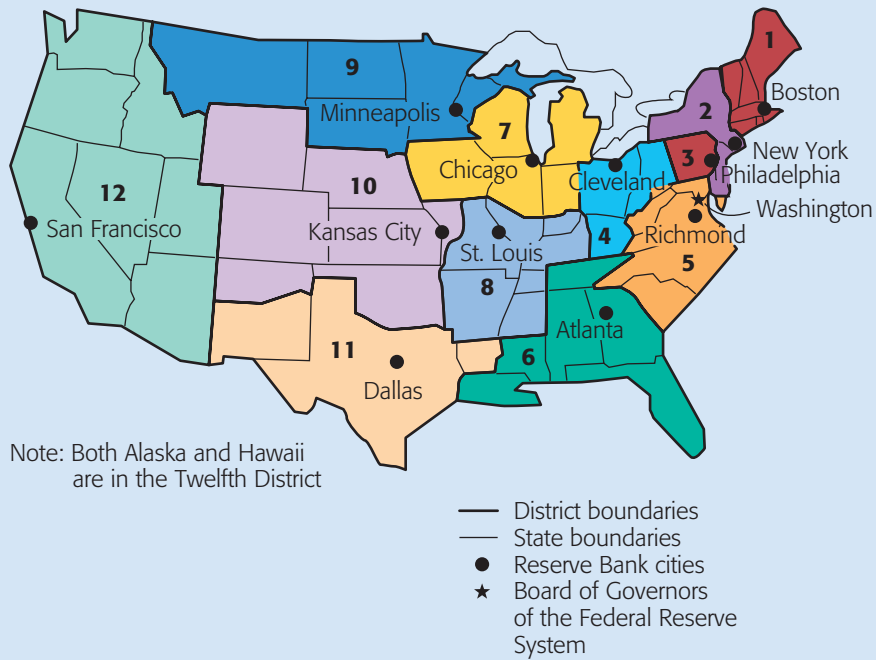


FIGURE 2

The United States is divided into 12 Federal Reserve districts, each with its own Federal Reserve Bank.

Why did it take the United States so long to create a central bank? Part of the reason is the suspicion of central authority that has always been part of U.S. politics and culture. Another reason is the large size and extreme diversity of our country, and the fear that a powerful central bank might be dominated by the interests of one region to the detriment of others. These special American characteristics help explain why our own central bank is different in form from its European counterparts.

One major difference is indicated in the very name of the institution—the Federal Reserve System. It does not have the word “central” or “bank” anywhere in its title, making it less suggestive of centralized power.

Another difference is the way the system is organized. Instead of a single central bank, the United States is divided into 12 Federal Reserve districts, each one served by its own Federal Reserve Bank. The 12 districts and the Federal Reserve Banks that serve them are shown in Figure 2. For example, the Federal Reserve Bank of Dallas serves a district consisting of Texas and parts of New Mexico and Louisiana, while the Federal Reserve Bank of Chicago serves a district including Iowa and parts of Illinois, Indiana, Wisconsin, and Michigan.

Another interesting feature of the Federal Reserve System is its peculiar status within the government. Strictly speaking, it is not even a *part* of any branch of government. But the *Fed* (as the system is commonly called) was created by Congress, and could be eliminated by Congress if it so desired. Second, both the president and Congress exert some influence on the Fed through their appointments of key officials in the system. Finally, the Fed’s mission is not to make a profit like an ordinary corporation, but rather to serve the general public.

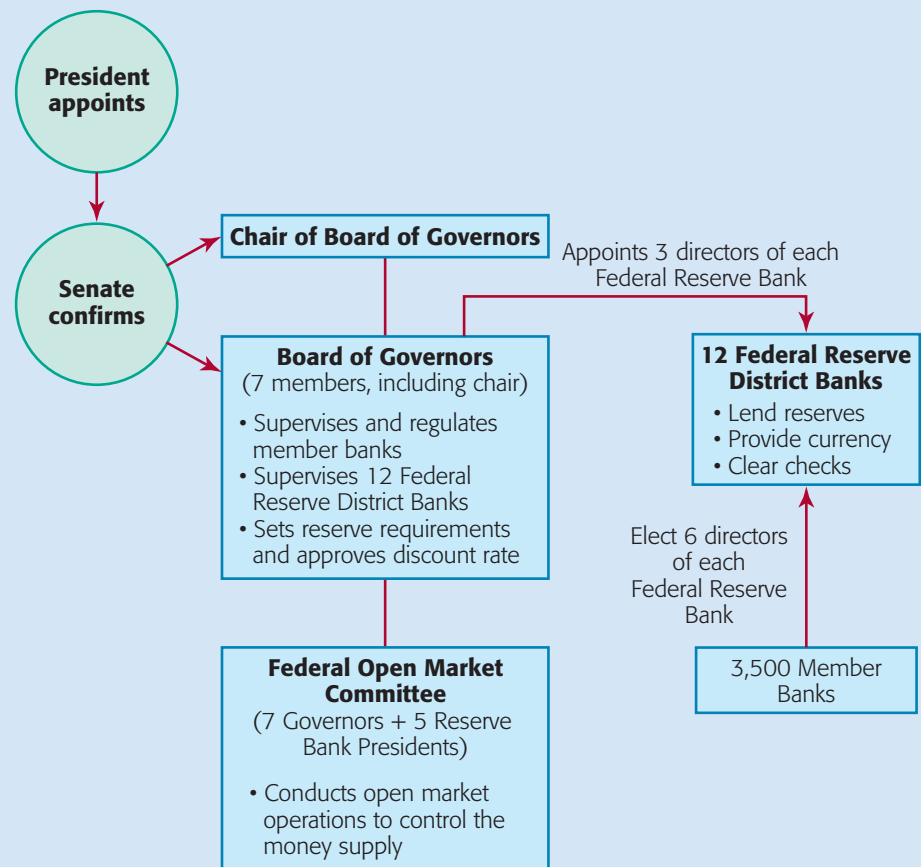


The Federal Open Market Committee meets in this room, inside the Fed’s headquarters in Washington DC. The meetings are highly secretive. No one from the media, and no one representing Congress or the President, is permitted in the room during the meetings.

FIGURE 3

Principal decision-making power at the Fed is vested in the Board of Governors, who are appointed by the president and confirmed by the Senate. Monetary policy is set by the Federal Open Market Committee, which consists of the 7 governors plus 5 of the presidents of Federal Reserve Banks.

THE STRUCTURE OF THE FEDERAL RESERVE SYSTEM



THE STRUCTURE OF THE FED

Figure 3 shows the organizational structure of the Federal Reserve System. Near the top is the Board of Governors, consisting of seven members who are appointed by the president and confirmed by the Senate for a 14-year term. The most powerful person at the Fed is the chairman of the Board of Governors—one of the seven governors who is appointed by the president, with Senate approval, to a four-year term as chair. In order to keep any president or Congress from having too much influence over the Fed, the four-year term of the chair is *not* coterminous with the four-year term of the president. As a result, every newly elected president inherits the Fed chair appointed by his predecessor, and may have to wait several years before making an appointment of his own.

Each of the 12 Federal Reserve Banks is supervised by nine directors, three of whom are appointed by the Board of Governors. The other six are elected by private commercial banks—the official stockholders of the system. The directors of each Federal Reserve Bank choose a president of that bank, who manages its day-to-day operations.

Notice that Figure 3 refers to “member banks.” Only about 3,500 of the 9,000 or so commercial banks in the United States are members of the Federal Reserve

System. But they include all 2,500 *national banks* (those chartered by the federal government) and about 1,000 *state banks* (chartered by their state governments). All of the largest banks in the United States (e.g., Citibank, Bank of America, and BankBoston) are nationally chartered banks and therefore member banks as well.

THE FEDERAL OPEN MARKET COMMITTEE

Finally, we come to what most economists regard as the most important part of the Fed—the **Federal Open Market Committee (FOMC)**. As you can see in Figure 3, the FOMC consists of all 7 governors of the Fed, along with 5 of the 12 district bank presidents.² The committee meets about eight times a year to discuss current trends in inflation, unemployment, output, interest rates, and international exchange rates. After determining the current state of the economy, the FOMC sets the general course for the nation’s money supply.

The word “open” in the FOMC’s name is ironic, since the committee’s deliberations are private. Summaries of its meetings are published only after a delay of a month or more. In some cases, the committee will release a brief public statement about its decisions on the day they are made. But not even the president of the United States knows the details behind the decisions, or what the FOMC actually discussed at its meeting, until the summary of the meeting is finally released. The reason for the word “open” is that the committee exerts control over the nation’s money supply by buying and selling bonds in the public (“open”) bond market. Later, we will discuss how and why the FOMC does this.

THE FUNCTIONS OF THE FEDERAL RESERVE

The Federal Reserve, as the overseer of the nation’s monetary system, has a variety of important responsibilities. Some of the most important are:

Supervising and Regulating Banks. We’ve already seen that the Fed sets and enforces reserve requirements, which all banks—not just Fed members—must obey. The Fed also sets standards for establishing new banks, determines what sorts of loans and investments banks are allowed to make, and closely monitors banks’ financial activities.

Acting as a “Bank for Banks.” Commercial banks use the Fed in much the same way that ordinary citizens use commercial banks. For example, we’ve already seen that banks hold most of their reserves in reserve accounts with the Fed. In addition, banks can borrow from the Fed, just as we can borrow from our local bank. The Fed charges a special interest rate, called the **discount rate**, on loans that it makes to member banks. In times of financial crisis, the Fed is prepared to act as *lender of last resort*, to make sure that banks have enough reserves to meet their obligations to depositors.

Issuing Paper Currency. The Fed doesn’t actually *print* currency; that is done by the government’s Bureau of Engraving and Printing. But once printed, it is shipped to the Fed (under *very* heavy guard). The Fed, in turn, puts this currency

Federal Open Market Committee

A committee of Federal Reserve officials that establishes U.S. monetary policy.

Discount rate The interest rate the Fed charges on loans to banks.

² Although all Reserve Bank presidents attend FOMC meetings, only 5 of the 12 presidents can vote on FOMC decisions. The president of the Federal Reserve Bank of New York has a permanent vote because New York is such an important financial center. But the remaining four votes rotate among the other district presidents.

into circulation. This is why every U.S. bill carries the label *Federal Reserve Note* on the top.

Check Clearing. Suppose you write a check for \$500 to pay your rent. Your building's owner will deposit the check into *his* checking account, which is probably at a different bank than yours. Somehow, your rent payment must be transferred from your bank account to your landlord's account at the other bank—a process called *check clearing*. In some cases, the services are provided by private clearinghouses. But in many other cases—especially for clearing out-of-town checks—the Federal Reserve system performs the service by transferring funds from one bank's reserve account to another's.

Controlling the Money Supply. The Fed, as the nation's monetary authority, is responsible for controlling the money supply. Since this function is so important in macroeconomics, we explore it in detail in the next section.

THE FED AND THE MONEY SUPPLY

Suppose the Fed wants to change the nation's money supply. (*Why* would the Fed want to do this? The answer will have to wait until the next chapter.) There are many ways this could be done. To increase the money supply, the Fed could print up currency and give it to Fed officials, letting them spend it as they wish. Or it could hold a lottery and give all of the newly printed money to the winner. To decrease the money supply, the Fed could require that all citizens turn over a portion of their cash to Fed officials who would then feed it into paper shredders.

These and other methods would certainly work, but they hardly seem fair or orderly. In practice, the Fed uses a more organized, less haphazard method to change the money supply: *open market operations*.

When the Fed wishes to increase or decrease the money supply, it buys or sells government bonds to bond dealers, banks, or other financial institutions. These actions are called open market operations.

Open market operations Purchases or sales of bonds by the Federal Reserve System.

We'll make two special assumptions to keep our analysis of open market operations simple for now.

1. Households and businesses are satisfied holding the amount of cash they are currently holding. Any additional funds they might acquire are deposited in their checking accounts. Any decrease in their funds comes from their checking accounts.
2. Banks never hold reserves in excess of those legally required by law.

Later, we'll discuss what happens when these simplifying assumptions do not hold. We'll also assume that the required reserve ratio is 0.1, so that each time deposits rise by \$1,000 at a bank, its required reserves rise by \$100.

HOW THE FED INCREASES THE MONEY SUPPLY

To increase the money supply, the Fed will *buy* government bonds. This is called an *open market purchase*. Suppose the Fed buys a government bond worth \$1,000 from Salomon Brothers, a bond dealer that has a checking account at First National

Bank.³ The Fed will pay Salomon Brothers with a \$1,000 check, which the firm will deposit into its account at First National. First National, in turn, will send the check to the Fed, which will credit First National's reserve account by \$1,000.

These actions will change First National's balance sheet as follows:

CHANGES IN FIRST NATIONAL BANK'S BALANCE SHEET

Action	Changes in Assets	Changes in Liabilities
Fed buys \$1,000 bond from Salomon Brothers, which deposits \$1,000 check from Fed into its checking account.	+\$1,000 in reserves	+\$1,000 in demand deposits

Notice that here we show only *changes* in First National's balance sheet. Other balance-sheet items—such as property and buildings, loans, government bonds, or net worth—are not immediately affected by the open market purchase, so they are not listed here. As you can see, First National gains an asset—reserves—so we enter “+\$1,000 in reserves” on the left side of the table. But there are also additional liabilities—the \$1,000 that is now in Salomon Brothers' checking account and which First National owes to that firm. The additional liabilities are represented by the entry “+\$1,000 in demand deposits” on the right side. Since First National's balance sheet was in balance before Salomon Brothers' deposit, and since assets and liabilities both grew by the same amount—\$1,000—we know that the balance sheet is still in balance. Total assets are again equal to total liabilities plus net worth.

Before we go on, let's take note of two important things that have happened. First, the Fed, by conducting an open market purchase, has injected *reserves* into the banking system. So far, these reserves are being held by First National, in its reserve account with the Fed.

The second thing to notice is something that is easy to miss: *The money supply has increased*. How do we know? Because demand deposits are part of the money supply, and they have increased by \$1,000. As you are about to see, even more demand deposits will be created before our story ends.

To see what will happen next, let's take the point of view of First National Bank's manager. He might reason as follows: “My demand deposits have just increased by \$1,000. Since the required reserve ratio is 0.1, I must now hold $0.1 \times \$1,000 = \100 in additional reserves. But my *actual* reserves have gone up by more than \$100; in fact, they have gone up by \$1,000. Therefore, I have **excess reserves**—reserves above those I'm legally required to hold—equal to $\$1,000 - \100 , or \$900. Since these excess reserves are earning no interest, I should lend them out.” Thus, we can expect First National, in its search for profit, to lend out \$900 at the going rate of interest.

How will First National actually make the loan? It could lend out \$900 in *cash* from its vault. It would be more typical, however, for the bank to issue a \$900 *check* to the borrower. When the borrower deposits the \$900 check into his own bank account (at some other bank), the Federal Reserve—which keeps track of these transactions for the banking system—will deduct \$900 from First National's reserve account and transfer it to the other bank's reserve account. This will cause a further change in First National's balance sheet, as follows:

Excess reserves Reserves in excess of required reserves.

³ We'll limit our analysis to commercial banks, which hold demand deposits, although our story would be similar if other types of depository institutions were involved.

CHANGES IN FIRST NATIONAL BANK'S BALANCE SHEET

Action	Changes in Assets	Changes in Liabilities
Fed buys \$1,000 bond from Salomon Brothers, which deposits \$1,000 check from Fed into its checking account.	+\$1,000 in reserves	+\$1,000 in demand deposits
First National lends out \$900 in excess reserves.	-\$900 in reserves +\$900 in loans	
The total effect on First National from beginning to end.	+\$100 in reserves +\$900 in loans	+\$1,000 in demand deposits

Look at the boldface entries in the table. By making the loan, First National has given up an asset—\$900 in reserves. This causes assets to change by $-\$900$. But First National also gains an asset of equal value—the \$900 loan. (Remember: While loans are liabilities to the borrower, they are assets to banks.) This causes assets to change by $+\$900$. Both of these changes are seen on the assets side of the balance sheet.

Now look at the bottom row of the table. This tells us what has happened to First National from beginning to end. We see that, after making its loan, First National has \$100 more in reserves than it started with, and \$900 more in loans, for a total of \$1,000 more in assets. But it also has \$1,000 more in liabilities than it had before—the additional demand deposits that it owes to Salomon Brothers. Both assets and liabilities have gone up by the same amount. Notice, too, that First National is once again holding exactly the reserves it must legally hold. It now has \$1,000 more in demand deposits than it had before, and it is holding $0.1 \times \$1,000 = \100 more in reserves than before. First National is finished (“loaned up”) and cannot lend out any more reserves.

But there is still more to our story. Let’s suppose that First National lends the \$900 to the owner of a local business, Paula’s Pizza, and that Paula deposits her loan check into *her* bank account at Second Federal Bank. Then, remembering that the Fed will transfer \$900 in reserves from First National’s reserve account to that of Second Federal, we’ll see the following changes in Second Federal’s balance sheet:

CHANGES IN SECOND FEDERAL'S BALANCE SHEET

Action	Changes in Assets	Changes in Liabilities
Paula deposits \$900 loan check into her checking account.	+\$900 in reserves	+\$900 in demand deposits

Second Federal now has \$900 more in assets—the increase in its reserve account with the Federal Reserve—and \$900 in additional liabilities—the amount added to Paula’s checking account.

Now consider Second Federal’s situation from its manager’s viewpoint. He reasons as follows: “My demand deposits have risen by \$900, which means my required reserves have risen by $0.1 \times \$900 = \90 . But my reserves have *actually* increased by \$900. Thus, I have *excess reserves* of $\$900 - \$90 = \$810$, which I will lend out.” After making the \$810 loan, Second Federal’s balance sheet will change once again (look at the boldface entries):

CHANGES IN SECOND FEDERAL'S BALANCE SHEET

Action	Changes in Assets	Changes in Liabilities
Paula deposits \$900 loan check into her checking account.	+\$900 in reserves	+\$900 in demand deposits
Second Federal lends out \$810 in excess reserves.	–\$810 in reserves +\$810 in loans	
The total effect on Second Federal from beginning to end.	+\$ 90 in reserves +\$810 in loans	+\$900 in demand deposits

In the end, as you can see in the bottom row of the table, Second Federal has \$90 more in reserves than it started with, and \$810 more in loans. Its demand deposit liabilities have increased by \$900. Notice, too, that the money supply has increased once again—this time, by \$900.

Are you starting to see a pattern? Let's carry it through one more step. Whoever borrowed the \$810 from Second Federal will put it into his or her checking account at, say, Third State Bank. This will give Third State excess reserves that it will lend out. As a result, its balance sheet will change as shown.

CHANGES IN THIRD STATE'S BALANCE SHEET

Action	Changes in Assets	Changes in Liabilities
Borrower from Second Federal deposits \$810 loan check into checking account.	+\$810 in reserves	+\$810 in demand deposits
Third State lends out \$729 in excess reserves.	–\$729 in reserves +\$729 in loans	
The total effect on Third State from beginning to end.	+\$ 81 in reserves +\$729 in loans	+\$810 in demand deposits

As you can see, demand deposits increase each time a bank lends out excess reserves. In the end, they will increase by a *multiple* of the original \$1,000 in reserves injected into the banking system by the open market purchase. Does this process sound familiar? It should. It is very similar to the explanation of the *expenditure multiplier* in the previous chapter, where in each round, an increase in spending led to an increase in income, which caused spending to increase again in the next round. Here, instead of spending, it is the *money supply*—or more specifically, *demand deposits*—that increase in each round.

THE DEMAND DEPOSIT MULTIPLIER

By how much will demand deposits increase in total? If you look back at the balance sheet changes we've analyzed, you'll see that each bank creates less in demand deposits than the bank before. When Salomon Brothers deposited its \$1,000 check from the Fed at First National, \$1,000 in demand deposits was created. This led to an additional \$900 in demand deposits created by Second Federal, another \$810 created by Third State, and so on. In each round, a bank lent 90 percent of the deposit it received. Eventually the additional demand deposits will become so small that we can safely ignore them. When the process is complete, how much in additional demand deposits have been created?

TABLE 2

CUMULATIVE INCREASES IN DEMAND DEPOSITS AFTER A \$1,000 CASH DEPOSIT

Round	Additional Demand Deposits Created by This Bank	Additional Demand Deposits Created by All Banks
First National Bank	\$1,000	\$ 1,000
Second Federal	\$ 900	\$ 1,900
Third State	\$ 810	\$ 2,710
Bank 4	\$ 729	\$ 3,439
Bank 5	\$ 656	\$ 4,095
Bank 6	\$ 590	\$ 4,685
Bank 7	\$ 531	\$ 5,216
Bank 8	\$ 478	\$ 5,694
Bank 9	\$ 430	\$ 6,124
Bank 10	\$ 387	\$ 6,511
Bank 11	\$ 349	\$ 6,860
Bank 12	\$ 314	\$ 7,174
...		
All Other Banks	very close to \$2,826	
Total		\$10,000

Table 2 provides the answer. Each row of the table shows the additional demand deposits created at each bank, as well as the running total. The last row shows that, in the end, \$10,000 in new demand deposits has been created.

Let's go back and summarize what happened in our example. The Fed, through its open market purchase, injected \$1,000 of reserves into the banking system. As a result, demand deposits rose by \$10,000—10 times the injection in reserves. As you can verify, if the Fed had injected twice the amount of reserves (\$2,000), demand deposits would have increased by 10 times *that* amount (\$20,000). In fact, *whatever* the injection of reserves, demand deposits will increase by a factor of 10, so we can write

$$\Delta DD = 10 \times \text{reserve injection}$$

where “DD” stands for demand deposits. The injection of reserves must be *multiplied* by the number 10 in order to get the change in demand deposits that it causes. For this reason, 10 is called the *demand deposit multiplier* in this example.

Demand deposit multiplier The number by which a change in reserves is multiplied to determine the resulting change in demand deposits.

The demand deposit multiplier is the number by which we must multiply the injection of reserves to get the total change in demand deposits.

The size of the demand deposit multiplier depends on the value of the required reserve ratio set by the Fed. If you look back at Table 2, you will see that each round of additional deposit creation would have been smaller if the required reserve ratio had been larger. For example, with a required reserve ratio of 0.2 instead of 0.1, Second Federal would have created only \$800 in deposits, Third State would have created only \$640, and so on. The result would have been a smaller cumulative change in deposits, and a smaller multiplier.

Now let's derive the formula we can use to determine the demand deposit multiplier for *any* required reserve ratio. We'll start with our example in which the re-

quired reserve ratio is 0.1. If \$1,000 in reserves is injected into the system, the total change in deposits can be written as follows:

$$\Delta DD = \$1,000 + \$900 + \$810 + \$729 + \dots$$

Factoring out \$1,000, this becomes

$$\Delta DD = \$1,000 \times [1 + 0.9 + 0.9^2 + 0.9^3 + \dots]$$

In this equation, \$1,000 is the initial injection of reserves ($\Delta \text{reserves}$), and 0.9 is the fraction of reserves that each bank loans out, which is 1 minus the required reserve ratio ($1 - 0.1 = 0.9$). To find the change in deposits that applies to *any* change in reserves and *any* required reserve ratio (RRR), we can write

$$\Delta DD = \Delta \text{Reserves} \times [1 + (1 - RRR) + (1 - RRR)^2 + (1 - RRR)^3 + \dots]$$

Now we can see that the term in brackets—the infinite sum $1 + (1 - RRR) + (1 - RRR)^2 + (1 - RRR)^3 + \dots$ —is our demand deposit multiplier. But what is its value?

Recall from the last chapter that an infinite sum

$$1 + H + H^2 + H^3 + \dots$$

always has the value $1/(1 - H)$ as long as H is a fraction between zero and 1. In the last chapter, we replaced H with the MPC to get the expenditure multiplier. But here, we will replace H with $1 - RRR$ (which is always between zero and 1) to obtain a value for the deposit multiplier of $1/[1 - (1 - RRR)] = 1/RRR$.

For any value of the required reserve ratio (RRR), the formula for the demand deposit multiplier is $1/RRR$.

In our example, the RRR was equal to 0.1, so the deposit multiplier had the value $1/0.1 = 10$. If the RRR had been 0.2 instead, the deposit multiplier would have been equal to $1/0.2 = 5$.

Using our general formula for the demand deposit multiplier, we can restate what happens when the Fed injects reserves into the banking system as follows:

$$\Delta DD = \left(\frac{1}{RRR} \right) \times \Delta \text{Reserves.}$$

Since we've been assuming that the amount of cash in the hands of the public (the other component of the money supply) does not change, we can also write

$$\Delta \text{Money Supply} = \left(\frac{1}{RRR} \right) \times \Delta \text{Reserves.}$$

THE FED'S INFLUENCE ON THE BANKING SYSTEM AS A WHOLE

We can also look at what happened to total demand deposits and the money supply from another perspective. When the Fed bought the \$1,000 bond from Salomon Brothers, it injected \$1,000 of reserves into the banking system. That was the only increase in reserves that occurred in our story. Where did the additional \$1,000 in

reserves end up? If you go back through the changes in balance sheets, you'll see that First National ended up with \$100 in additional reserves, Second Federal ended up with \$90, Third Savings with \$81, and so on. Each of these banks is required to hold more reserves than initially, because its demand deposits have increased. In the end, *the additional \$1,000 in reserves will be distributed among different banks in the system as required reserves.*

After an injection of reserves, the demand deposit multiplier stops working—and the money supply stops increasing—only when all the reserves injected are being held by banks as required reserves.

This observation helps us understand the demand deposit multiplier in another way. In our example, the deposit-creation process will continue until the entire injection of \$1,000 in reserves becomes *required* reserves. But with a *RRR* of 0.1, each dollar of reserves entitles a bank to have \$10 in demand deposits. Therefore, by injecting \$1,000 of reserves into the system, the Fed has enabled banks, in total, to hold \$10,000 in additional demand deposits. Only when \$10,000 in deposits has been created will the process come to an end.

Just as we've looked at balance sheet changes for each bank, we can also look at the change in the balance sheet of the *entire banking system*. The Fed's open market purchase of \$1,000 has caused the following changes:

CHANGES IN THE BALANCE SHEET OF THE ENTIRE BANKING SYSTEM

Changes in Assets	Changes in Liabilities
+\$1,000 in reserves +\$9,000 in loans	+\$10,000 in demand deposits

In the end, total reserves in the system have increased by \$1,000—the amount of the open market purchase. Each dollar in reserves supports \$10 in demand deposits, so we know that total deposits have increased by \$10,000. Finally, we know that a balance sheet always balances. Since liabilities increased by \$10,000, loans must have increased by \$9,000 to increase total assets (loans and reserves) by \$10,000.



Demand deposits are a means of payment, and banks create them. This is why we say that banks “create deposits” and “create money.” But don't fall into the trap of thinking that banks create *wealth*. No one gains any additional wealth as a result of money creation.

To see why, think about what happened in our story when Salomon Brothers deposited the \$1,000 check from the Fed into its account at First National. *Salomon Brothers* was no wealthier: It gave up a \$1,000 check from the Fed and ended up with \$1,000 more in its checking account, for a net gain of zero. Similarly, the *bank* gained no additional wealth: It had \$1,000 more in cash, but it also *owed* Salomon Brothers \$1,000—once again, a net gain of zero.

The same conclusion holds for any other step in the money-creation process. When Paula borrows \$900 and deposits it into her checking account at Second Federal, she is no wealthier: She has \$900 more in her account, but owes \$900 to First National. And once again, the bank is no wealthier: It has \$900 more in demand deposits, but owes this money to Paula.

Always remember that while banks can “create money,” they cannot create wealth.

HOW THE FED DECREASES THE MONEY SUPPLY

Just as the Fed can increase the money supply by purchasing government bonds, it can also *decrease* the money supply by *selling* government bonds—an *open market sale*.

Where does the Fed get the government bonds to sell? It has trillions of dollars' worth of government bonds from open market *purchases* it has conducted in the past. Since, on average, the Fed tends to increase the money supply

each year, it conducts more open market purchases than open market sales, and its stock of bonds keeps growing. So we needn't worry that the Fed will run out of bonds to sell.

Suppose the Fed sells a \$1,000 government bond to a bond dealer, Merrill Lynch, which—like Salomon Brothers in our earlier example—has a checking account at First National Bank. Merrill Lynch pays the Fed for the bond with a \$1,000 check drawn on its account at First National. When the Fed gets Merrill Lynch's check, it will present the check to First National and deduct \$1,000 from First National's reserve account. In turn, First National will deduct \$1,000 from Merrill Lynch's checking account.

After all of this has taken place, First National's balance sheet will show the following changes:

CHANGES IN FIRST NATIONAL BANK'S BALANCE SHEET

Action	Changes in Assets	Changes in Liabilities
Fed sells \$1,000 bond to Merrill Lynch, which pays with a \$1,000 check drawn on First National.	−\$1,000 in reserves	−\$1,000 in demand deposits

Now First National has a problem. Since its demand deposits have decreased by \$1,000, it can legally decrease its reserves by 10 percent of that, or \$100. But its reserves have *actually* decreased by \$1,000, which is \$900 more than they are allowed to decrease. First National has *deficient reserves*—reserves smaller than those it is legally required to hold. How can it get the additional reserves it needs?

First National will have to *call in a loan*—that is, ask for repayment—in the amount of \$900.⁴ A loan is usually repaid with a check drawn on some other bank. When First National gets this check, the Federal Reserve will add \$900 to its reserve account, and deduct \$900 from the reserve account at the other bank. This is how First National brings its reserves up to the legal requirement. After it calls in the \$900 loan, First National's balance sheet will change as follows:

CHANGES IN FIRST NATIONAL BANK'S BALANCE SHEET

Action	Changes in Assets	Changes in Liabilities
Fed sells \$1,000 bond to Merrill Lynch, which pays with a \$1,000 check drawn on First National.	−\$1,000 in reserves	−\$1,000 in demand deposits
First National calls in loans worth \$900.	+\$ 900 in reserves −\$ 900 in loans	
The total effect on First National from beginning to end.	−\$ 100 in reserves −\$ 900 in loans	−\$1,000 in demand deposits

Look at the boldfaced terms. After First National calls in the loan, the composition of its assets will change: \$900 more in reserves, and \$900 less in loans. The

⁴ In reality, bank loans are for specified time periods, and a bank cannot actually demand that a loan be repaid early. But most banks have a large volume of loans outstanding, with some being repaid each day. Typically, the funds will be lent out again the very same day they are repaid. But a bank that needs additional reserves will simply reduce its rate of new lending on that day, thereby reducing its total amount of loans outstanding. This has the same effect as “calling in a loan.”

last row of the table shows the changes to First National's balance sheet from beginning to end. Compared to its initial situation, First National has \$100 less in reserves (it lost \$1,000 and then gained \$900), \$900 less in loans, and \$1,000 less in demand deposits.

As you might guess, this is not the end of the story. Remember that whoever paid back the loan to First National did so by a check drawn on another bank. That other bank, which we'll call Second United Bank, will lose \$900 in reserves and experience the following changes in its balance sheet:

CHANGES IN SECOND UNITED BANK'S BALANCE SHEET

Action	Changes in Assets	Changes in Liabilities
Someone with an account at Second United Bank writes a \$900 check to First National.	-\$900 in reserves	-\$900 in demand deposits

Now Second United Bank is in the same fix that First National was in. Its demand deposits have decreased by \$900, so its reserves can legally fall by \$90. However, its actual reserves have decreased by \$900—which is \$810 too much. Now it is Second United's turn to call in a loan. (On your own, fill in the rest of the changes in Second United Bank's balance sheet as it successfully brings its reserves up to the legal requirement.)

As you can see, the process of calling in loans will involve many banks. Each time a bank calls in a loan, demand deposits are destroyed—the same amount as were created in our earlier story, in which each bank *made* a new loan. The total decline in demand deposits will be a multiple of the initial withdrawal of reserves. Keeping in mind that a withdrawal of reserves is a *negative change in reserves*, we can still use our demand deposit multiplier— $1/(RRR)$ —and our general formula:

$$\Delta DD = \left(\frac{1}{RRR} \right) \times \Delta \text{Reserves.}$$

Applying it to our example, we have

$$\Delta DD = \left[\frac{1}{0.1} \right] \times (-\$1,000) = -\$10,000.$$

In words, the Fed's \$1,000 open market sale causes a \$10,000 decrease in demand deposits. Since we assume that the public's cash holdings do not change, the money supply decreases by \$10,000 as well.

To the banking system as a whole, the Fed's bond sale has done the following:

CHANGES IN BALANCE SHEET FOR THE ENTIRE BANKING SYSTEM

Changes in Assets	Changes in Liabilities
-\$1,000 in reserves -\$9,000 in loans	-\$10,000 in demand deposits

SOME IMPORTANT PROVISOS ABOUT THE DEMAND DEPOSIT MULTIPLIER

Although the process of money creation and destruction as we've described it illustrates the basic ideas, our formula for the demand deposit multiplier— $1/RRR$ —is

oversimplified. In reality, the multiplier is likely to be smaller than our formula suggests, for two reasons.

First, we've assumed that as the money supply changes, the public does *not* change its holdings of cash. But in reality, as the money supply increases, the public typically will want to hold part of the increase as demand deposits, and part of the increase as cash. As a result, in each round of the deposit-creation process, some reserves will be *withdrawn* in the form of cash. This will lead to a smaller increase in demand deposits than in our story.

Second, we've assumed that banks will always lend out all of their excess reserves. In reality, banks often *want* to hold excess reserves, for a variety of reasons. For example, they may want some flexibility to increase their loans in case interest rates—their reward for lending—rise in the near future. Or they may prefer not to lend the maximum legal amount during a recession, because borrowers are more likely to declare bankruptcy and not repay their loans. If banks increase their holdings of excess reserves as the money supply expands, they will make smaller loans than in our story, and in each round, demand deposit creation will be smaller.



In this section, you learned how the Fed sells government bonds to decrease the money supply. It's easy to confuse this with another type of government bond sale, which is done by the U.S. Treasury.

The U.S. Treasury is the branch of government that collects tax revenue, disburses money for government purchases and transfer payments, and borrows money to finance any government budget deficit. The Treasury borrows funds by issuing *new* government bonds and *selling* them to the public—to banks, other financial institutions, and bond dealers. What the public pays for these bonds is what they are lending the government.

When the Fed conducts open market operations, however, it does not buy or sell *newly* issued bonds, but “secondhand bonds”—those already issued by the Treasury to finance past deficits. Thus, open market sales are *not* government borrowing; they are strictly an operation designed to change the money supply, and they have no direct effect on the government budget.

OTHER TOOLS FOR CONTROLLING THE MONEY SUPPLY

Open market operations are the Fed's primary means of controlling the money supply. But there are two other tools that the Fed can use to increase or decrease the money supply.

- *Changes in the required reserve ratio.* In principle, the Fed can set off the process of deposit creation, similar to that described earlier, by lowering the required reserve ratio. Look back at Table 1, which showed the balance sheet of a bank facing a required reserve ratio of 0.1 and holding exactly the amount of reserves required by law—\$10 million. Now suppose the Fed lowered the required reserve ratio to 0.05. Suddenly, the bank would find that its required reserves were only \$5 million; the other \$5 million in reserves it holds would become excess reserves. To earn the highest profit possible, the bank would increase its lending by \$5 million. At the same time, all other banks in the country would find that some of their formerly required reserves were now excess reserves, and they would increase their lending. The money supply would increase.

On the other hand, if the Fed *raised* the required reserve ratio, the process would work in reverse: All banks would suddenly have reserve deficiencies and be forced to call in loans. The money supply would decrease.

- *Changes in the discount rate.* The discount rate, mentioned earlier, is the rate the Fed charges banks when it lends them reserves. In principle, a lower discount rate—enabling banks to borrow reserves from the Fed more cheaply—might encourage banks to borrow more. An increase in borrowed reserves works just like any other injection of reserves into the banking system: It increases the money supply.

On the other side, a rise in the discount rate would make it more expensive for banks to borrow from the Fed, and decrease the amount of borrowed reserves in the system. This withdrawal of reserves from the banking system would lead to a decrease in the money supply.

Changes in either the required reserve ratio or the discount rate *could* set off the process of deposit creation or deposit destruction in much the same way outlined in this chapter. In reality, neither of these policy tools is used very often. The most recent change in the required reserve ratio was in April 1992, when the Fed lowered the required reserve ratio for most demand deposits from 12 percent to 10 percent. Changes in the discount rate are more frequent, but it is not unusual for the Fed to leave the discount rate unchanged for a year or more.

Why are these other tools used so seldom? Part of the reason is that they can have such unpredictable effects. When the required reserve ratio changes, all banks in the system are affected simultaneously. Even a tiny error in predicting how a typical bank will respond can translate into a huge difference for the money supply.

A change in the discount rate has uncertain effects as well. Many bank managers do not like to borrow reserves from the Fed, since it puts them under closer Fed scrutiny. And the Fed discourages borrowing of reserves unless the bank is in difficulty. Thus, a small change in the discount rate is unlikely to have much of an impact on bank borrowing of reserves, and therefore on the money supply.

Open market operations, by contrast, have more predictable impacts on the money supply. They can be fine-tuned to any level desired. Another advantage is that they are covert. No one knows exactly what the FOMC decided to do to the money supply at its last meeting. And no one knows whether it is conducting more open market purchases or more open market sales on any given day (it always does a certain amount of both to keep bond traders guessing). By maintaining secrecy, the Fed can often change its policies without destabilizing financial markets, and also avoid the pressure that Congress or the president might bring to bear if its policies are not popular.

While other tools can affect the money supply, open market operations have two advantages over them: precision and secrecy. This is why open market operations remain the Fed's primary means of changing the money supply.

The Fed's ability to conduct its policies in secret—and its independent status in general—is controversial. Some argue that secrecy and independence are needed so that the Fed can do what is best for the country—keeping the price level stable—without undue pressure from Congress or the president. Others argue that there is something fundamentally undemocratic about an independent Federal Reserve, whose governors are not elected and who can, to some extent, ignore the popular will. In recent years, because the Fed has been so successful in guiding the economy, the controversy has largely subsided.

BANK FAILURES AND BANKING PANICS



A bank failure occurs when a bank is unable to meet the requests of its depositors to withdraw their funds. Typically, the failure occurs when depositors begin to worry about the bank's financial health. They may believe that their bank has made unsound loans that will not be repaid, so that it does not have enough assets to cover its demand deposit liabilities. In that case, everyone will want to be first in

line to withdraw cash, since banks meet requests for withdrawals on a first-come, first-served basis. Those who wait may not be able to get any cash at all. This can lead to a **run on the bank**, with everyone trying to withdraw funds simultaneously.

Ironically, a bank can fail even if it is in good financial health, with more than enough assets to cover its liabilities, just because people *think* the bank is in trouble. Why should a false rumor be a problem for the bank? Because many of its assets are illiquid, such as long-term loans. These cannot be sold easily or quickly enough to meet the unusual demands for withdrawal during a run on the bank.

For example, look back at Table 1, which shows a healthy bank with more assets than liabilities. But notice that the bank has only \$2 million in vault cash. Under normal circumstances, that would be more than enough to cover a day of heavy withdrawals. But suppose that depositors hear a rumor that the bank has made many bad loans, and they want to withdraw \$40 million. The bank would soon exhaust its \$2 million in cash. It could then ask the Federal Reserve for more cash, using the \$8 million in its reserve account, and the Fed would likely respond quickly, perhaps even delivering the cash the same day. The bank could also sell its \$25 million in government bonds and obtain more cash within a few days. But all together, this will give the bank only \$35 million with which to honor requests for withdrawals. What then? Unless the bank is lucky enough to have many of its long-term loans coming due that week, it will be unable to meet its depositors' requests for cash. A false rumor can cause a bank to fail.

A **banking panic** occurs when many banks fail simultaneously. In the past, a typical panic would begin with some unexpected event, such as the failure of a large bank. During recessions, for example, many businesses go bankrupt, so fewer bank loans are repaid. A bank that had an unusual number of “bad loans” would be in trouble, and if the public found out about this, there might be a run on that bank. The bank would fail, and many depositors would find that they had lost their deposits.

But that would not be the end of the story. Hearing that their neighbors' banks were short of cash might lead others to question the health of their own banks. Just to be sure, they might withdraw their own funds, preferring to ride out the storm and keep their cash at home. As we've seen, even healthy banks can fail under the pressure of a bank run. They, too, would have to close their doors, stoking the rumor mill even more, and so on.

Banking panics can cause serious problems for the nation. First, there is the hardship suffered by people who lose their accounts when their bank fails. Second, even when banks do not fail, the withdrawal of cash decreases the banking system's reserves. As we've seen, the withdrawal of reserves leads—through the demand deposit multiplier—to a larger decrease in the money supply. In the next chapter, you will learn that a decrease in the money supply can cause a recession. In a banking panic, the money supply can decrease suddenly and severely, causing a serious recession.

There were five major banking panics in the United States from 1863 to 1907. Indeed, it was the banking panic of 1907 that convinced Congress to establish the Federal Reserve System. From the beginning, one of the Fed's primary functions was to act as a lender of last resort, providing banks with enough cash to meet their obligations to depositors.

But the creation of the Fed did not, in itself, solve the problem. Figure 4 shows the number of bank failures each year since 1921. As you can see, banking panics continued to plague the financial system even after the Fed was created. The Fed did

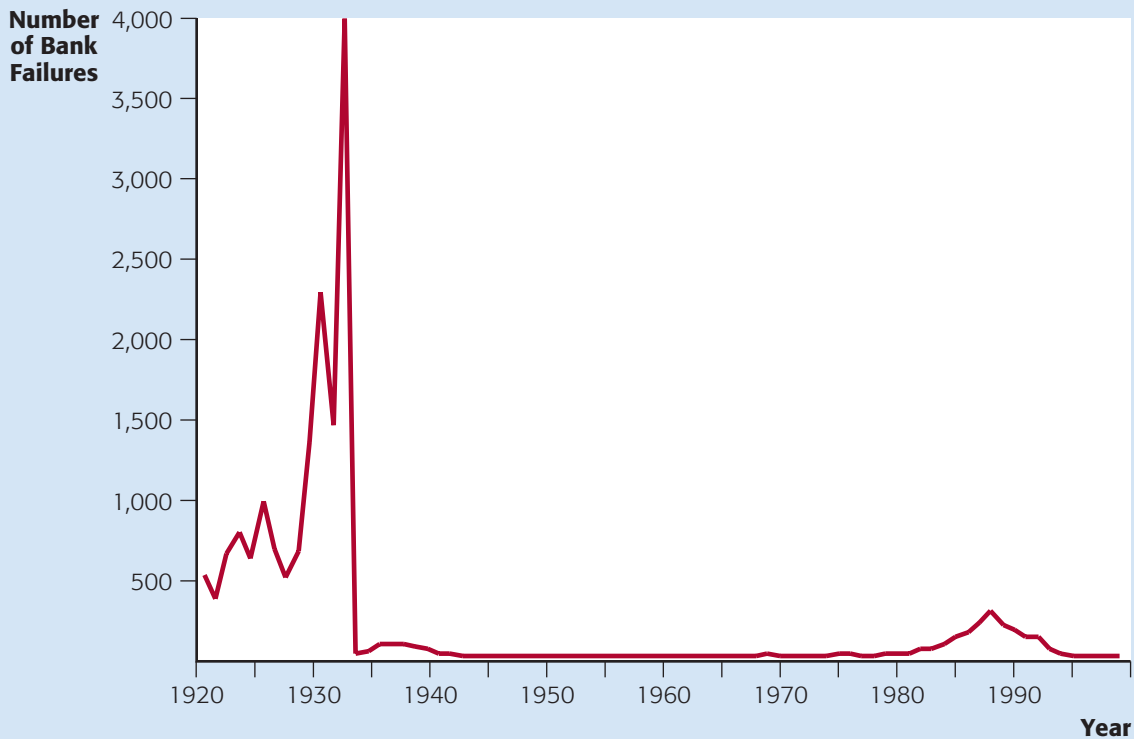


Run on the bank An attempt by many of a bank's depositors to withdraw their funds.

Banking panic A situation in which depositors attempt to withdraw funds from many banks simultaneously.

FIGURE 4

BANK FAILURES IN THE UNITED STATES, 1921–1999



Bank failures continued after the Fed was created in 1913. During the Great Depression, a large number of banks failed. The Fed learned a hard lesson: It needed to inject reserves into the banking system whenever a crisis threatened. The creation of the Federal Deposit Insurance Corporation in 1933 also strengthened faith in the stability of the banking system. Few banks have failed since that time.

not always act forcefully enough or quickly enough to prevent the panic from spreading.

The Great Depression is a good example of this problem. In late 1929 and 1930, many banks began to fail because of bad loans. Then, from October 1930 until March 1933, more than one-third of all banks failed as frantic depositors stormed bank after bank, demanding to withdraw their funds—even from banks that were in reasonable financial health. Many economists believe that the banking panic of 1930–1933 turned what would have been just a serious recession into the Great Depression. Officials of the Federal Reserve System, not quite grasping the seriousness of the problem, stood by and let it happen.⁵

As you can see in Figure 4, banking panics were largely eliminated after 1933. Indeed, except for the moderate increase in failures during the late 1980s and early 1990s, the system has been almost failure free. Why the dramatic improvement?

Largely for two reasons. First, the Federal Reserve learned an important lesson from the Great Depression, and it now stands ready to inject reserves into the system more quickly in a crisis. Moreover, in 1933 Congress created the Federal De-

⁵ Milton Friedman and Anna Jacobson Schwartz, *A Monetary History of the United States, 1867–1960* (Princeton University Press, 1963), especially p. 358.

posit Insurance Corporation (FDIC) to reimburse those who lose their deposits. If your bank is insured by the FDIC (today, accounts are covered in 99 percent of all banks) and cannot honor its obligations for any reason—bad loans, poor management, or even theft—the FDIC will reimburse you up to the first \$100,000 you lose in each of your bank accounts. (If you have more than \$100,000 in a single bank account, you are not insured for the amount over \$100,000.)

The FDIC has had a major impact on the psychology of the banking public. Imagine that you hear your bank is about to go under. As long as you have less than \$100,000 in your account, you will not care. Why? Because even if the rumor turns out to be true, you will be reimbursed in full. The resulting calmness on your part, and on the part of other depositors, will prevent a run on the bank. This makes it very unlikely that bank failures will spread throughout the system.

FDIC protection for bank accounts has not been costless. Banks must pay insurance premiums to the FDIC, and they pass this cost on to their depositors and borrowers by charging higher interest rates on loans and higher fees for their services. And there is a more serious cost. If you are thoroughly protected in the event of a bank failure, your bank's managers have little incentive to develop a reputation for prudence in lending funds, since you will be happy to deposit your money there anyway. Without government regulations, banks could act irresponsibly, taking great risks with your money, and you would remain indifferent. Many more banks would fail, the FDIC would have to pay off more depositors, and banks—and their customers—would bear the burden of higher FDIC premiums. This is the logic behind the Fed's continuing regulation of bank lending. Someone must watch over the banks to keep the failure rate low, and if the public has no incentive to pay attention, the Fed must do so. Most economists believe that if we want the freedom from banking panics provided by the FDIC, we must also accept the strict regulation and close monitoring of banks provided by the Fed and other agencies.

Look again at Figure 4 and notice the temporary rise in bank failures of the late 1980s and the early 1990s. Most of these failures occurred in state-chartered banks. These banks are less closely regulated by the Fed, and are often insured by state agencies instead of the FDIC. When a few banks went bankrupt because highly speculative loans turned sour, insurance funds in several states were drained. Citizens in those states began to fear that insufficient funds were left to insure their own deposits, and the psychology of banking panics took over. To many observers, the experience of the late 1980s and early 1990s was a reminder of the need for a sound insurance system and close monitoring of the banking system.



Fred Furlong and Simon Kwan, in "Rising Bank Risk?" (<http://www.frbsf.org/econsrch/wklyltr/wklyltr99/e199-32.html>) explore recent developments in bank behavior.

S U M M A R Y

In the United States, the standard measure of money—M1—includes currency, checking account balances, and travelers checks. Each of these assets is liquid and widely acceptable as a means of payment. Other, broader measures go beyond M1 to include funds in savings accounts and time deposits.

The amount of money circulating in the economy is controlled by the Federal Reserve, operating through the banking system. Banks and other financial intermediaries are profit-seeking firms that collect loanable funds from households and businesses, then repackage them to make loans to other households, businesses, and governmental agencies,

The Federal Reserve injects money into the economy by altering banks' balance sheets. In a balance sheet, assets al-

ways equal liabilities plus net worth. One important kind of asset is *reserves*—funds that banks are required to hold in proportion to their demand deposit liabilities. When the Fed wants to increase the money supply, it buys bonds in the open market and pays for them with a check. This is called an *open market purchase*. When the Fed's check is deposited in a bank, the bank's balance sheet changes. On the asset side, reserves increase; on the liabilities side, demand deposits (a form of money) also increase. The bank can lend some of the reserves, and the money loaned will end up in some other banks where it supports creation of still more demand deposits. Eventually, demand deposits, and the M1 money supply, increase by some multiple of the original injection of reserves by the Fed. The

demand deposit multiplier—the inverse of the required reserve ratio—gives us that multiple.

The Fed can decrease the money supply by selling government bonds—an *open market sale*—causing demand deposits

to shrink by a multiple of the initial reduction in reserves. The Fed can also change the money supply by changing either the required reserve ratio or the discount rate it charges when it lends reserves to banks.

KEY TERMS

liquidity	financial intermediary	required reserve ratio	open market operations
cash in the hands of the public	balance sheet	net worth	excess reserves
demand deposits	bond	central bank	demand deposit multiplier
M1	loan	Federal Open Market Committee	run on the bank
M2	reserves	discount rate	banking panic
	required reserves		

REVIEW QUESTIONS

- Describe the main characteristics of money. What purpose does money serve in present-day economies?
- Which of the following is considered part of the U.S. money supply?
 - A \$10 bill you carry in your wallet
 - A \$100 travelers check you bought but did not use
 - A \$100 bill in a bank teller's till
 - The \$325.43 balance in your checking account
 - A share of General Motors stock worth \$40
- Given the following data, calculate the value of the M1 money supply (the data are in billions of dollars):

Bank reserves	50
Cash in the hands of the public	400
Demand deposits	400
Noninstitutional MMMF balances	880
Other checkable deposits	250
Savings-type account balances	1,300
Small time deposits	950
Travelers checks	10
- What is a depository institution? Give an example of each of the four types of depository institutions.
- What are reserves? What determines the amount of reserves that a bank holds? Explain the difference between required reserves and excess reserves.
- What are the main functions of the Federal Reserve System?
- Explain how the Federal Reserve can use open market operations to change the level of bank reserves. How does a change in reserves affect the money supply? (Give answers for both an increase and a decrease in the money supply.)
- Suppose that the money supply is \$1 trillion. Decision makers at the Federal Reserve decide that they wish to reduce the money supply by \$100 billion, or by 10 percent. If the required reserve ratio is 0.05, what does the Fed need to do to carry out the planned reduction?
- How does a “run on a bank” differ from a “banking panic”? What are their implications for the economy? What steps have been taken to reduce the likelihood of bank runs and bank panics?

PROBLEMS AND EXERCISES

- Suppose the required reserve ratio is 0.2. If an extra \$20 billion in reserves is injected into the banking system through an open market purchase of bonds, by how much can demand deposits increase? Would your answer be different if the required reserve ratio were 0.1?
- Suppose bank reserves are \$100 billion, the required reserve ratio is 0.2, and excess reserves are zero. Now suppose that the required reserve ratio is lowered to 0.1 and
 - that banks once again become fully “loaned up” with no excess reserves. What is the new level of demand deposits?
- For each of the following situations, determine whether the money supply will increase, decrease, or stay the same.
 - Depositors become concerned about the safety of depository institutions.
 - The Fed lowers the required reserve ratio.

- c. The economy enters a recession and banks have a hard time finding credit-worthy borrowers.
 - d. The Fed sells \$100 million of bonds to First National Bank of Ames, Iowa.
4. Suppose that the Fed decides to increase the money supply. It purchases a government bond worth \$1,000 from a private citizen. He deposits the check in his account at First National Bank, as in the chapter example. But now, suppose that the required reserve ratio is 0.2, rather than 0.1 as in the chapter.
- a. Trace the effect of this change through three banks—First National, Second Federal, and Third State. Show the changes to each bank's balance sheet as a result of the Fed's action.
 - b. By how much does the money supply change in each of these first three rounds?
 - c. What will be the ultimate change in demand deposits in the entire banking system?

C H A L L E N G E Q U E S T I O N

1. Sometimes banks wish to hold reserves in excess of the legal minimum. Suppose the Fed makes an open market purchase of \$100,000 in government bonds. The required reserve ratio is 0.1, but each bank decides to hold additional reserves equal to 5 percent of its deposits.
 - a. Trace the effect of the open market purchase of bonds through the first three banks in the money expansion process. Show the changes to each bank's balance sheet.
 - b. Derive the demand deposit multiplier in this case. Is it larger or smaller than when banks hold no excess reserves?
 - c. What is the ultimate change in demand deposits in the entire banking system?

E X P E R I E N T I A L E X E R C I S E S

1. The *Journal of Internet Banking and Commerce* at <http://www.arraydev.com/commerce/jibc/current.htm> is a Web-based magazine devoted to online banking and related issues. Take a look at the current edition and see if you can determine any problems that electronic banking might cause the Fed. Also see what you can learn about the status of Internet banking outside the United States.



2. If you have access to the Interactive Edition of *The Wall Street Journal*, you can use the Briefing Books feature to obtain data on over 10,000 public companies. Find the Briefing Book on a large commercial bank in your area. Look at some of its press releases to determine how this bank has been influenced by Federal Reserve regulations and operations.

CHAPTER

25

THE MONEY MARKET AND THE INTEREST RATE

CHAPTER OUTLINE

The Demand for Money

An Individual's Demand for Money
The Economy-Wide Demand for Money

The Supply of Money

Equilibrium in the Money Market

How the Money Market Reaches Equilibrium

What Happens When Things Change?

How the Fed Changes the Interest Rate
The Fed in Action
How Do Interest Rate Changes Affect the Economy?
Fiscal Policy (and Other Spending Changes) Revisited

Are There Two Theories of the Interest Rate?

Using the Theory: Expectations and the Fed

Expectations and Money Demand
Managing Expectations

Which of the following two newspaper headlines might you see in your daily paper?

1. “Motorists Fear Department of Energy Will Raise Gasoline Prices”
2. “Wall Street Expects Fed to Raise Interest Rates”

You probably know the answer—the first headline is entirely unrealistic. The Department of Energy, the government agency that makes energy policy, has no authority to set prices in any market. The Federal Reserve, by contrast, has full authority to influence the interest rate—the price of borrowing money. And it exercises this authority every day. This is why headlines such as the second one appear in newspapers so often.

In this chapter, you will learn how the Fed, through its control of the money supply, also controls the interest rate. We'll continue our focus on the short run, postponing any discussion about longer time horizons until the next chapter.

THE DEMAND FOR MONEY

Re-read the title of this section. Does it appear strange to you? Don't people always want as much money as possible?

Indeed, they do. But when we speak about the *demand* for something, we don't mean the amount that people would desire if they could have all they wanted, without having to sacrifice anything for it. Instead, economic decision makers always face constraints: They must sacrifice one thing in order to have more of another. Thus, the *demand for money* does not mean how much money people would like to have in the best of all possible worlds. Rather, it means *how much money people would like to hold, given the constraints that they face*. Let's first consider the demand for money by an individual, and then turn our attention to the demand for money in the entire economy.

Identify Goals and Constraints



AN INDIVIDUAL'S DEMAND FOR MONEY

Money is one of the forms in which people hold their wealth. Unfortunately, at any given moment, the total amount of wealth we have is given; we can't just snap our

fingers and have more of it. Therefore, if we want to hold more wealth in the form of money, we must hold less wealth in other forms—savings accounts, money market funds, time deposits, stocks, bonds, and so on. Indeed, people exchange one kind of wealth for another millions of times a day—in banks, stock markets, and bond markets. If you sell shares in the stock market, for example, you give up wealth in the form of corporate stock and acquire money. The buyer of your stock gives up money and acquires the stock.

These two facts—that wealth is given, and that you must give up one kind of wealth in order to acquire more of another—determine an individual's **wealth constraint**. Whenever we speak about the demand for money, the wealth constraint is always in the background, as in the following statement:

An individual's quantity of money demanded is the amount of wealth that the individual chooses to hold as money, rather than as other assets.

Why do people want to hold some of their wealth in the form of money? The most important reason is that money is a *means of payment*; you can buy things with it. Other forms of wealth, by contrast, are *not* used for purchases. (For example, we don't ordinarily pay for our groceries with shares of stock.) However, the other forms of wealth provide a financial return to their owners. For example, bonds, savings deposits, and time deposits pay interest, while stocks pay dividends and may also rise in value (which is called a *capital gain*). Money, by contrast, pays either very little interest (some types of checking accounts) or none at all (cash and most checking accounts). Thus,

when you hold money, you bear an opportunity cost—the interest you could have earned.

Each of us must continually decide how to divide our total wealth between money and other assets. The upside to money is that it can be used as a means of payment. The more of our wealth we hold as money, the easier it is to buy things at a moment's notice, and the less often we will have to pay the costs (in time, trouble, and commissions to brokers) to change our other assets into money. The downside to money is that it pays little or no interest.

To keep our analysis as simple as possible, we'll use bonds as our representative nonmoney asset. We'll also assume money pays *no* interest at all. In our discussion, therefore, people will choose between two assets that are mirror images of each other. Specifically,

individuals choose how to divide wealth between two assets: (1) money, which can be used as a means of payment but earns no interest; and (2) bonds, which earn interest, but cannot be used as a means of payment.

This choice involves a clear trade-off: The more wealth we hold as money, the less often we will have to go through the inconvenience of changing our bonds into money . . . but the less interest we will earn on our wealth.



You've been reminded several times, but since it's a very common mistake, another reminder won't hurt. Money and wealth are *stock* variables, not flow variables. They refer to amounts held *at a particular moment in time*. Do not confuse them with flow variables such as *income* or *saving*. Your income is what you earn *over a period of time*. Your saving is the part of your disposable income that you do not spend *over a period of time*.

Wealth constraint At any point in time, wealth is fixed.

What determines how much money an individual will decide to hold? While tastes vary from person to person, three key variables have rather predictable impacts on most of us.

- *The price level.* The greater the number of dollars you spend in a typical week or month, the more money you will want to have on hand to make your purchases. A rise in the price level, which raises the dollar cost of your purchases, should therefore increase the amount of money you want to hold.
- *Real income.* Suppose the price level remains unchanged, but your income increases. Your purchasing power or *real* income will increase, and so will the number of dollars you spend in a typical week or month. Once again, since you are spending more dollars, you will choose to hold more of your wealth in the form of money.
- *The interest rate.* Interest payments are what you give up when you hold money—the *opportunity cost* of money. The greater the interest rate, the greater the opportunity cost of holding money. Thus, a rise in the interest rate *decreases* your quantity of money demanded.

The effect of the interest rate on the quantity of money demanded will play a key role in our analysis. But before we go any further, you may be wondering whether it is realistic to think that changes in the interest rate—which are usually rather small—would have any effect at all. Here, as in many aspects of economic life, you may not find yourself consciously thinking about the interest rate in deciding how to adjust your money-holding habits. Just as you don't rethink all your habits about using lights and computers every time the price of electricity changes, you may respond to interest rates more casually. But when we add up everybody's behavior, we find a noticeable and stable tendency for people to hold less money when it is more expensive to hold money—that is, when the interest rate is higher.

Identify Goals and Constraints



The Demand for Money by Businesses. Our discussion of money demand has focused on the typical individual. But some money (not a lot in comparison to what individuals hold) is held by businesses. Stores keep some currency in their cash registers, and firms generally keep funds in business checking accounts. Businesses face the same types of constraints as individuals: They have only so much wealth, and they must decide how much of it to hold as money rather than other assets. The quantity of money demanded by businesses follows the same principles we have developed for individuals: They want to hold more money when real income or the price level is higher, and less money when the opportunity cost (the interest rate) is higher.

THE ECONOMY-WIDE DEMAND FOR MONEY

When we use the term *demand for money* without the word *individual*, we mean the total demand for money by all wealth holders in the economy—businesses and individuals. And just as each person and each firm in the economy has only so much wealth, so, too, there is a given amount of wealth in the economy as a whole at any given time. In our analysis, this total wealth must be held in one of two forms: money or bonds.

The (economy-wide) quantity of money demanded is the amount of total wealth in the economy that all households and businesses, together, choose to hold as money rather than as bonds.

THE DEMAND FOR MONEY

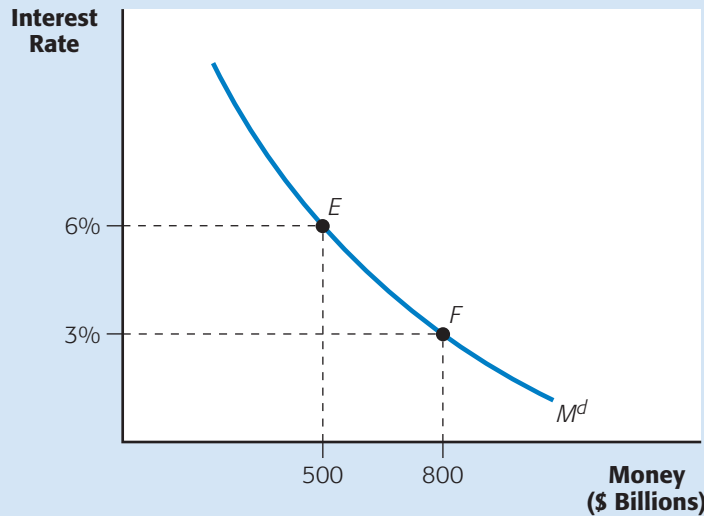


FIGURE 1

The downward-sloping money demand curve shows that, for given real GDP and a given price level, the amount of money demanded by households and firms is inversely related to the interest rate. At an interest rate of 6 percent, \$500 billion of money is demanded; at the lower interest rate of 3 percent, \$800 billion is demanded.

The demand for money in the economy depends on the same three variables that we discussed for individuals. In particular, (1) a rise in the price level will increase the demand for money; (2) a rise in real income (real GDP) will increase the demand for money; and (3) a rise in the interest rate will *decrease* the quantity of money demanded.

The Money Demand Curve. Figure 1 shows a **money demand curve**, which tells us *the total quantity of money demanded in the economy at each interest rate*. Notice that the curve is downward sloping. As long as the other influences on money demand don't change, a drop in the interest rate—which lowers the opportunity cost of holding money—will increase the quantity of money demanded.

Point *E*, for example, shows that when the interest rate is 6 percent, the quantity of money demanded is \$500 billion. If the interest rate falls to 3 percent, we move to point *F*, where the quantity demanded is \$800 billion. As we move along the money demand curve, the interest rate changes, but other determinants of money demand (such as the price level and real income) are assumed to remain unchanged.

Shifts in the Money Demand Curve. What happens when something *other* than the interest rate changes the quantity of money demanded? Then the curve shifts. For example, suppose that real income increases. Then, at each interest rate, individuals and businesses will want to hold *more* of their wealth in the form of money. The entire money demand curve will shift rightward. This is illustrated in Figure 2, where the money demand curve shifts rightward from M_1^d to M_2^d . At an interest rate of 6 percent, the quantity of money demanded rises from \$500 billion to \$700 billion; if the interest rate were 3 percent, the amount of money demanded would rise from \$800 billion to \$1,000 billion.

Money demand curve A curve indicating how much money will be willingly held at each interest rate.

A change in the interest rate moves us along the money demand curve. A change in money demand caused by something other than the interest rate (such as real income or the price level) will cause the curve to shift.

FIGURE 2

An increase in real GDP or in the price level will shift the money demand curve to the right. At each interest rate, more money will be demanded.

A SHIFT IN THE DEMAND FOR MONEY

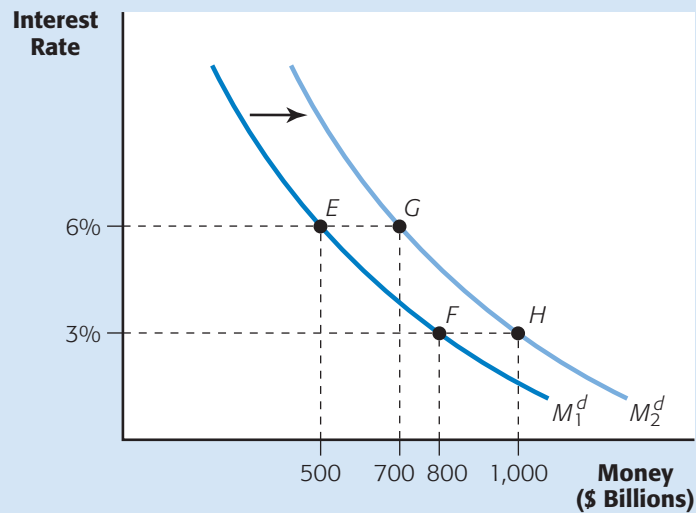


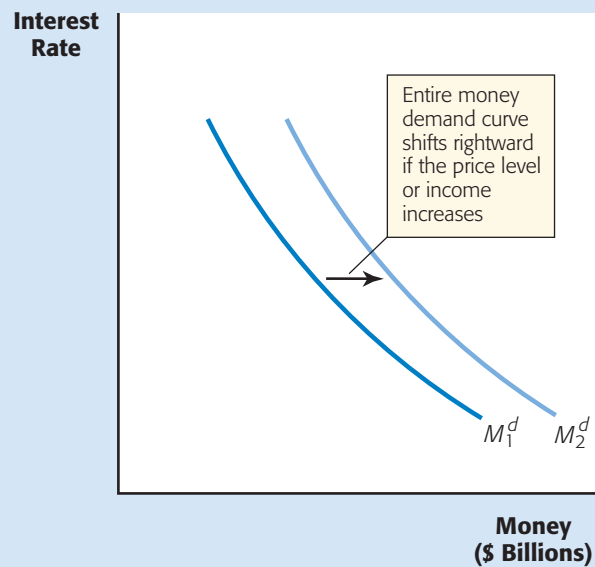
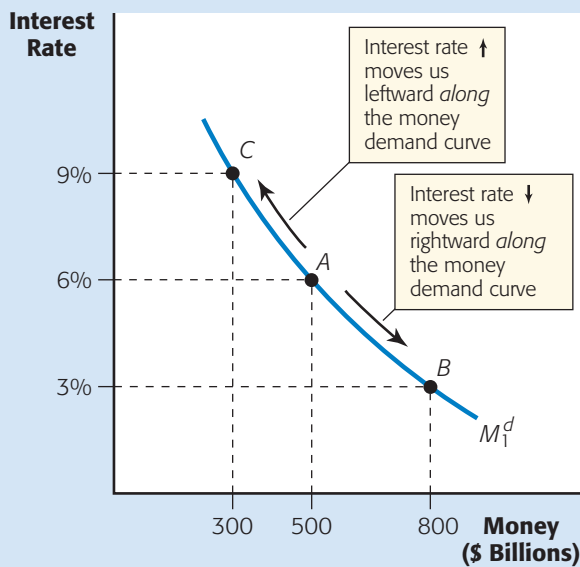
Figure 3 summarizes how the key variables we've discussed so far affect the demand for money.

THE SUPPLY OF MONEY

Just as we did for money demand, we would like to draw a curve showing the quantity of money *supplied* at each interest rate. In the previous chapter, you learned how the Fed controls the money supply: It uses open market operations to inject or

FIGURE 3

SHIFTS AND MOVEMENTS ALONG THE MONEY DEMAND CURVE: A SUMMARY



THE SUPPLY OF MONEY

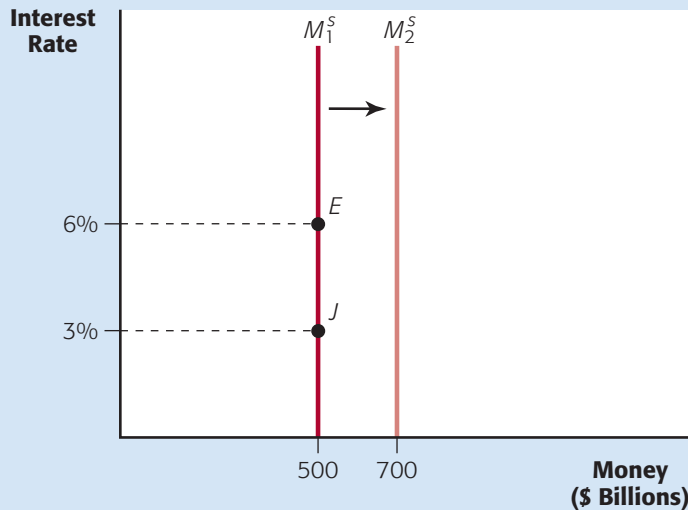


FIGURE 4

Once the Fed sets the money supply, it remains constant until the Fed changes it. The vertical supply curve labeled M_1^s shows a money supply of \$500 billion, regardless of the interest rate. An increase in the money supply to \$700 billion is depicted in a rightward shift of the money supply curve to M_2^s .

withdraw reserves from the banking system and then relies on the demand deposit multiplier to do the rest. Since the Fed decides what the money supply will be, we treat it as a fixed amount. That is, the interest rate can rise or fall, but the money supply will remain constant unless and until the Fed decides to change it.

Look at the vertical line labeled M_1^s in Figure 4. This is the economy's **money supply curve**, which shows the total amount of money supplied at each interest rate. The line is vertical because once the Fed sets the money supply, it remains constant until the Fed changes it. In the figure, the Fed has chosen to set the money supply at \$500 billion. A rise in the interest rate from, say, 3 percent to 6 percent would move us from point J to point E along the solid money supply curve, leaving the money supply unchanged.

Now suppose the Fed, for whatever reason, were to *change* the money supply. Then there would be a *new* vertical line, showing a different quantity of money supplied at each interest rate. Recall from the previous chapter that the Fed raises the money supply by purchasing bonds in an open market operation. For example, if the demand deposit multiplier is 10, and the Fed purchases government bonds worth \$20 billion, the money supply increases by $10 \times \$20 \text{ billion} = \200 billion . In this case, the money supply curve shifts rightward, to the vertical line labeled M_2^s in the figure.

Open market purchases of bonds inject reserves into the banking system, and shift the money supply curve rightward by a multiple of the reserve injection. Open market sales have the opposite effect: They withdraw reserves from the system and shift the money supply curve leftward by a multiple of the reserve withdrawal.

Money supply curve A line showing the total quantity of money in the economy at each interest rate.

EQUILIBRIUM IN THE MONEY MARKET

Now we are ready for Key Step #3: to combine what you've learned about money demand and money supply to find the equilibrium interest rate in the economy. But

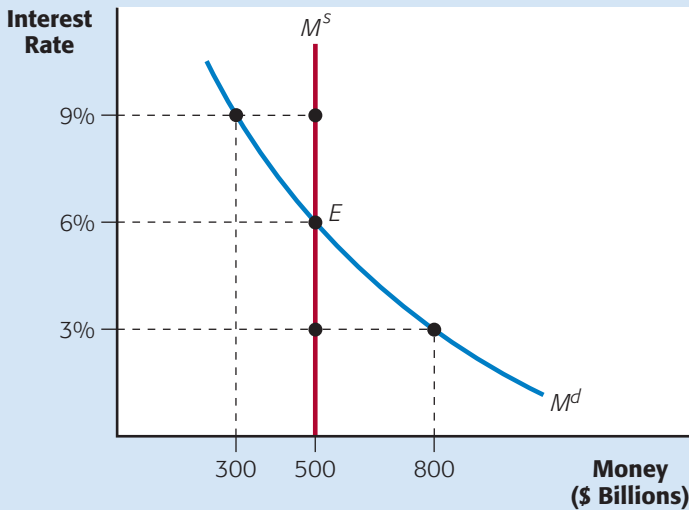


Find the Equilibrium

FIGURE 5

Money market equilibrium occurs when households and firms are content to hold the amount of money they are actually holding. At point E —at an interest rate of 6 percent—the quantity of money demanded equals the quantity supplied, and the market is in equilibrium. At a higher interest rate, such as 9 percent, there would be an excess supply of money, and the interest rate would fall. At a lower interest rate, such as 3 percent, there would be an excess demand for money, and the interest rate would rise.

MONEY MARKET EQUILIBRIUM



before we do, a question may have occurred to you. Haven't we already discussed how the interest rate is determined? Indeed, we have. The classical model tells us that the interest rate is determined by equilibrium in the *loanable funds market*—where a flow of loanable funds is offered by lenders to borrowers. But remember: The classical model tells us how the economy operates in the *long run*. We can rely on its mechanisms to work only over long periods of time. Here, we are interested in how the interest rate is determined in the *short run*, so we must change our perspective. Toward the end of the chapter, we'll come back to the classical model and explain why its theory of the interest rate does not apply in the short run.

In the short run—our focus here—we look for the equilibrium interest rate in the *money market*: the interest rate at which the quantity of money demanded and the quantity of money supplied are equal. Figure 5 combines the money supply and demand curves. Equilibrium occurs at point E , where the two curves intersect. At this point, the quantity of money demanded and the quantity supplied are both equal to \$500 billion, and the equilibrium interest rate is 6 percent.

It is important to understand what equilibrium in the money market actually means. First, remember that the money supply curve tells us the quantity of money, determined by the Fed, that *actually exists* in the economy. Every dollar of this money—either in cash or in checking account balances—is held by *someone*. Thus, the money supply curve, in addition to telling us the quantity of money supplied by the Fed, also tells us the quantity of money that people are actually holding at any given moment. The money demand curve, on the other hand, tells us how much money people *want* to hold at each interest rate. Thus, when the quantity of money supplied and the quantity demanded are equal, all of the money in the economy is being *willingly held*. That is, people are *satisfied* holding the money that they are *actually* holding.

Equilibrium in the money market occurs when the quantity of money people are actually holding (quantity supplied) is equal to the quantity of money they want to hold (quantity demanded).

Can we have faith that the interest rate will reach its equilibrium value in the money market, such as 6 percent in our figure? Indeed we can. In the next section, we explore the forces that drive the money market toward its equilibrium.

HOW THE MONEY MARKET REACHES EQUILIBRIUM

To understand how the money market reaches its equilibrium, suppose that the interest rate, for some reason, were *not* at its equilibrium value of 6 percent in Figure 5. For example, suppose the interest rate were 9 percent. As the figure shows, at this interest rate the quantity of money demanded would be \$300 billion, while the quantity supplied would be \$500 billion. Or, put another way, people would *actually* be holding \$500 billion of their wealth as money, but they would *want* to hold only \$300 billion as money. There would be an **excess supply of money** (the quantity of money supplied would exceed the quantity demanded) equal to \$500 billion – \$300 billion = \$200 billion.

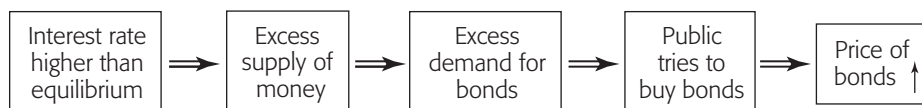
Now comes an important point. Remember that in our analysis, money and bonds are the only two assets available. If people want to hold *less* money than they are currently holding, then, by definition, they must want to hold *more* in bonds than they are currently holding—an **excess demand for bonds**.

When there is an excess supply of money in the economy, there is also an excess demand for bonds.

To understand this more clearly, imagine that instead of the money market, which can seem rather abstract, we were discussing something more concrete: the arrangement of books in a bookcase. Suppose that you have a certain number of books, and you have only two shelves on which to hold all of them—top and bottom. One day, you look at the shelves and decide that, the way you've arranged things, the top shelf has *too many* books. Then, by definition, you must also feel that the bottom shelf has *too few* books. That is, an excess supply of books on the top shelf (it has more books than you want there) is the same as an excess demand for books on the bottom shelf (it has fewer books than you want there).

A similar conclusion applies to the money market. People allocate a given amount of wealth between two different assets: money and bonds. Too much in one asset implies too little in the other.

So far, we've established that if the interest rate were 9 percent, which is higher than its equilibrium value, there would be an excess supply of money, and an excess demand for bonds. What would happen? The public would try to convert the undesired money into bonds. That is, people would try to *buy* bonds. Just as there is a market for money, there is also a market for bonds. And as the public begins to demand more bonds, making them scarcer, *the price of bonds will rise*. We can illustrate the steps in our analysis so far as follows:



We conclude that, when the interest rate is higher than its equilibrium value, the price of bonds will rise. Why is this important? In order to take our story further, we must first take a detour for a few paragraphs.

Excess supply of money The amount of money supplied exceeds the amount demanded at a particular interest rate.

Excess demand for bonds The amount of bonds demanded exceeds the amount supplied at a particular interest rate.



When bond traders—such as those pictured here—try to buy more bonds, the price of bonds rises, and the interest rate on those bonds falls.

An Important Detour: Bond Prices and Interest Rates. A bond, in the simplest terms, is a promise to pay back borrowed funds at a certain date or dates in the future. There are many types of bonds. Some promise to make payments each month or each year for a certain period and then pay back a large sum at the end. Others promise to make just one payment—perhaps 1, 5, 10, or more years from the date the bond is issued. When a large corporation or the government wants to borrow money, it issues a new bond and sells it in the marketplace; the amount borrowed is equal to the price of the bond.

Let's consider a very simple example: a bond that promises to pay to its holder \$1,000 exactly one year from today. Suppose that you purchase this bond from the issuer—a firm or government agency—for \$800. Then you are lending \$800 to the issuer, and you will be paid back \$1,000 one year later. What interest rate are you earning on your loan? Let's see: You will be getting back \$200 more than you lent, so that is your *interest payment*. The *interest rate* is the interest payment divided by the amount of the loan, or $\$200/\$800 = 0.25$ or 25 percent.

Now, what if instead of \$800, you paid a price of \$900 for this very same bond. The bond still promises to pay \$1,000 one year from now, so your interest payment would now be \$100, and your interest rate would be $\$100/\$900 = 0.11$ or 11 percent—a considerably lower interest rate. As you can see, the interest rate that you will earn on your bond depends entirely on the *price* of the bond. *The higher the price, the lower the interest rate.*

This general principle applies to virtually all types of bonds, not just the simple one-time-payment bond we've considered here. Bonds promise to pay various sums to their holders at different dates in the future. Therefore, the more you pay for any bond, the lower your overall rate of return, or interest rate, will be. Thus:

When the price of bonds rises, the interest rate falls; when the price of bonds falls, the interest rate rises.¹

The relationship between bond prices and interest rates helps explain why the government, the press, and the public are so concerned about the *bond market*, where bonds issued in previous periods are bought and sold. This market is sometimes called the *secondary* market for bonds, to distinguish it from the *primary* market where newly issued bonds are bought and sold. When you hear that “the bond market rallied” on a particular day of trading, it means that prices rose in the secondary bond market. This is good news for bond holders. But it is also good news for any person or business that wants to borrow money. When prices rise in the secondary market, they immediately rise in the primary market as well, since newly issued bonds and previously issued bonds are almost perfect substitutes for each other. Therefore, a bond market rally not only means lower interest rates in the secondary market, it also means lower interest rates in the primary market, where firms borrow money by issuing new bonds. Sooner or later, it will also lead to a drop in the interest rate on mortgages, car loans, credit card balances, and even many student loans. This is good news for borrowers. But it is bad news for anyone wishing to lend money, for now they will earn less interest.

¹ In our macroeconomic model of the economy, we refer to *the* interest rate. In the real world, there are many types of interest rates—a different one for each type of bond, and still other rates on savings accounts, time deposits, car loans, mortgages, and more. However, all of these interest rates move up and down together, even though some may lag behind a few days, weeks, or months. Thus, when bond prices rise, interest rates *generally* will fall, and vice versa.

Now that you understand the relationship between bond prices and interest rates, let's return to our analysis of the money market.

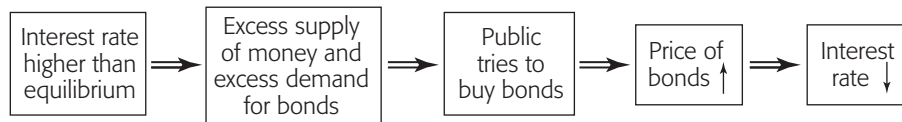
Back to the Money Market.

Look back at Figure 5, and let's recap what you've learned so far. If the interest rate were 9 percent, there would be an excess supply of money, and therefore an excess demand for bonds. The public would try to buy bonds, and the price of bonds would rise. Now we can complete the story. As you've just learned, a rise in the price of bonds means a *decrease* in the interest rate. The complete sequence of events is



We've shown that when the money market is not in equilibrium, the public *tries* to buy or sell bonds. The word *tries* is important. On any given day, the total number of bonds—like the money stock—is some fixed amount. (We ignore the relatively small number of newly issued bonds added to the market each day.) Therefore, it is impossible for the public as a whole to acquire more bonds, or to get rid of them. A single individual may be able to acquire bonds or money by exchanging with another individual. But the total amount of bonds and money held by the public will remain unchanged.

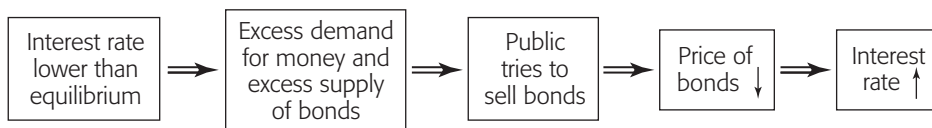
How, then, does the money market achieve equilibrium? When many people simultaneously try to sell bonds, they cause the price of bonds to fall. The price of bonds stops falling only when the public, as a whole, is happy holding the same bonds they were holding originally. When many people simultaneously try to acquire bonds, they cause the price of bonds to rise until the public is, once again, satisfied holding what it started with. *Individuals* may buy and sell bonds, but the public, as a whole, can only *try* to.



Thus, if the interest is 9 percent in our figure, it will begin to fall. Therefore, 9 percent is *not* the equilibrium interest rate.

How far will the interest rate fall? As long as there continues to be an excess supply of money, and an excess demand for bonds, the public will still be trying to acquire bonds and the interest rate will continue to fall. But notice what happens in the figure as the interest rate falls: The quantity of money demanded *rises*. Finally, when the interest rate reaches 6 percent, the excess supply of money, and therefore the excess demand for bonds, is eliminated. At this point, there is no reason for the interest rate to fall further, so 6 percent is, indeed, our equilibrium interest rate.

We can also do the same analysis from the other direction. Suppose the interest rate were *lower* than 6 percent in the figure. Then, as you can see in Figure 5, there would be an *excess demand for money*, and an *excess supply of bonds*. In this case, the following would happen:



The interest rate would continue to rise until it reached its equilibrium value: 6 percent.

WHAT HAPPENS WHEN THINGS CHANGE?

Now that we have seen how the interest rate is *determined* in the money market, we turn our attention to *changes* in the interest rate. We'll focus on two questions:

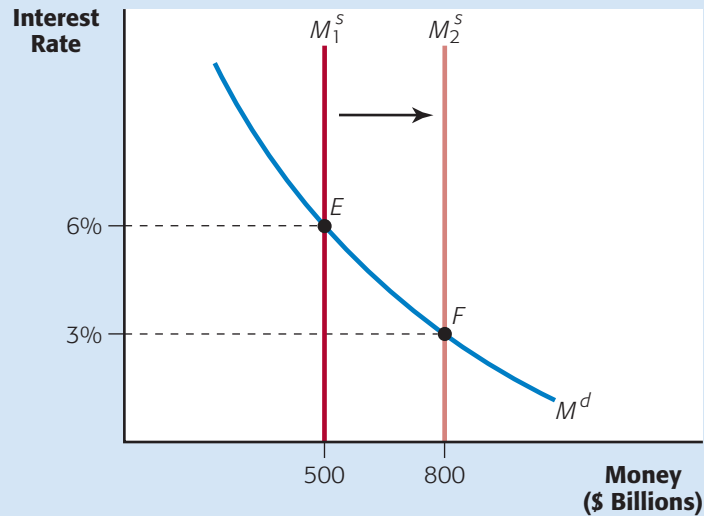


What Happens When Things Change?

FIGURE 6

If the Fed wishes to lower the interest rate, it can do so by increasing the money supply. At point *E*, the money market is in equilibrium at an interest rate of 6 percent. To lower the rate, the Fed could increase the money supply to \$800 billion. At the original interest rate, there would be an excess supply of money (and an excess demand for bonds). Bond prices would rise, and the interest rate would fall until a new equilibrium is established at point *F* with an interest rate of 3 percent.

AN INCREASE IN THE MONEY SUPPLY



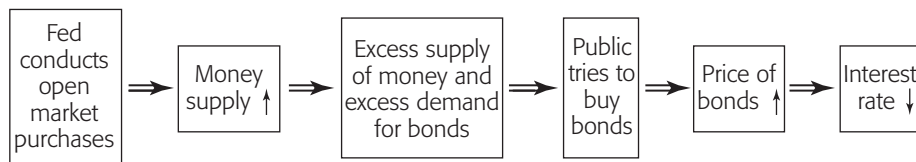
(1) What *causes* the equilibrium interest rate to change? and (2) What are the *consequences* of a change in the interest rate? As you are about to see, the Fed can change the interest rate as a matter of policy, or the interest rate can change on its own, as a by-product of other events in the economy. We'll begin with the Fed.

HOW THE FED CHANGES THE INTEREST RATE

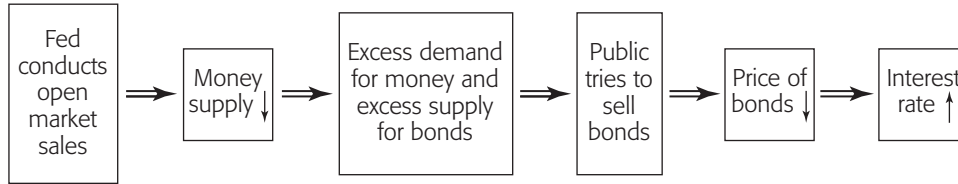
Changes in the interest rate from day to day, or week to week, are often caused by the Fed. Later in this chapter, you'll learn *why* the Fed often wants to manipulate the interest rate. For now, we'll focus on *how* the Fed does this.

Suppose the Fed wants to *lower* the interest rate. Fed officials cannot just *declare* that the interest rate should be lower. To change the interest rate, the Fed must change the *equilibrium* interest rate in the money market, and it does this by changing the money supply.

Look at Figure 6. Initially, with a money supply of \$500 billion, the money market is in equilibrium at point *E*, with an interest rate of 6 percent. To lower the interest rate, the Fed *increases* the money supply through open market purchases of bonds. In the figure, the Fed raises the money supply to \$800 billion, shifting the money supply curve rightward. (This is a much greater shift than the Fed would ever actually engineer in practice, but it makes the graph easier to read.) At the old interest rate of 6 percent, there would be an excess supply of money and an excess demand for bonds. This will drive the interest rate down until it reaches its new equilibrium value of 3 percent, at point *F*. The process works like this:



The Fed can *raise* the interest rate as well, through open market *sales* of bonds. In this case, the money supply curve in Figure 6 would shift leftward (not shown), setting off the following sequence of events:



If the Fed increases the money supply by buying government bonds, the interest rate falls. If the Fed decreases the money supply by selling government bonds, the interest rate rises. By controlling the money supply through purchases and sales of bonds, the Fed can also control the interest rate.

THE FED IN ACTION

When the Fed tries to achieve a macroeconomic goal by controlling or manipulating the money supply, it is conducting *monetary policy*. During periods of economic calm, such as 1993 through 1999, the Fed's monetary policy tends to be stable, and the interest rate remains at about the same level from year to year. Occasionally, however, the Fed sees the need to act dramatically—to adjust the money stock aggressively and engineer large changes in interest rates. Such an episode occurred in the period from mid-1999 to early 2000. In 1999, the Fed believed that the economy was becoming overheated and that it needed to be slowed down by a rise in the interest rate (you'll learn why a higher interest rate slows the economy in the next section).

Figure 7 shows what happened. Starting in June 1999, the Fed began to conduct open market sales of bonds, withdrawing reserves from the banking system. As you can see in panel (a) of the figure, from mid-1999 to early 2000, banking system reserves fell by about \$2.9 billion. This, in turn, shrank demand deposits and similar checking account balances by about \$29.3 billion—10 times the withdrawal of reserves. (The previous chapter explained why the decrease in checking-type accounts is greater than the decrease in reserves.)

Because checking account balances are part of the money supply, the Fed's action shifted the money supply curve leftward. This, in turn, caused the interest rate to rise. Panel (c) of the figure shows changes in the *federal funds rate*—the interest rate that the Fed watches the most closely when it conducts monetary policy. The **federal funds rate** is the interest rate that banks with excess reserves charge for lending reserves to other banks. Although it is just an interest rate for lending among banks, it varies closely with other interest rates in the economy, so it gives us a good idea of how interest rates in general were changing during this period. As you can see, the federal funds rate rose by a full percentage point, from 4.75 percent to 5.75 percent, over the period. From March to May, 2000 (not shown), after the graphs in Figure 7 were drawn, the Fed continued its tightening of the money supply, and the federal funds rate rose even higher, to 6.5 percent.

The contraction of the money supply and the rise in interest rates from mid-1999 and into 2000 raise some important questions. Why would the Fed feel the need to raise interest rates in the first place? Why does it do so gradually, rather than all at once? And how does the Fed know how much to tighten? We'll be answering



You can find recent and historical data on the money supply and interest rates at the Fed's Web site: <http://www.bog.frb.fed.us/releases>.

Federal funds rate The interest rate charged for loans of reserves among banks.

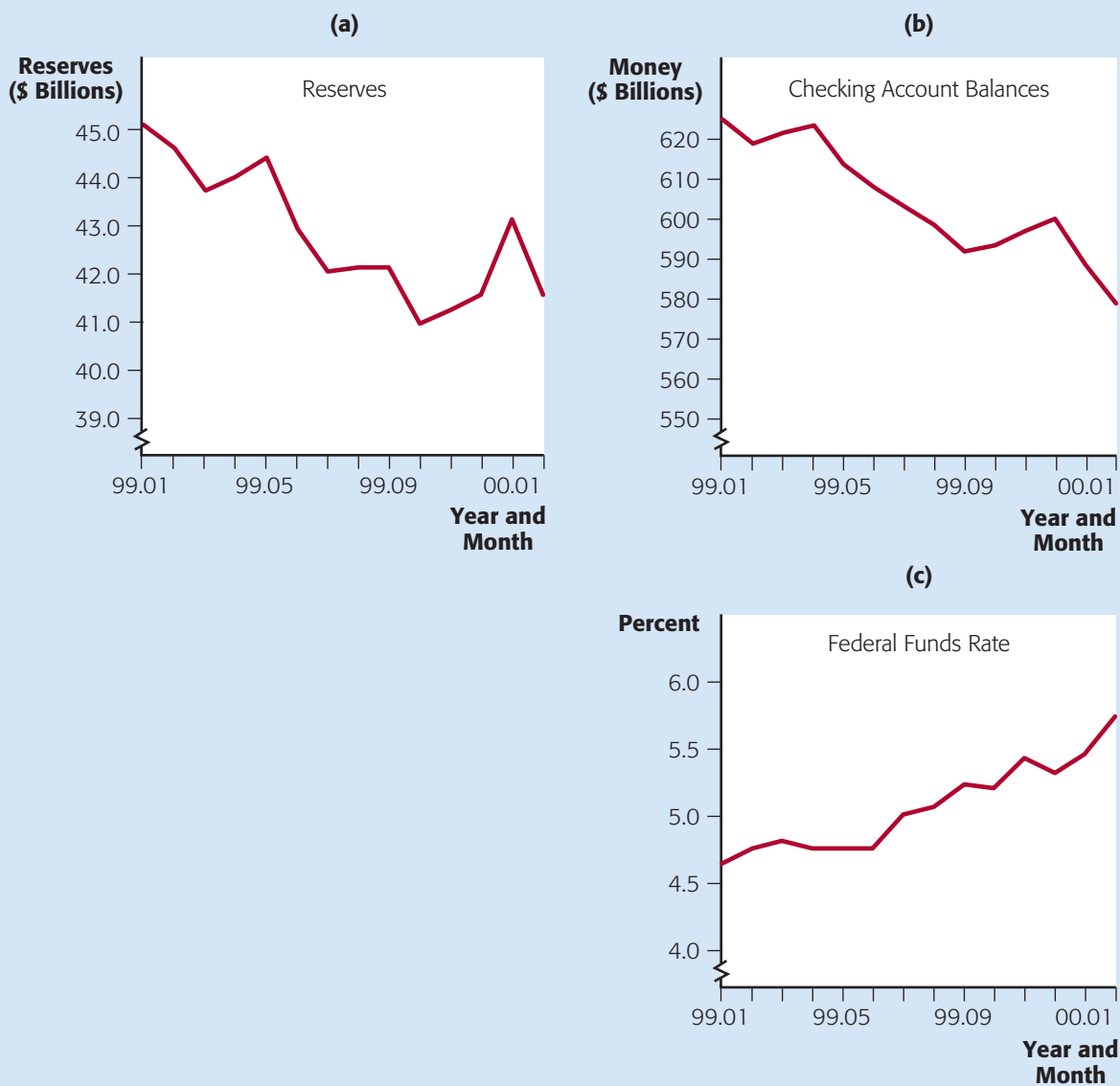
questions like these in the next two chapters. But we can begin to understand the Fed's motives by learning how interest rate changes affect the economy, which is the subject of the next section.

HOW DO INTEREST RATE CHANGES AFFECT THE ECONOMY?

Suppose the Fed increases the money supply through open market purchases of bonds. The interest rate falls, for the reasons discussed earlier in this chapter, and strongly confirmed by the data shown in Figure 7. But what then? How is the

FIGURE 7

THE FED IN ACTION



In June 1999, the Fed began to sell bonds and withdraw reserves from the banking system. As a result, checking account balances fell, the federal funds rate increased, and other interest rates in the economy (not shown) increased as well.

macroeconomy affected? The answer is: *A drop in the interest rate will boost several different types of spending in the economy.*

How the Interest Rate Affects Spending. First, a lower interest rate stimulates business spending on plant and equipment. This idea came up a few chapters ago in the classical model, but we will go back over it here.

Remember that the interest rate is one of the key costs of any investment project. If a firm must borrow funds, it will have to pay for them at the going rate of interest—for example, by selling a bond at the going price. If the firm uses its *own* funds, so it doesn't have to borrow, the interest rate *still* represents a cost: Each dollar spent on plant and equipment *could* have been lent to someone else at the going interest rate. Thus, the interest rate is the *opportunity cost* of the firm's own funds when they are spent on plant and equipment.

A firm deciding whether to spend on plant and equipment compares the benefits of the project—the increase in future income—with the costs of the project. With a lower interest rate, the costs of funding investment projects are lower, so more projects will get the go-ahead. Other variables affect investment spending as well. But for given values of these other variables, a drop in the interest rate will cause an increase in spending on plant and equipment.

Interest rate changes also affect another kind of investment spending: spending on new houses and apartments that are built by developers or individuals. Most people borrow to buy houses or condominiums, and most developers borrow to build apartment buildings. The loan agreement for housing is called a *mortgage*, and mortgage interest rates move closely with other interest rates. Thus, when the Fed lowers the interest rate, families find it more affordable to buy homes, and landlords find it more profitable to build new apartments. Total investment in new housing increases.

Finally, in addition to investment spending, the interest rate affects consumption spending on big ticket items such as new cars, furniture, and dishwashers. Economists call these *consumer durables* because they usually last several years. People often borrow to buy consumer durables, and the interest rate they are charged tends to rise and fall with other interest rates in the economy. Spending on new cars, the most expensive durable that most of us buy, is especially sensitive to interest rate changes. When the interest rate falls, consumption spending rises at *any* level of disposable income. It causes a *shift* of the consumption function, not a movement along it. Therefore, we consider this impact on consumption to be a rise in autonomous consumption spending, called *a* in our discussion of the consumption function.

We can summarize the impact of monetary policy as follows:

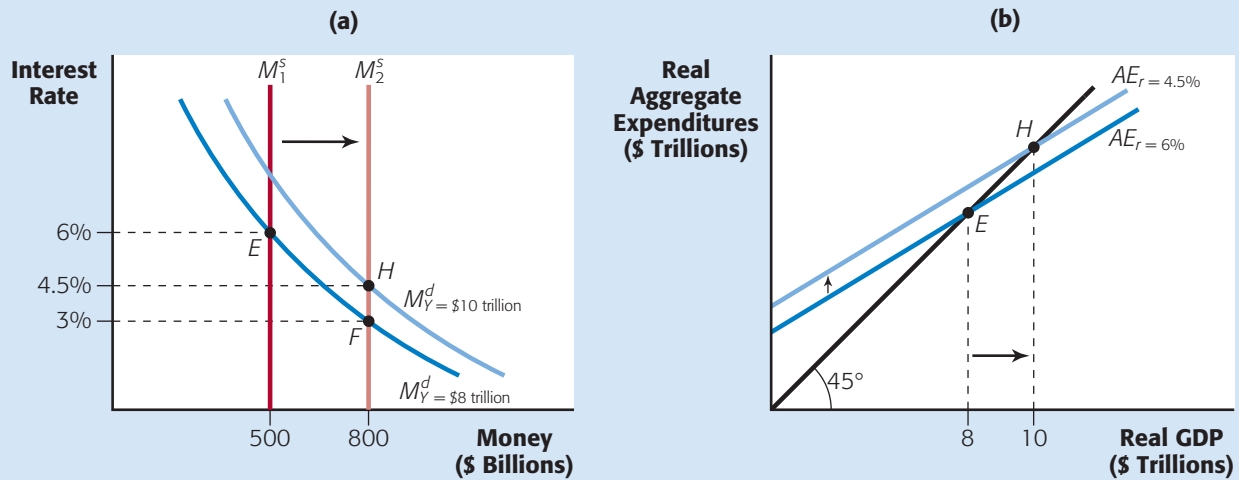
When the Fed increases the money supply, the interest rate falls, and spending on three categories of goods increases: plant and equipment, new housing, and consumer durables (especially automobiles). When the Fed decreases the money supply, the interest rate rises, and these categories of spending fall.

Monetary Policy and the Economy. Now we can finally see how monetary policy affects the economy overall. The only remaining step is one you learned two chapters ago: how a change in spending affects output and employment. This is what our short-run macro model was all about.

In Figure 8, we revisit the short-run macro model, but we now include the money market in our analysis. In panel (a), the Fed has initially set the money supply at \$500 billion. Equilibrium is at point *E*, with an interest rate (*r*) of 6 percent.

FIGURE 8

MONETARY POLICY AND THE ECONOMY



Monetary policy involves an interaction between the interest rate and equilibrium real GDP. Initially, the Fed has set the money supply at \$500 billion, so the interest rate is 6 percent (point E). Given that interest rate, aggregate expenditure is $AE_{r=6\%}$ in panel (b), and real GDP is \$8 trillion (point E).

If the Fed increases the money supply to \$800 billion, money market equilibrium moves temporarily to point F in panel (a). The interest rate falls, stimulating interest-sensitive spending and driving aggregate expenditures upward in panel (b). Through the multiplier process, real GDP increases. As it does, the money demand curve shifts rightward in panel (a). In the new equilibrium, real GDP is \$10 trillion and the interest rate is 4.5 percent (point H).

Panel (b) shows the familiar short-run aggregate expenditure diagram, with equilibrium at point E , and equilibrium GDP equal to \$8 trillion.

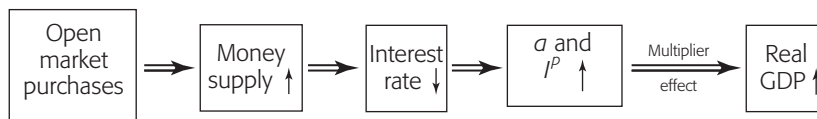
But notice the new labels in the figure. The aggregate expenditure line has the subscript “ $r = 6\%$,” and the money demand curve has the subscript “ $Y = \$8$ trillion.” These are necessary because of the *interdependence* between the interest rate and equilibrium GDP. Recall that the money demand curve will shift if there is a change in real income. Therefore, our money demand curve is drawn for a particular level of real income—the level determined in panel (b), or \$8 trillion. Similarly, as you are about to see, a change in the interest rate will cause the aggregate expenditure line to shift. Therefore, our aggregate expenditure line is drawn for a particular interest rate—the one determined in the money market, or 6 percent. As you can see, the equilibrium in each panel depends on the equilibrium in the other panel.

Now we suppose that the Fed increases the money supply to \$800 billion. (Again, this is an unrealistically large change in the money supply, but it makes it easier to see the change in the figure.) In panel (a), the money market equilibrium moves from point E to point F , and the interest rate begins to drop. (It would drop all the way down to 3 percent, except that the money demand curve will start shifting as well before we are finished.) The drop in the interest rate causes planned investment spending on plant and equipment and on new housing to rise. It also causes an increase in consumption spending—especially on consumer durables like automobiles—to rise at any level of income. This is an increase in autonomous consumption spending (a). In panel (b), the rise in spending causes the aggregate expenditure line to shift upward, setting off the multiplier effect and increasing equilib-

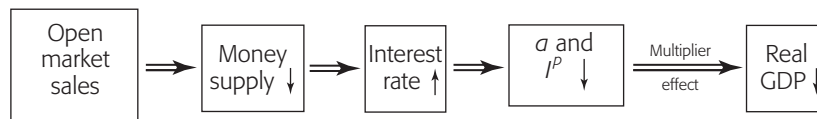
rium GDP. The rise in income causes the money demand curve to shift rightward, since the demand for money is greater when income is higher.

To find the final equilibrium in the economy, we would need quite a bit of information about how sensitive spending is to the drop in the interest rate, as well as how changes in income feed back into the money market to affect the interest rate. In Figure 8, we've illustrated just one possibility, in which the new equilibrium is at point *H* in both the money market and the aggregate expenditure diagrams. At this new equilibrium, the interest rate ends up at 4.5 percent, so the higher aggregate expenditure line is labeled " $r = 4.5\%$." Equilibrium GDP has risen to \$10 trillion, so the new higher money demand curve is labeled " $Y = \$10$ trillion." In the end, we see that the Fed, by increasing the money supply and lowering the interest rate, has increased the level of output.

We've covered a lot of ground to reach our conclusion, so let's review the highlights of how monetary policy works. This is what happens when the Fed conducts open market purchases of bonds:



Open market *sales* by the Fed have exactly the opposite effects. In this case, the money supply curve in Figure 8 would shift leftward (not shown), driving the interest rate up. The rise in the interest rate would cause a decrease in interest-sensitive spending (*a* and *I*), shifting the aggregate expenditure line downward. Equilibrium GDP would fall by a multiple of the initial decrease in spending.



FISCAL POLICY (AND OTHER SPENDING CHANGES) REVISITED

Two chapters ago, we discussed how fiscal policy affects the economy in the short run. For example, an increase in government purchases causes output to rise, and in successive rounds of the multiplier, spending and output rise still more. Now that we've added the money market to our analysis, it's time to revisit fiscal policy. As you'll see, its effects are now a bit more complicated.

Figure 9 shows the money market and the familiar short-run aggregate expenditure diagram. Initially, we have equilibrium in both panels. In panel (a), the money market equilibrium is point *E*, with the interest rate at 6 percent. In panel (b), the solid aggregate expenditure line, labeled " $r = 6\%$," is consistent with the interest rate we've found in the money market. As you can see, with this aggregate expenditure line, the equilibrium is at

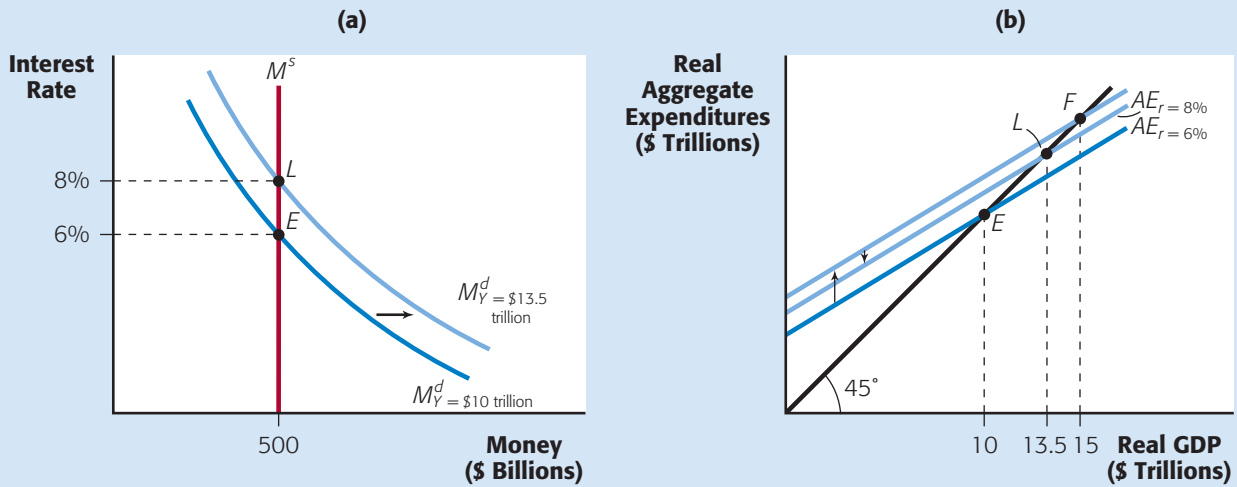


When thinking about the effects of monetary policy, try not to confuse movements *along* the aggregate expenditure line with *shifts* of the line itself. We move along the line only when a change in *income* causes spending to change. The line shifts when something *other* than a change in income causes spending to change.

When the Fed changes the interest rate, both types of changes occur, but it's important to keep the order straight. *First*, the drop in the interest rate (something other than income) causes interest-sensitive spending to change, *shifting* the aggregate expenditure line. *Then*, increases in income in each round of the multiplier cause further increases in spending, moving us *along* the new aggregate expenditure line.

FIGURE 9

FISCAL POLICY AND THE MONEY MARKET



The economy is initially in equilibrium with an interest rate of 6 percent in panel (a) and real GDP of \$10 trillion in panel (b). An increase in government purchases shifts the aggregate expenditure line upward, triggering the multiplier process. If the interest rate did not change, equilibrium would be reestablished at point *F* in panel (b) with real GDP of \$15 trillion. But the increase in GDP stimulates money demand in panel (a), driving the interest rate upward to 8 percent at point *L*. That reduces interest-sensitive spending, lowering aggregate expenditure to $AE_{r=8\%}$ in panel (b) so that the real GDP at the new equilibrium is \$13.5 trillion (point *L*).

point *E*, with real GDP equal to \$10 trillion, just as we assumed when we drew the money demand curve in panel (a).

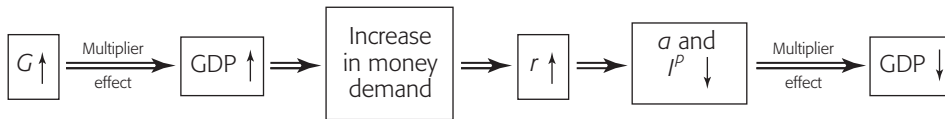
An Increase in Government Purchases. Now let's see what happens when the government changes its fiscal policy, say, by increasing government purchases by \$2 trillion. Panel (b) shows the initial effect: The aggregate expenditure line shifts upward, by \$2 trillion, to the topmost aggregate expenditure line. This new aggregate expenditure line is drawn for the same interest rate as the original line: $r = 6\%$. The shift illustrates what *would* happen if there were no change in the interest rate, as in our analysis of fiscal policy two chapters ago.

As you've learned, the increase in government purchases will set off the multiplier process, increasing GDP and income in each round. *If this were the end of the story*, the result would be a rise in real GDP equal to $[1/(1 - MPC)] \times \Delta G$. In our example, with an *MPC* of 0.6, the multiplier would be $1/(1 - 0.6) = 2.5$. The new equilibrium would be at point *F*, with GDP equal to \$15 trillion—a rise of \$5 trillion.

But point *F* is *not* the end of our story—not when we include effects in the money market. As income increases, the money demand curve in panel (a) will shift rightward, raising the interest rate. As a result, autonomous consumption (*a*) and investment spending (*I*) will decrease and shift the aggregate expenditure line downward. That is,

an increase in government purchases, which by itself shifts the aggregate expenditure line upward, also sets in motion forces that shift it downward.

We can outline these forces as follows:



Thus, at the same time as the increase in government purchases has a *positive* multiplier effect on GDP, the decrease in a and I have *negative* multiplier effects. Which effect dominates? The positive multiplier effect. Why? Because the only force pulling GDP down—a higher interest rate—*depends upon* a rise in GDP. (It is the rise in GDP that shifts the money demand curve and drives up the interest rate.) If the negative effect on GDP were stronger, GDP would actually decrease in the end, so the interest rate would be lower, not higher, and there would be no force pulling GDP down at all.

Thus, we know that an increase in government purchases causes GDP to rise. But the rise is smaller than the simple multiplier formula suggests. That’s because the simple multiplier ignores the moderating effect of a rise in the interest rate on GDP.

In the short run, an increase in government purchases causes real GDP to rise, but not by as much as if the interest rate had not increased.

Let’s sum up the characteristics of the new equilibrium after an increase in government purchases:

- The aggregate expenditure line is higher, but by less than ΔG .
- Real GDP and real income are higher, but the rise is less than $[1/(1 - MPC)] \times \Delta G$.
- The money demand curve has shifted rightward, because real income is higher.
- The interest rate is higher, because money demand has increased.
- Autonomous consumption and investment spending are lower, because the interest rate is higher.

Figure 9 indicates one possible result that is consistent with all of these requirements. In the figure, the new equilibrium occurs at point L in both panels, with the new equilibrium GDP at \$13.5 trillion and the new equilibrium interest rate at 8 percent. Notice that real GDP has risen, but by only \$3.5 trillion—not the \$5 trillion suggested by the simple multiplier formula. Moreover, the two panels of the diagram are consistent with each other. The aggregate expenditure line (labeled $r = 8\%$) corresponds to the equilibrium interest rate in the money market. The money demand curve (labeled “ $Y = \$13.5$ trillion”) corresponds to the equilibrium GDP in the aggregate expenditure diagram.

Crowding Out Once Again. Our analysis illustrates an interesting by-product of fiscal policy. Comparing our initial equilibrium (point E in both panels) to the final equilibrium (point L), we see that government purchases increase, but—because of the rise in the interest rate—*investment spending has decreased*.

What about consumption spending? It is influenced by two opposing forces. The rise in the interest rate causes *some* types of consumption spending (e.g., on automobiles) to decrease, but the rise in *income* makes other types of consumption spending *increase*. Thus, an increase in government purchases may increase or decrease consumption spending, depending on which effect is stronger.

Summing up:

When effects in the money market are included in the short-run macro model, an increase in government purchases raises the interest rate and crowds out some private investment spending. It may also crowd out consumption spending.

This should sound familiar. In the classical, long-run model, an increase in government purchases also causes crowding out. But there is one important difference between crowding out in the classical model and the effects we are outlining here. In the classical model, there is *complete crowding out*: Investment spending and consumption spending fall by the same amount that government purchases rise. As a result, total spending does not change at all, and neither does GDP. This is why, in the long run, we expect fiscal policy to have no effect on equilibrium GDP.

In the short run, however, our conclusion is somewhat different. While we expect *some* crowding out from an increase in government purchases, *it is not complete*. Investment spending falls, and consumption spending *may* fall, but together, they do not drop by as much as the rise in government purchases. In the short run, real GDP rises.

Other Spending Changes. So far, we've focused on the impact on the economy of a change in government purchases. But our analysis extends to *any* shock that shifts the aggregate expenditure line. Positive shocks would shift the aggregate expenditure line upward, just as in Figure 9. More specifically:

Increases in government purchases, investment, and autonomous consumption, as well as decreases in taxes, all shift the aggregate expenditure line upward. Real GDP rises, but so does the interest rate. The rise in equilibrium GDP is smaller than if the interest rate remained constant.

For example, a \$2 trillion increase in investment spending shifts the aggregate expenditure line upward by \$5 trillion, as in Figure 9. If there were no rise in the interest rate, real GDP would rise according to the simple multiplier of $1/(1 - MPC) = 2.5$. Applying this multiplier to a \$2 trillion increase in investment tells us that real GDP would rise by a full \$5 trillion. But once again, the rise in GDP does drive up the interest rate in the money market, which works to decrease investment and interest-sensitive consumption. And once again, GDP will rise, but not by as much as the simple multiplier suggests.

Negative shocks shift the aggregate expenditure line *downward*. More specifically:

Decreases in government purchases, investment, and autonomous consumption, as well as increases in taxes, all shift the aggregate expenditure line downward. Real GDP falls, but so does the interest rate. The decline in equilibrium GDP is smaller than if the interest rate remained constant.

What About the Fed? In our analysis of spending shocks, we've made an implicit but important assumption. Look back at Figure 9. Notice that, from beginning to end, the money supply curve never shifted. This implies that the Fed just stands by, not interfering at all with the changes we've been describing. More specifically, we've been assuming that *the Fed does not change the money supply in response to shifts in the aggregate expenditure line*.

While this assumption has helped us focus on the impact of spending shocks, it is not the way the Fed has conducted policy during the past few decades. Instead, the Fed has usually responded to neutralize the impact of spending shocks. That is, it has used monetary policy to prevent spending shocks from changing GDP at all. You'll learn why, and how, the Fed does this when we revisit monetary policy in the chapter after next.

ARE THERE TWO THEORIES OF THE INTEREST RATE?

At the beginning of this chapter, you were reminded that you had already learned a different theory of how the interest rate is determined in the economy. In the classical model, the interest rate is determined in the *market for loanable funds*. In this chapter, you learned that the interest rate is determined in the *money market*, where people make decisions about holding their wealth as money and bonds. Which theory is correct?

The answer is: Both are correct. The classical model, you remember, tells us what happens in the economy in the *long run*. Therefore, when we ask what changes the interest rate over long periods of time—many years or even a decade—we should think about the market for loanable funds. But over shorter time periods—days, weeks, or months—we should use the money market model presented in this chapter.

Why don't we use the classical loanable funds model to determine the interest rate in the short run? Because, as you've seen, the economy behaves differently in the short run than it does in the long run. For example, in the classical model, output is automatically at full employment. But in the short run, output changes as the economy goes through booms and recessions. These changes in output affect the loanable funds market in ways that the classical model does not consider. For example, flip back to the chapter on the classical model and look at Figure 9 there. Recessions, which decrease household income, also decrease household saving at any given interest rate: With less income, households will spend less *and* save less. The supply of loanable funds curve would shift leftward in the diagram, and the interest rate would rise. The classical model—because it ignores recessions—ignores these short-run changes in the supply of loanable funds.

The classical model also ignores an important idea discussed in this chapter: that the public continuously chooses how to divide its wealth between money and bonds. In the short run, the public's preferences over money and bonds can change, and this, in turn, can change the interest rate. This idea does not appear in the classical model.

Of course, in the long run, the classical model gives us an accurate picture of how the economy—and the interest rate—behaves. Recessions and booms don't last forever, so the economy returns to full employment. Thus, in the long run we needn't worry about recessions causing shifts in the supply of loanable funds curve. Also, changes in preferences for holding money and bonds are rather short lived. We can ignore these changes when we take a long-run view.

Our view of the interest rate depends on the time period we are considering. In the long run, we view the interest rate as determined in the market for loanable funds, where household saving is lent to businesses and the government. In the short run, we view the interest rate as determined in the money market, where wealth holders adjust their wealth between money and bonds, and the Fed participates by controlling the money supply.

Using the THEORY



EXPECTATIONS AND THE FED

So far, we've considered changes in the interest rate engineered by the Fed, or caused by a spending shock that shifts the aggregate expenditure line upward or downward. Here, we discuss one additional source of interest rate changes: a *shift in the money demand curve*. Note that you've already seen what happens when the money demand curve shifts as a by-product of a spending shock (Figure 9). Here, we explore what happens when the *initial shock* to the economy is a shift in the money demand curve. More specifically, we'll look at what happens when a change in expectations about future interest rates shifts the demand for money curve.

EXPECTATIONS AND MONEY DEMAND

Why should expectations about the future interest rate affect money demand *today*? Because bond prices and interest rates are negatively related. If you expect the interest rate to rise in the future, then you also expect the price of bonds to fall in the future.

To see this more clearly, imagine (pleasantly) that you hold a bond promising to pay you \$100,000 in exactly one year and that the going interest rate is currently 5 percent. The price for your bond will be \$95,238. Why? At that price, a buyer would earn $\$100,000 - \$95,238 = \$4,762$ in interest. Since the bond cost \$95,238, the buyer's rate of return would be $\$4,762/\$95,238 = 0.05$, or 5 percent—the going rate of interest. If you tried to charge more than \$95,238 for the bond, its rate of return would be less than 5 percent, so no one would buy it—they could always earn 5 percent by buying another bond that pays the going rate of interest.

Now suppose that you *expect* the interest rate to rise to 10 percent in the near future, say, next week. (This is an unrealistically large change in the interest rate in so short a time, but it makes the point dramatically.) Then next week, the going price for your bond would be only about \$90,909. At that price, a buyer would earn $\$100,000 - \$90,909 = \$9,091$ in interest, so the buyer's rate of return would be $\$9,091/\$90,909 = 0.10$, or 10 percent. Thus, if you believe that the interest rate is about to rise from 5 to 10 percent, you also believe the price of your bond is about to fall from \$95,238 to \$90,909—a drop of over \$4,300.

What would you do?

Logically, you would want to sell your bond *now*, before the price drops. If you still want to hold this type of bond later, you can always buy it back next week at the lower price, and gain from the transaction. Thus, if you expect the interest rate to rise in the future, you will want to exchange your bonds for money *today*. Your demand for money will increase.

Of course, if *you* expect the interest rate to drop, and your expectation is reasonable, others will probably feel the same way. They, too, will want to trade in their bonds for money. Thus, if the expectation is widespread, there will be an increase in the demand for money economy-wide.

A general expectation that interest rates will rise (bond prices will fall) in the future will cause the money demand curve to shift rightward in the present.

Notice that when people expect the interest rate to rise, we *shift* the money demand curve, rather than move along it. People will want to hold more money at any *current* interest rate.

INTEREST RATE EXPECTATIONS

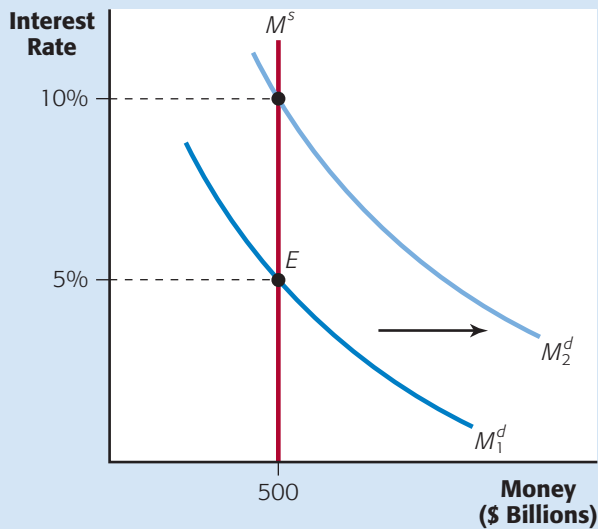


FIGURE 10

If households and firms expect the interest rate to rise in the future, their demand for money will increase today. Starting from equilibrium at point E , an expected increase in the interest rate from 5 percent to 10 percent will increase money demand to M_2^d . The result is a self-fulfilling prophecy: The interest rate increases to 10 percent *today*.

Figure 10 shows what will happen in the money market when people expect the interest rate to rise. Initially, with the money supply equal to \$500 billion, the equilibrium is at point E and the interest rate is 5 percent. But the expected rise in the interest rate shifts the money demand curve rightward. After the shift, there is an excess demand for money and an excess supply of bonds at the original interest rate of 5 percent. As the public attempts to sell bonds, the price of bonds will fall, which means the interest rate will rise.

How far will the interest rate rise? That depends. Imagine a simple case where *everyone* in the economy expected the interest rate to rise to 10 percent next week. Then *no one* would want to hold bonds at any *current* interest rate less than 10 percent. For example, if the interest rate rose to 9 percent, people would still expect it to rise further, so they would still want to sell their bonds. Therefore, to return the money market to equilibrium, the interest rate would rise to exactly the level that people expected. This is the case we've illustrated in Figure 10, where the money demand curve shifts rightward by just enough to raise the interest rate to 10 percent. More generally:

When the public as a whole expects the interest rate to rise in the future, they will drive up the interest rate in the present.

When information comes along that makes people believe that interest rates will rise and bond prices fall in the near future, the result is an immediate rise in the interest rate and a fall in bond prices. This principle operates even if the information is false and there is ultimately no reason for the interest rate to rise. Thus, a general expectation that interest rates will rise can be a *self-fulfilling prophecy*: Because people believe it, it actually happens. Their expectation alone is enough to drive up the interest rate.

This immediate response to information about the future—and the possibility of a self-fulfilling prophecy—works in the opposite direction as well:

When the public expects the interest rate to drop in the future, they will drive down the interest rate in the present.

In this case, the public expects bond prices to rise, so they try to shift their wealth from money to bonds. In Figure 10, the money demand curve would shift leftward (not shown). The price of bonds would rise, and the interest rate would fall, just as was originally expected.

MANAGING EXPECTATIONS

Changes in interest rates due to changes in expectations can have important consequences. First, fortunes can be won and lost depending on how people bet on the future. For example, suppose you believe the interest rate is about to drop, so you buy bonds, thinking their price is about to rise. But suppose the interest rate actually *rises* instead. Then your bonds will immediately drop in price and be worth less than what you paid for them. In fact, it is not unusual for major bondholders—such as pension funds or money market mutual funds—to gain or lose millions of dollars in a single day based on a good or a bad bet.

Another consequence is one we discussed earlier in this chapter: Changes in the interest rate affect aggregate expenditure, and therefore output. Fortunately, the Fed can counteract these changes with open market purchases or sales of bonds, as needed, and we'll discuss this a bit later.

Still, the public's ever-changing expectations about future interest rates make the Fed's job more difficult. Expectations can change interest rates, and changes in interest rates can affect individual fortunes as well as the economy as a whole. This observation helps explain some seemingly mysterious Fed behavior. Public policy statements made by the Fed's chair (currently Alan Greenspan) or by other Fed officials are remarkably tentative, and sometimes downright confusing. You can read them again and again and still have no idea what the Fed intends to do about interest rates in the future.

For example, on July 9, 1993, the Federal Reserve's Open Market Committee (FOMC) released a summary of the minutes of its May 1993 meeting. Here is the part of the statement explaining the Fed's future intentions regarding the money supply and interest rates. See if you can tell what the Fed planned to do.

In the view of a majority of the members . . . developments over recent months were sufficiently worrisome to warrant positioning policy for a move toward restraint should signs of continuing inflation continue to multiply. . . . Slightly greater reserve restraint would or slightly lesser reserve restraint might be acceptable.²

And here is the key sentence from a more recent FOMC statement, released after its May 1999 meeting:

. . . [T]he Committee recognizes the need to be alert to developments over coming months that might indicate that financial conditions may no longer be consistent with containing inflation.³

This is the kind of writing that gives English teachers indigestion. But from the Fed's point of view, the obfuscation is understandable. If the officials of the FOMC had given stronger hints about their thinking, the money and bond markets might have gone into overdrive, as people rushed to buy or sell bonds in order to profit (or avoid loss) from the Fed's action. On rare occasions, Fed officials—by speaking

² *New York Times*, July 10, 1993.

³ Federal Reserve Board Press Release, May 18, 1999 (<http://www.bog.frb.fed.us/BoardDocs/Press/General/1999/19990518/DEFAULT.HTM>).

more clearly—have given unintentionally strong hints and then had to quickly undo the damage with further statements or open market operations.

But secrecy about the Fed's intentions leads to surprises when the Fed finally acts, and surprises, too, create turmoil in the financial markets. In late 1998—after urging by government officials and the financial community—the Fed began an experiment: Immediately after its meetings, the FOMC would reveal if it had any significant “bias” toward either raising or lowering rates at its next meeting. (The Fed had been deciding on such a bias since the 1960s, but—until 1998—had kept the information secret until weeks later.) The idea was to cushion the blow when the interest rate move finally came, so that the financial markets would react more gradually. But the experiment failed because the public reacted as if the Fed's bias was really its *plan* for interest rates, despite Fed statements to the contrary. Thus, each announcement of bias caused a frenzy of activity in financial markets.

In February 2000, the Fed abandoned its experiment with *bias*, and began a new experiment: It would just state how the FOMC viewed *risks* to the economy, rather than hint at future changes in interest rates. This experiment represented a new kind of compromise between clarity and secrecy: The FOMC would inform the public of *which* policy direction was *more likely*, but provide no information on *how likely* the policy was. For example, here is the statement released by the FOMC immediately after its meeting on February 2, 2000, the first such release under the new experiment:

*Against the background of its long-run goals of price stability and sustainable economic growth and of the information currently available, the Committee believes the risks are weighted mainly toward conditions that may generate heightened inflation pressures in the foreseeable future.*⁴

As you'll see in the chapter after next, the Fed usually responds to inflationary dangers by raising interest rates. By stating that it viewed inflation as a greater danger than recession, the FOMC was informing the public that it was more likely to raise rates than to lower them. However, it was not saying that it *planned* to raise interest rates or even that a rise was likely. That would depend on how *important* the FOMC viewed the inflationary dangers, something that was not revealed in the statement.

⁴ Federal Reserve Board Press Release, February 2, 2000 (<http://www.bog.frb.fed.us/BoardDocs/Press/General/2000/20000202/DEFAULT.HTM>).



The minutes of the most recent FOMC meeting are available at <http://www.bog.frb.fed.us/FOMC/minutes>.

S U M M A R Y

The interest rate is a key macroeconomic variable. This chapter explores how the supply and demand for money interact to determine the interest rate in the short run, and how the Federal Reserve can adjust the money supply to change the interest rate.

An individual's demand for money indicates the fraction of wealth that person wishes to hold in the form of money, for different interest rates. Money is useful as a means of payment, but holding money means sacrificing the interest that could be earned by holding bonds instead. The higher the interest rate, the larger the fraction of their wealth people will hold in the form of bonds, and the smaller the fraction they will hold as money.

The demand for money is sensitive to the interest rate, but it also depends on the price level, real income, and expecta-

tions. An increase in the price level, higher real income, or an increase in the expected future interest rate can each shift the money demand curve to the right.

The money supply is under the control of the Fed and is independent of the interest rate. Equilibrium in the money market occurs at the intersection of the downward-sloping money demand curve and the vertical money supply curve. The interest rate will adjust so that the quantity of money demanded by households and firms just equals the quantity of money supplied by the Fed and the banking system.

Conditions in the money market mirror conditions in the bond market. If the interest rate is above equilibrium in the money market, there will be an excess supply of money there. People *want to* hold less money than they actually *do* hold,

which means that they wish to hold more bonds than they do hold. An excess supply of money means an excess demand for bonds. As people try to obtain more bonds, the price of bonds rises, and the interest rate falls. Thus, an excess supply of money will cause the interest rate to fall. Similarly, an excess demand for money will cause the interest rate to rise.

The Fed can increase the money supply through an open market purchase of bonds, and decrease it through an open market sale. An increase in the money stock creates an excess supply of money. Very quickly, the interest rate will fall so that the public is willing to hold the now-higher money supply. A decrease in the money stock will drive up the interest rate.

Changes in the interest rate affect interest-sensitive forms of spending—firms' spending on plant and equipment, new housing constructions, and households' purchases of "big ticket" consumer durables. By lowering the interest rate, the Fed can stimulate aggregate expenditures and increase GDP through the multiplier process.

Finally, expectations of future interest rate changes can become self-fulfilling prophecies, as well as create undesirable turmoil in financial markets.

KEY TERMS

wealth constraint
money demand curve

money supply curve
excess supply of money

excess demand for bonds

federal funds rate

REVIEW QUESTIONS

- Why do individuals choose to hold some of their wealth in the form of money? Besides individual tastes, what factors help determine how much money an individual holds?
- Why does the money demand curve slope downward? Which of the following result in a shift of the money demand curve and which result in a movement along the curve? If there is a shift, in which direction?
 - The Fed lowers interest rates.
 - The Fed raises interest rates.
 - The price level falls.
 - The price level rises.
 - Income increases.
 - Income decreases.
- Why is the economy's money supply curve vertical? What causes the money supply curve to shift?
- What sequence of events brings the money market to equilibrium if there is an excess supply of money? An excess demand for money?
- The text mentions that starting in June 1999 the Fed began selling government bonds, and as a result, the interest rate rose. Explain how the Fed's sale of bonds led to a lower interest rate.
- Describe how an increase in the interest rate affects spending on the following:
 - plant and equipment
 - new housing
 - consumer durables
- Does a change in expectations about the interest rate result in a shift in the money demand curve or a movement along it? Explain what happens in the money market when people expect the interest rate to fall.
- Why do we have both a short-run and a long-run theory of the interest rate? Briefly, what determines the interest rate in the short run? In the long run?

PROBLEMS AND EXERCISES

- Assume the demand deposit multiplier is 10. For each of the following, state the impact on the money supply curve (the direction it will shift, and the amount of the shift).
 - The Fed purchases bonds worth \$10 billion.
 - The Fed sells bonds worth \$5 billion.
- A bond promises to pay \$500 one year from now. For the following prices, find the corresponding interest payments and interest rates that the bond offers.

Price	Amount Paid		Interest Rate
	in One Year	Interest Payment	
\$375	\$500	_____	_____
\$425	\$500	_____	_____
\$450	\$500	_____	_____
\$500	\$500	_____	_____

As the price of the bond rises, what happens to the bond's interest rate?

3. “A general expectation that the interest rate will fall can be a self-fulfilling prophecy.” Explain what this means.
4. Suppose that, in an attempt to prevent the economy from overheating, the Fed raises the interest rate. Illustrate graphically, using a diagram similar to Figure 8 in this chapter, the effect on the money supply, interest rate, and GDP.
5. For each of the following events, state (1) the impact on the money demand curve, and (2) whether the Fed should increase or decrease the money supply if it wants to keep the interest rate unchanged. (*Hint*: It will help to draw a diagram of the money market for each case.)
 - a. People start making more of their purchases over the Internet, using credit cards.
 - b. Increasing fear of credit card fraud makes people stop buying goods over the Internet with credit cards, and discourages the use of credit cards in other types of purchases as well.
 - c. A new type of electronic account is created in which your funds are held in bonds up to the second you make a purchase. Then—when you buy something—just the right amount of bonds are transferred to the ownership of the seller. (*Hint*: Would you want to increase or decrease the amount of your wealth in the form of money after this new type of account were available?)
6. Determine whether monetary policy is *more* or *less* effective in changing GDP when autonomous consumption and investment spending are very sensitive to changes in the interest rate, and explain your reasoning.
7. In a later chapter, you will learn that a drop in the interest rate has *another* channel of influence on real GDP: It causes a depreciation of the dollar (that is, it makes the dollar cheaper to foreigners), which, in turn, increases our net exports.
 - a. When we take account of the effect on net exports, does a given change in the money supply have *more* or *less* of an impact on real GDP?
 - b. Suppose that the Fed wants to rein in real GDP as it did in late 1999 and early 2000. Should the Fed raise the interest rate by more or by less when it takes the impact on net exports into account (compared to the case of no impact on net exports)? Explain.

C H A L L E N G E Q U E S T I O N S

1. Determine whether *fiscal policy* is *more* or *less* effective in changing GDP when autonomous consumption and investment spending are very sensitive to changes in the interest rate, and explain your reasoning.
2. In Problem 7, you were asked how the *net export* effect changes the potency of monetary policy. Answer the same question about fiscal policy (that is, does the net export effect make fiscal policy more or less potent in changing GDP?).

E X P E R I E N T I A L E X E R C I S E

1. A favorite activity of many macroeconomists is “Fed watching.” Go to the Federal Reserve System’s Web site and look for the most recent Congressional testimony of the board chairman (<http://www.bog.frb.fed.us/boarddocs/testimony>). Is the Fed currently signalling that it will raise or lower interest rates, or that it will hold them constant? Is the Fed stating its intentions directly, or hiding them with vague language?





CHAPTER

26

AGGREGATE DEMAND AND AGGREGATE SUPPLY

CHAPTER OUTLINE

The Aggregate Demand Curve

The Price Level and the Money Market
Deriving the Aggregate Demand Curve
Understanding the *AD* Curve
Movements Along the *AD* Curve
Shifts of the *AD* Curve

The Aggregate Supply Curve

Prices and Costs in the Short Run
Deriving the Aggregate Supply Curve
Movements Along the *AS* Curve
Shifts of the *AS* Curve

AD and *AS* Together: Short-Run Equilibrium

What Happens When Things Change?

Demand Shocks in the Short Run
Demand Shocks: Adjusting to the Long Run
The Long-Run Aggregate Supply Curve
Supply Shocks

Some Important Provisos About the *AS* Curve

Using the Theory: The Recession and Recovery of 1990–92

Economic fluctuations are facts of life. If you need a reminder, look back at Figure 1 in the chapter titled “Economic Fluctuations.” There you can see that while potential GDP trends upward year after year—due to economic growth—*actual* GDP tends to rise above and fall below potential over shorter periods.

But the figure also reveals another important fact about the economy: Deviations from potential output don’t last forever. When output dips below or rises above potential, the economy returns to potential output after a few quarters or years. True, in some of these episodes, government policy—either fiscal or monetary—helped the economy return to full employment more quickly. But even without corrective policies—such as during long parts of the Great Depression of the 1930s—the economy shows a remarkable tendency to begin moving back toward potential output. Why? And what, exactly, is the mechanism that brings us back to our potential when we have strayed from it? These are the questions we will address in this chapter. And we’ll address them by studying the behavior of a variable that we’ve put aside for several chapters: the price level.

The chapter begins by exploring the relationship between the price level and output. This is a two-way relationship, as you can see in Figure 1 in *this* chapter. On the one hand, changes in the price level cause changes in real GDP. This causal relationship is illustrated by the *aggregate demand curve*, which we will discuss shortly. On the other hand, changes in real GDP cause changes in the price level. This relationship is summarized by the *aggregate supply curve*, to which we will turn later.

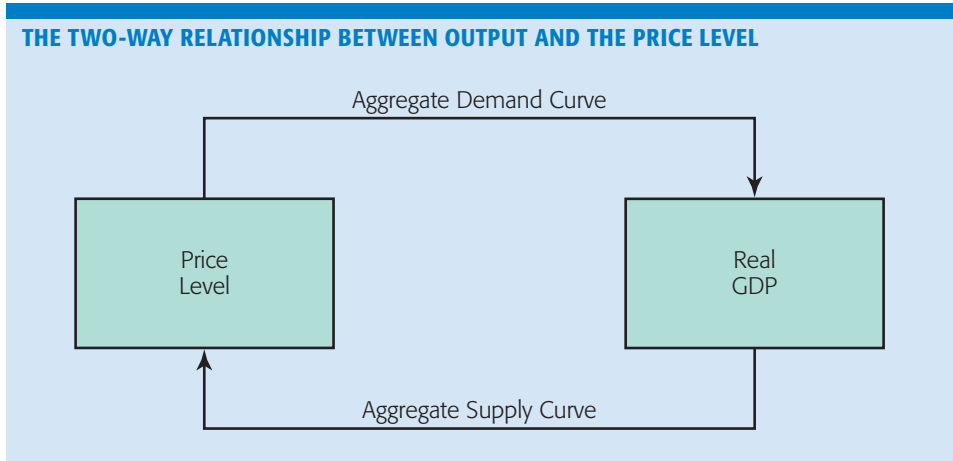
Once we’ve developed the aggregate demand and supply curves, we’ll be able to use them to understand how changes in the price level—sometimes gently, other times more harshly—steer the economy back toward potential output.

THE AGGREGATE DEMAND CURVE

In this section, we’ll focus on how changes in the price level affect equilibrium real GDP. We’ll postpone until later the question of *why* the price level might change.

THE TWO-WAY RELATIONSHIP BETWEEN OUTPUT AND THE PRICE LEVEL

FIGURE 1



THE PRICE LEVEL AND THE MONEY MARKET

Our first step in understanding how price level changes affect the economy is their impact in the money market. And this impact is straightforward: When the price level rises, the money demand curve shifts rightward. Why? Remember that the money demand curve tells us how much of their wealth people want to hold as money (as opposed to bonds) at each interest rate. People hold bonds because of the interest they pay; people hold money because of its convenience. Each day, as we make purchases, we need cash or funds in our checking account to pay for them. If the price level rises, and the average purchase becomes more expensive, we'll need to hold more of our wealth as money just to achieve the same level of convenience. Thus, at any given interest rate, the demand for money increases, and the money demand curve shifts rightward.

The shift in money demand, and its impact on the economy, is illustrated in Figure 2. Panel (a) has our familiar money market diagram. We'll assume that, initially, the price level in the economy is equal to 100. With this price level, the money market is in equilibrium at point *E*, with an interest rate of 6 percent.

In panel (b), equilibrium GDP is at point *E*, with output equal to \$10 trillion. The aggregate expenditure line is marked " $r = 6\%$," which is the equilibrium interest rate we just found in the money market.

Now let's imagine a rather substantial rise in the price level—from 100 to 140. What will happen in the economy? The initial impact is in the money market. The money demand curve will start to shift rightward, and the interest rate will rise. Next, in panel (b), the higher interest rate decreases interest-sensitive spending—business investment, new housing, and consumer durables. The aggregate expenditure line shifts downward, and equilibrium real GDP decreases. All of these changes continue until we reach a new, consistent equilibrium in both panels. Compared with our initial position, this new equilibrium has the following characteristics:

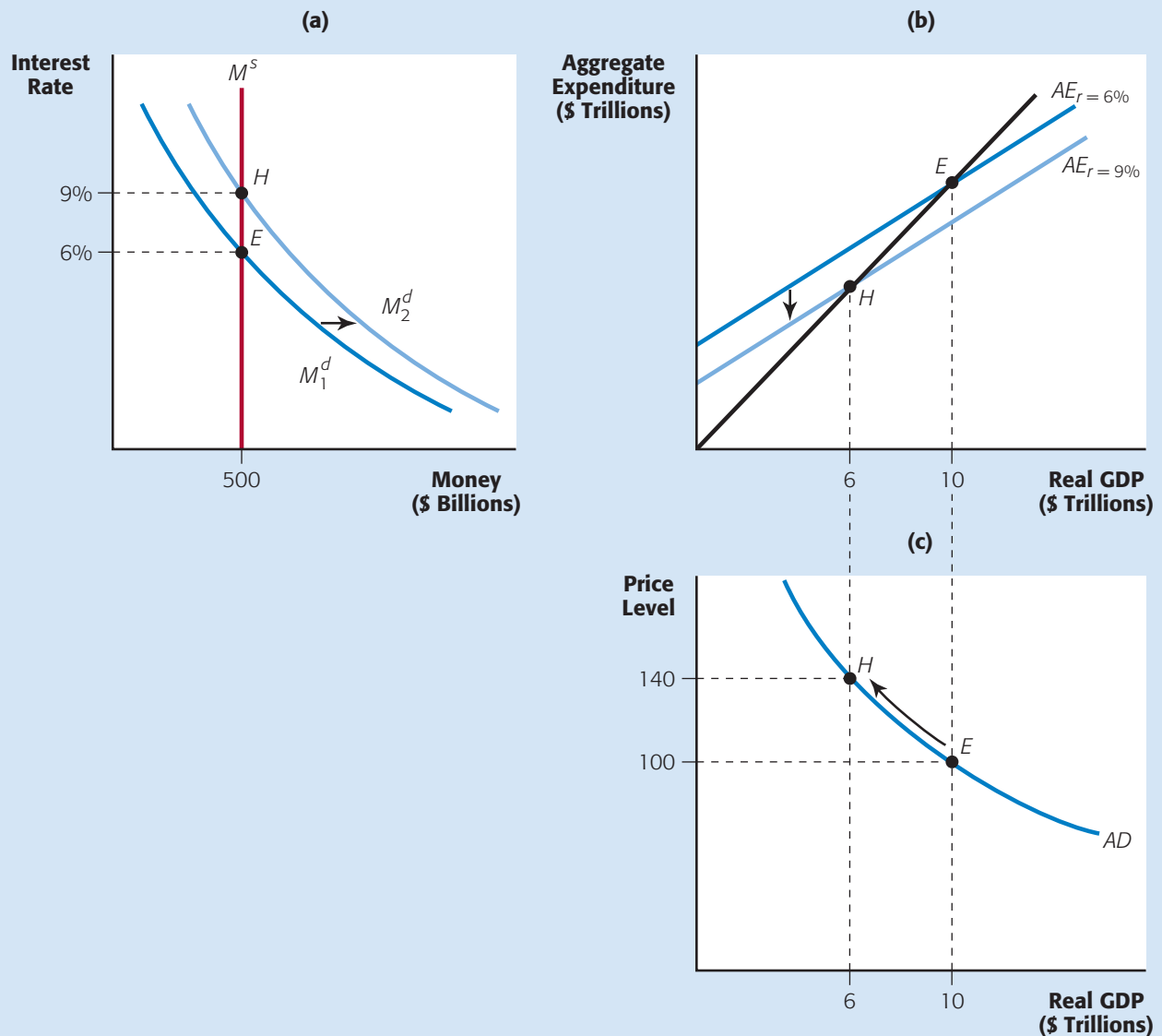
- The money demand curve has shifted rightward.
- The interest rate is higher.
- The aggregate expenditure line has shifted downward.
- Equilibrium GDP is lower.

Remember that all of these changes are caused by a rise in the price level.

The points labeled *H* in panels (a) and (b) show one possible new equilibrium consistent with these requirements. In panel (a), the money demand curve has

FIGURE 2

DERIVING THE AGGREGATE DEMAND CURVE



Initially, the money market is in equilibrium at point E in panel (a), and aggregate expenditure equals real GDP at point E in panel (b). That price-output combination determines point E in panel (c). A higher price level—140—increases money demand, raises the interest rate, reduces interest-sensitive spending, and lowers aggregate expenditure. Through the multiplier process, equilibrium real GDP falls to \$6 trillion. The new price-output combination determines point H in panel (c). Connecting points like E and H yields the downward-sloping aggregate demand (AD) curve.

shifted to M_2^d . The interest rate has risen to 9 percent. The aggregate expenditure line has shifted downward, to the one marked “ $r = 9\%$.” Finally, equilibrium output has fallen to \$6 trillion.

Now recall the initial event that caused real GDP to fall: a rise in the price level. We’ve thus established an important principle:

A rise in the price level causes a decrease in equilibrium GDP.

DERIVING THE AGGREGATE DEMAND CURVE

In panel (c), we introduce a new curve that summarizes the negative relationship between the price level and equilibrium GDP more directly. In this panel, the price level is measured along the vertical axis, while real GDP is on the horizontal. Point *E* represents our initial equilibrium, with $P = 100$ and equilibrium GDP = \$10 trillion. Point *H* represents the new equilibrium, with $P = 140$ and equilibrium GDP = \$6 trillion. If we continued to change the price level to other values—raising it further to 150, lowering it to 85, and so on—we would find that each different price level results in a different equilibrium GDP. This is illustrated by the downward-sloping curve in the figure, which we call the *aggregate demand curve*.

The aggregate demand (AD) curve tells us the equilibrium real GDP at any price level.

Aggregate demand (AD) curve A curve indicating equilibrium GDP at each price level.

UNDERSTANDING THE AD CURVE

The *AD* curve is unlike any other curve you've encountered in this text. In all other cases, our curves have represented simple behavioral relationships. For example, the demand curve for maple syrup shows us how a change in price affects the behavior of buyers in a market. Similarly, the aggregate expenditure line shows how a change in income affects total spending in the economy.

But the *AD* curve represents more than just a behavioral relationship between two variables. Each point on the curve represents a short-run *equilibrium* in the economy. For example, point *E* on the *AD* curve in Figure 2 tells us that when the price level is 100, *equilibrium* GDP is \$10 trillion. Thus, point *E* doesn't just tell us that total spending is \$10 trillion; rather, it tells us that when $P = 100$, and when spending and output have the same value, then *both* are equal to \$10 trillion.

As you can see, a better name for the *AD* curve would be the "equilibrium output at each price level" curve—not a very catchy name. The *AD* curve gets its name because it *resembles* the demand curve for an individual product. It's a downward-sloping curve, with the price level (instead of the price of a single good) on the vertical axis and *equilibrium total output* (instead of the quantity of a single good demanded) on the horizontal axis. But there the similarity ends. The *AD* curve is not a demand curve at all, in spite of its name.



Watch out for two very common mistakes about the aggregate demand curve. The first is thinking that it is simply a "total demand" or "total spending" curve for the economy, telling us the total quantity of output that purchasers want to buy at each price level. This is an oversimplification. Rather, the *AD* curve tells us the *equilibrium real GDP* at each price level.

Remember that equilibrium GDP is the level of output at which total spending *equals* total output. Thus, total spending is only part of the story behind the *AD* curve: the other part is the requirement that total spending and total output be equal.

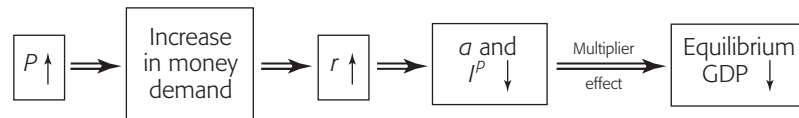
A second, related mistake is thinking that the *AD* curve slopes downward for the same reason that a microeconomic demand curve slopes downward. This, too, is wrong: *microeconomic* demand curves for individual products rely on an entirely different mechanism than the one we've described for the *AD* curve. In the market for maple syrup, for example, a rise in price causes quantity demanded to decrease mostly because people switch to *other* goods that are now relatively cheaper. But along the *AD* curve, a rise in the price level generally causes the prices of all goods to increase *together*. In this case, there are no relatively cheaper goods to switch to!

The *AD* curve works in an entirely different way from microeconomic demand curves. Along the *AD* curve, an increase in the price level raises the interest rate in the money market, which decreases spending on interest-sensitive goods, causing a drop in equilibrium GDP.

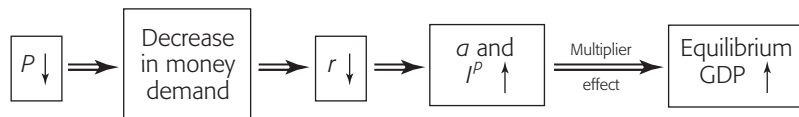
MOVEMENTS ALONG THE AD CURVE

As you will see later in this chapter, a variety of events can cause the price level to change, and move us *along* the AD curve. It's important to understand what happens in the economy as we make such a move.

Look again at the AD curve in panel (c) of Figure 2. Suppose the price level rises, and we move from point E to point H along this curve. Then the following sequence of events occurs: The rise in the price level increases the demand for money, raises the interest rate, decreases autonomous consumption (a) and investment spending (I^p), and works through the multiplier to decrease equilibrium GDP. The process can be summarized as follows:



The opposite sequence of events will occur if the price level falls, moving us rightward along the AD curve:



SHIFTS OF THE AD CURVE

When we move along the AD curve in Figure 2, we assume that the price level changes, but that other influences on equilibrium GDP are constant. When any of these other influences on GDP changes, the AD curve will shift. The distinction between movements along the AD curve and shifts of the curve itself is very important. Always keep the following rule in mind:

When a change in the price level causes equilibrium GDP to change, we move along the AD curve. Whenever anything other than the price level causes equilibrium GDP to change, the AD curve itself shifts.

What are these other influences on GDP? They are the very same changes you learned about in previous chapters. Specifically, equilibrium GDP will change whenever there is a change in any of the following:

- government purchases
- taxes
- autonomous consumption spending
- investment spending
- net exports
- the money supply

Let's consider some examples and see how each causes the AD curve to shift.

Spending Shocks. Spending shocks initially affect the economy by shifting the aggregate expenditure line. Here, we'll see how these spending shocks—which you've encountered several times in this book already—shift the AD curve.

In Figure 3, we assume that the economy begins at a price level of 100. In the money market (not shown), the equilibrium interest rate is 6 percent, and equilib-

A SPENDING SHOCK SHIFTS THE AD CURVE

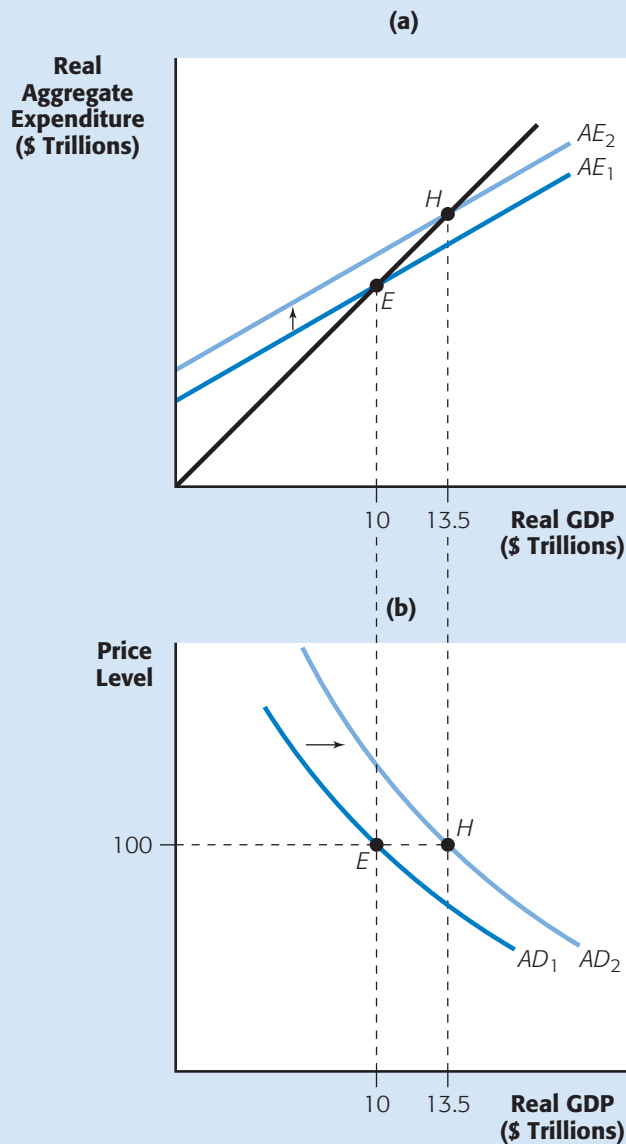


FIGURE 3

A positive spending shock, such as an increase in government purchases, shifts the aggregate expenditure line upward in panel (a), leading to a new level of equilibrium GDP. At each price level, GDP is higher than before, indicating that the AD curve has shifted to the right.

rium output—given by point E in panel (a)—is \$10 trillion. Panel (b) shows the same equilibrium as represented by point E on AD_1 .

Now let's repeat an experiment from the previous chapter: We'll increase government purchases by \$2 trillion and ask what happens if the price level remains at 100. If you flip back to Figure 9 in the previous chapter, you'll see that this rise in government purchases caused the AE line to shift upward, but it also caused the equilibrium interest rate to rise to 8 percent, causing the AE line to shift back downward a bit. The result was that equilibrium GDP rose to \$13.5 trillion. This new equilibrium is also shown in panel (a) of Figure 3. The aggregate expenditure line shifts upward to AE_2 , and the equilibrium moves to point H . With the price level remaining at 100, equilibrium GDP increases.

Now look at panel (b) in Figure 3. There, the new equilibrium is represented by point H ($P = 100$, real GDP = \$13.5 trillion). This point lies to the right of our original curve AD_1 . Point H , therefore, must lie on a *new AD curve*—a curve that tells us equilibrium GDP at any price level *after the increase in government spending*. The new AD curve is the one labeled AD_2 , which goes through point H . What about the other points on AD_2 ? They tell us that, if we had started at any *other* price level, an increase in government spending would have increased equilibrium GDP at that price level, too. We conclude that *an increase in government purchases shifts the entire AD curve rightward*.

Other spending shocks that shift the aggregate expenditure line upward shift the AD curve rightward, just as in Figure 3. More specifically,

the AD curve shifts rightward when government purchases, investment spending, autonomous consumption spending, or net exports increase, or when taxes decrease.

Our analysis also applies in the other direction. For example, at any given price level, a *decrease* in government spending shifts the aggregate expenditure line *downward*, decreasing equilibrium GDP. This in turn shifts the AD curve leftward.

More generally,

the AD curve shifts leftward when government purchases, investment spending, autonomous consumption spending, or net exports decrease, or when taxes increase.

Changes in the Money Supply. Changes in the money supply will also shift the aggregate demand curve. To see why, let's imagine that the Fed conducts open market operations to *increase* the money supply. As you learned in the previous chapter, this will cause the interest rate to decrease, increasing investment spending and autonomous consumption spending. Together, these spending changes will shift the aggregate expenditure line upward, just as in the panel (a) of Figure 3, and increase equilibrium GDP. Since this change in equilibrium output is caused by something *other* than a change in the price level, the AD curve shifts. In this case, because the money supply *increased*, the AD curve shifts *rightward*, just as in panel (b) of Figure 3.

A decrease in the money supply would have the opposite effect: The interest rate would rise, the aggregate expenditure line would shift downward, and *equilibrium GDP at any price level would fall*. We conclude that

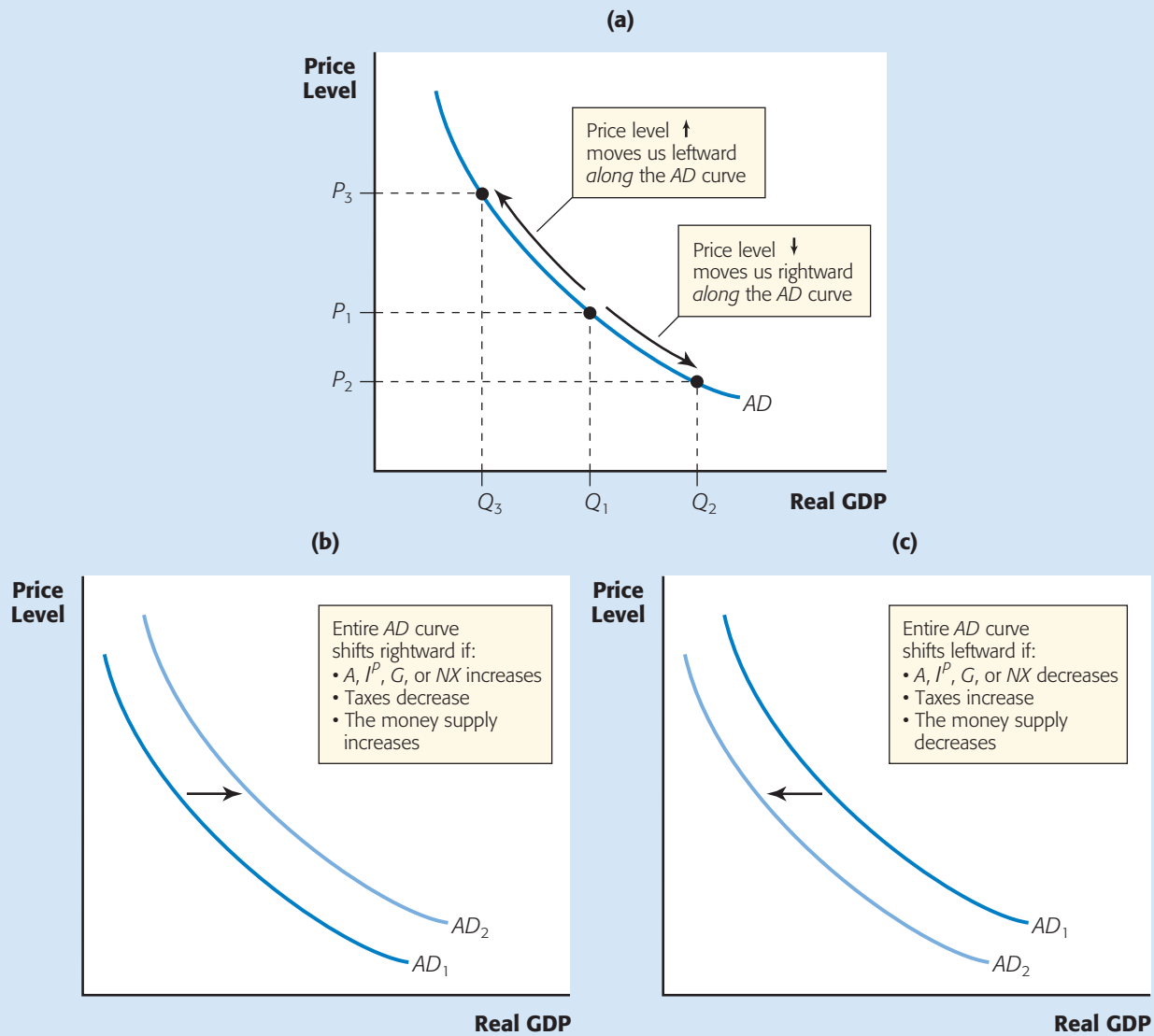
an increase in the money supply shifts the AD curve rightward. A decrease in the money supply shifts the AD curve leftward.

Shifts vs. Movements Along the AD Curve: A Summary. Figure 4 summarizes how some events in the economy cause a movement along the AD curve, and other events shift the AD curve. You can use the figure as an exercise, drawing diagrams similar to Figures 2 and 3 to illustrate why we move along or shift the AD curve in each case.

Notice that panels (b) and (c) of Figure 4 tell us how a variety of events affect the AD curve, but *not* how they affect *real* GDP. The reason is that, even if we know which AD curve we are on, we could be at *any point* along that curve, depending on the price level.

EFFECTS OF KEY CHANGES ON THE AGGREGATE DEMAND CURVE

FIGURE 4



But how is the price level determined? Our first step in answering that question is to understand the other side of the relationship between GDP and the price level.

THE AGGREGATE SUPPLY CURVE

Look back at Figure 1, which illustrates the *two-way* relationship between the price level and output. On the one hand, changes in the price level affect output. This is the relationship—summarized by the *AD curve*—that we explored in the previous section. On the other hand, changes in output affect the price level. This relationship—summarized by the *aggregate supply curve*—is the focus of this section.

The effect of changes in output on the price level is complex, involving a variety of forces. Current research is helping economists get a clearer picture of this relationship. Here, we will present a simple model of the aggregate supply curve that focuses on the link between prices and costs. Toward the end of the chapter, we'll discuss some additional ideas about the aggregate supply curve.



Burger King, like other firms in the economy, charges a markup over its costs per unit. The average markup in the economy is determined by competitive conditions, and tends to change slowly over time.

PRICES AND COSTS IN THE SHORT RUN

The price *level* in the economy results from the pricing behavior of millions of individual business firms. In any given year, some of these firms will raise their prices, and some will lower them. For example, during the 1990s, personal computers and long-distance telephone calls came down in price, while college tuition and the prices of movies rose. These types of price changes are subjects for *microeconomic* analysis, because they involve individual markets.

But often, all firms in the economy are affected by the same *macroeconomic* event, causing prices to rise or fall throughout the economy. This change in the price *level* is what interests us in macroeconomics.

To understand how macroeconomic events affect the price level, we begin with a very simple assumption:

A firm sets the price of its products as a markup over cost per unit.

For example, if it costs Burger King \$2.00, on average, to produce a Whopper (cost per unit is \$2.00), and Burger King's percentage markup is 10 percent, then it will charge $\$2.00 + (0.10 \times \$2.00) = \$2.20$ per Whopper.¹

The percentage markup in any particular industry will depend on the degree of competition there. If there are many firms competing for customers in a market, all producing very similar products, then we can expect the markup to be relatively small. Thus, we expect a relatively low markup on fast-food burgers or personal computers. In industries where there is less competition—such as daily newspapers or jet aircraft—we would expect higher percentage markups.

In macroeconomics, we are not concerned with how the markup differs in different industries, but rather with the *average percentage markup* in the economy:

The average percentage markup in the economy is determined by competitive conditions in the economy. The competitive structure of the economy changes very slowly, so the average percentage markup should be somewhat stable from year to year.

But a stable markup does not necessarily mean a stable price level, because unit costs can change. For example, if Burger King's markup remains at 10 percent, but the unit cost of a Whopper rises from \$2.00 to \$3.00, then the price of a Whopper will rise to $\$3.00 + (0.10 \times \$3.00) = \$3.30$. Extending this example to all firms in the economy, we can say:

In the short run, the price level rises when there is an economy-wide increase in unit costs, and the price level falls when there is an economy-wide decrease in unit costs.

¹ In microeconomics, you learn more sophisticated theories of how firms' prices are determined. But our simple markup model captures a central conclusion of those theories: that an increase in costs will result in higher prices.

Our primary concern in this chapter is the impact of *output* on unit costs and, therefore, on the price level. Why should a change in output affect unit costs and the price level? We'll focus on three key reasons.

As total output increases:

Greater amounts of inputs may be needed to produce a unit of output. As output increases, firms hire new, untrained workers who may be less productive than existing workers. Firms also begin using capital and land that are less well suited to their industry. As a result, greater amounts of labor, capital, land, and raw materials are needed to produce each unit of output. Even if the prices of these inputs remain the same, unit costs will rise. For example, imagine that Intel increases its output of computer chips. Then it will have to be less picky about the workers it employs, hiring some who are less well suited to chip production than those already working there. Thus, more labor hours will be needed to produce each chip. Intel may also have to begin using older, less-efficient production facilities, which require more silicon and other raw materials per chip. Even if the prices of all of these inputs remain unchanged, unit costs will rise.

The prices of non-labor inputs rise. This is especially true of inputs like land and natural resources, which may be available only in limited quantities in the short run. An increase in the output of final goods raises the demand for these inputs, causing their prices to rise. Firms that produce final goods experience an increase in unit costs, and raise their own prices accordingly.

The nominal wage rate rises. Greater output means higher employment, leaving fewer unemployed workers looking for jobs. As firms compete to hire increasingly scarce workers, they must offer higher nominal wage rates to attract them. Higher nominal wages increase unit costs, and therefore result in a higher price level. Notice that we use the nominal wage, rather than the real wage we've emphasized elsewhere in this book. That's because we are interested in explaining how firms' prices are determined. Since price is a nominal variable, it will be marked up over *nominal* costs.

A decrease in output affects unit costs through the same three forces, but with the opposite result. As output falls, firms can be more selective in hiring the best, most efficient workers and in choosing other inputs, decreasing input requirements per unit of output. Decreases in demand for land and natural resources will cause their prices to drop. And as unemployment rises, wages will fall as workers compete for jobs. All of these contribute to a drop in unit costs, and a decrease in the price level.

All three of our reasons are important in explaining why a change in output affects the price level. However, they operate within different time frames. When total output increases, new, less-productive workers will be hired rather quickly. Similarly, the prices of certain key inputs—such as lumber, land, oil, and wheat—may rise within a few weeks or months.

But our third explanation—changes in the nominal wage rate—is a different story. While wages in some lines of work might respond very rapidly, we can expect wages in many industries to change very little or not at all for a year or more after a change in output.

For a year or so after a change in output, changes in the average nominal wage are less important than other forces that change unit costs.

Here are some of the more important reasons why wages in many industries respond so slowly to changes in output:

- Many firms have union contracts that specify wages for up to three years. While wage increases are often built into these contracts, a rise in output will not affect the wage increase. When output rises or falls, these firms continue to abide by the contract.
- Wages in many large corporations are set by slow-moving bureaucracies.
- Wage changes in either direction can be costly to firms. Higher wages must be widely publicized in order to raise the number of job applicants at the firm. Lower wages can reduce the morale of workers—and their productivity. Thus, many firms are reluctant to change wages until they are reasonably sure that any change in demand for their output will be long lasting.
- Firms may benefit from developing reputations for paying stable wages. A firm that raises wages when output is high and labor is scarce may have to lower wages when output is low and labor is plentiful. Such a firm would develop a reputation for paying unstable wages, and have difficulty attracting new workers.

In this section, we focus exclusively on the short run—a time horizon of a year or so after a change in output. Since the average wage rate changes very little over the short run, we'll make the following simplifying assumption: *The nominal wage rate is fixed in the short run.* More specifically,

we assume that changes in output have no effect on the nominal wage rate in the short run.²

Keep in mind, though, that our assumption of a constant wage holds only in the *short run*. As you will see later, wage changes play a very important role in the economy's adjustment over the long run.

Since we assume a constant nominal wage in the short run, a change in output will affect unit costs through the other two factors we mentioned earlier. Specifically, in the short run, a rise in real GDP raises firms' unit costs because (1) the prices of non-labor inputs rise, and (2) input requirements per unit of output rise. With a constant percentage markup, the rise in unit costs translates into a rise in the price level. Thus,

in the short run, a rise in real GDP, by causing unit costs to increase, will also cause a rise in the price level.

In the other direction, a *drop* in real GDP lowers unit costs because (1) the prices of non-labor inputs fall, and (2) input requirements per unit of output fall. With a constant percentage markup, the drop in unit costs translates into a fall in the price level.

In the short run, a fall in real GDP, by causing unit costs to decrease, will also cause a decrease in the price level.

DERIVING THE AGGREGATE SUPPLY CURVE

Figure 5 summarizes our discussion about the effect of output on the price level in the short run. Suppose the economy begins at point A, with output at \$10 trillion

² This simplifying assumption is not entirely realistic. In some industries, wages will respond to changes in output, at least somewhat, even in the short run. However, assuming that the nominal wage remains constant in the short run makes our model much simpler, without affecting any of our essential conclusions.

THE AGGREGATE SUPPLY CURVE

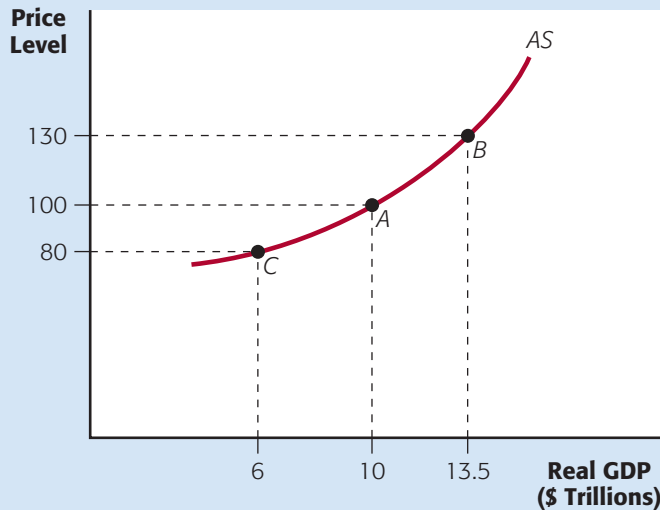


FIGURE 5

Beginning at point *A*, an increase in output will raise unit costs. For given percentage markups, firms will raise the prices they charge. An increase in output from \$10 trillion to \$13.5 trillion might raise the price level from 100 to 130 at point *B*. A decrease in output would lower unit costs and lead firms to lower their prices. The price level might fall to 80 at point *C*. Connecting points such as *A*, *B*, and *C* traces out the economy's AS curve.

and the price level at 100. Now suppose that output rises to \$13.5 trillion. What will happen in the short run? Even though wages are assumed to remain constant, the price level will rise because of the other forces we've discussed. In the figure, the price level rises to 130, indicated by point *B*. If, instead, output *fell* to \$7 trillion, the price level would fall—to 80 in the figure, indicated by point *C*.

As you can see, each time we change the level of output, there will be a new price level in the short run, giving us another point on the figure. If we connect all of these points, we obtain the economy's *aggregate supply curve*:

The aggregate supply curve (or AS curve) tells us the price level consistent with firms' unit costs and their percentage markups at any level of output over the short run.

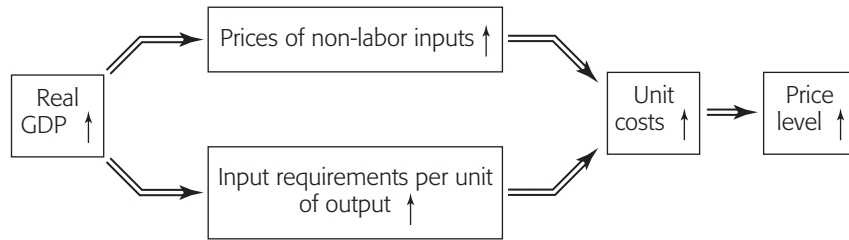
A more accurate name for the AS curve would be the “short-run-price-level-at-each-output-level” curve, but that is more than a mouthful. The AS curve gets its name because it *resembles* a microeconomic market supply curve. Like the supply curve for maple syrup we discussed in Chapter 3, the AS curve is upward sloping, and it has a price variable (the price level) on the vertical axis, and a quantity variable (total output) on the horizontal axis. But there, the similarity ends.

MOVEMENTS ALONG THE AS CURVE

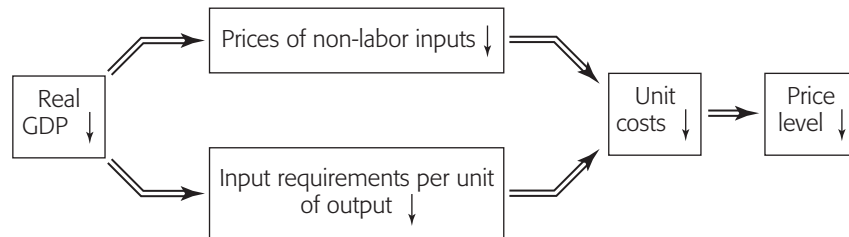
When a change in output causes the price level to change, we *move along* the economy's AS curve. But what happens in the economy as we make such a move?

Look again at the AS curve in Figure 5. Suppose we move from point *A* to point *B* along this curve in the short run. The increase in output raises the prices of raw materials and other (non-labor) inputs and also raises input requirements per unit of output at many firms. Both of these changes increase costs per unit. As long as the markup remains somewhat stable, the rise in unit costs will lead firms to raise their prices, and the price level will increase. Thus, as we move upward along the AS curve, we can represent what happens as follows:

Aggregate supply (AS) curve A curve indicating the price level consistent with firms' unit costs and markups for any level of output over the short run.



The opposite sequence of events occurs when real GDP falls, moving us downward along the AS curve:



SHIFTS OF THE AS CURVE

When we drew the AS curve in Figure 5, we assumed that a number of important variables remained unchanged. In particular, we assumed that the only changes in unit costs were those caused by a change in output. But in the real world, unit costs sometimes change for reasons *other* than a change in output. When this occurs, unit costs—and the price level—will change at *any* level of output, so the AS curve will shift.

In general, we distinguish between a movement along the AS curve, and a shift of the curve itself, as follows:

When a change in real GDP causes the price level to change, we move along the AS curve. When anything other than a change in real GDP causes the price level to change, the AS curve itself shifts.



A common mistake about the AS curve is thinking that it describes the same kind of relationship between price and quantity as a microeconomic supply curve. There are two reasons why this is wrong.

First, the direction of causation between price and output is reversed for the AS curve. For example, when we draw the supply curve for maple syrup, we view changes in the price of maple syrup as causing a change in output supplied. But along the AS curve, a change in output causes a change in the price level.

Second, the basic assumption behind the AS curve is very different from that behind a single market supply curve. When we draw the supply curve for an individual product, we assume that the prices of inputs used in producing the good remain fixed. This is a sensible thing to do, because an increase in production for a single good is unlikely to have much effect on input prices in the economy as a whole.

But when we draw the AS curve, we imagine an increase in *real GDP*, in which *all* firms are increasing their output. This will significantly raise the demand for inputs, so it is unrealistic to assume that input prices will remain fixed. Indeed, the rise in input prices as output increases is one of the important reasons for the AS curve's upward slope.

Figure 6 illustrates the logic of a shift in the AS curve. Suppose the economy's initial AS curve is AS_1 . Now suppose that some economic event *other* than a change in output—for the moment, we'll leave the event unnamed—causes firms to raise their prices. Then the price level will be higher at *any* level of output we might imagine, so the AS curve must shift *upward*—for example, to AS_2 in the figure. At an output level of \$10 trillion, the price level would rise from 100 to 140. At any other output level, the price level would also rise.

SHIFTS OF THE AGGREGATE SUPPLY CURVE

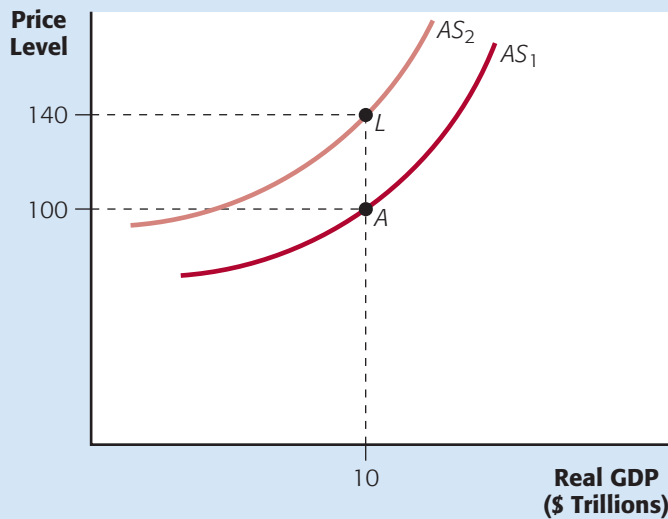


FIGURE 6

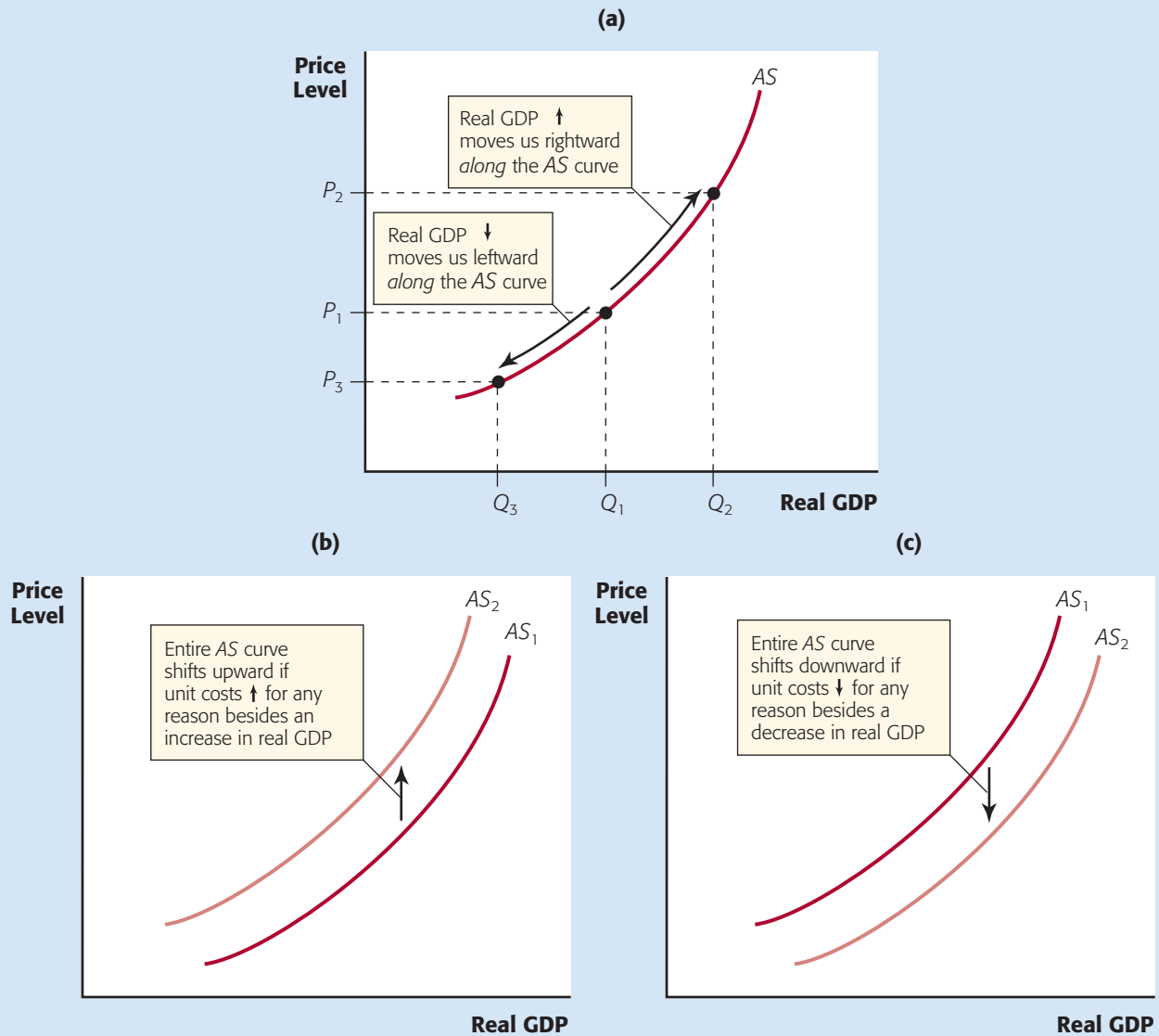
Any factor that changes firms' unit costs at any output level will shift the AS curve. For example, an increase in world oil prices or bad weather would shift the AS curve upward; the price level would be higher at each level of real GDP.

What can cause unit costs to change at any given level of output? The following are some important examples:

- *Changes in world oil prices.* Oil is traded on a world market, where prices can fluctuate even while output in the United States does not. And changes in world oil prices have caused major shifts in the AS curve. Three events over the past few decades—an oil embargo by Arab oil-producing nations in 1973–74, the Iranian revolution in 1978–79, Iraq's invasion of Kuwait in 1990—all caused large jumps in the price of oil. Each time, costs per unit rose for firms across the country, and they responded by charging higher prices than before for *any* output level they might produce. As in Figure 6, the AS curve shifted upward. Conversely, in 1991, the price of oil decreased dramatically. This caused unit costs to decrease at many firms, shifting the AS curve downward.
- *Changes in the weather.* Good crop-growing weather increases farmers' yields for any given amounts of land, labor, capital, and other inputs used. This decreases farms' unit costs, and the price of agricultural goods falls. Since many of these goods are final goods (such as fresh fruit and vegetables), the price drop will contribute directly to a drop in the price level, and a downward shift of the AS curve. Additionally, agricultural products are important inputs in the production of many other goods. (For example, corn is an input in beef production.) Good weather thus leads to a drop in input prices for many other firms in the economy, causing their unit costs—and their prices—to decrease. For these reasons, we can expect good weather to shift the AS curve downward. Bad weather, which decreases crop yields, increases unit costs at any level of output, and shifts the AS curve upward.
- *Technological change.* New technologies can enable firms to produce any given level of output at lower unit costs. In recent years, for example, we've seen revolutions in telecommunications, information processing, and medicine. The result has been steady downward shifts of the AS curve.
- *Adjustment to the Long Run.* We've assumed that, in the short run, the nominal wage remains unchanged as output changes. But as we extend our time horizon

FIGURE 7

EFFECTS OF KEY CHANGES ON THE AGGREGATE SUPPLY CURVE



beyond the first year after a change in output, our assumption of a constant wage becomes increasingly unrealistic. As you will see a bit later, when output is above its full-employment level, we can expect nominal wage rates to rise. This is part of the long-run adjustment process in the economy. Similarly, if output is below potential, wage rates will eventually fall. These adjustments in wages shift the AS curve, since we assumed a constant wage when we drew the curve.

Figure 7 summarizes how different events in the economy cause a movement along, or a shift in, the AS curve. But the AS curve tells only half of the economy's story: It shows us the price level *if* we know the level of output. The AD curve tells the other half of the story: It shows us the level of output *if* we know the economy's price level. In the next section, we finally put the two halves of the story together, allowing us to determine both the price level and output.

SHORT-RUN MACROECONOMIC EQUILIBRIUM

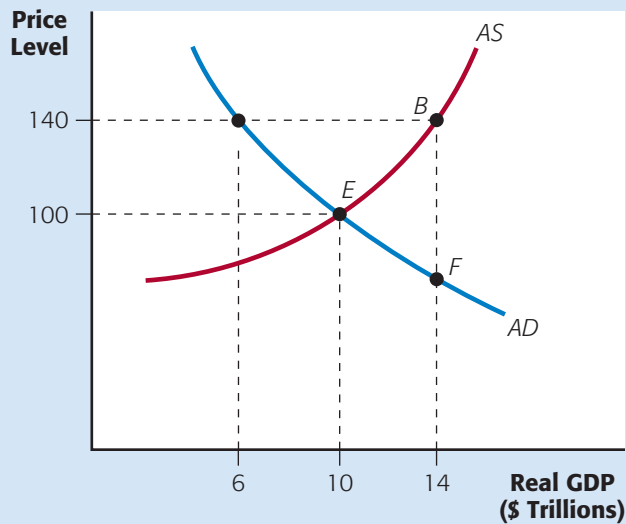


FIGURE 8

Short-run equilibrium occurs where the *AD* and *AS* curves intersect. At point *E*, the price level of 100 is consistent with an output of \$10 trillion along the *AD* curve. The output level of \$10 trillion is consistent with a price level of 100 along the *AS* curve. At any other combination of price level and output, such as point *F*, at least one condition for equilibrium will not be satisfied.

AD AND AS TOGETHER: SHORT-RUN EQUILIBRIUM

Where will the economy settle in the short run? That is, where is our **short-run macroeconomic equilibrium**? Figure 8 shows how to answer that question, using both the *AS* curve and the *AD* curve. If you suspect that the equilibrium is at point *E*, the intersection of these two curves, you are correct. At that point, the price level is 100, and output is \$10 trillion. But it's worth thinking about *why* point *E*—and only point *E*—is our short-run equilibrium.

First, we know that the economy must be at some point on the *AD* curve. Otherwise, real GDP would not be at its equilibrium value. For example, suppose the economy were at point *B*, which lies to the right of the *AD* curve. At this point, the price level is 140, and output is \$14 trillion. But the *AD* curve tells us that with a price level of 140, *equilibrium* output is \$6 trillion. Thus, at point *B*, real GDP would be greater than its equilibrium value. As you learned several chapters ago, this situation cannot persist for long, since inventories would pile up, and firms would be forced to cut back on their production. Thus, point *B* cannot be our short-run equilibrium.

Second, short-run equilibrium requires that the economy be operating on its *AS* curve. Otherwise, firms would not be charging the prices dictated by their unit costs and the average percentage markup in the economy. For example, point *F* lies *below* the *AS* curve. But the *AS* curve tells us that if output is \$14 trillion, based on the average percentage markup and unit costs, the price level should be 140 (point *B*), not something lower. That is, the price level at point *F* is *too low* for equilibrium. This situation will not last long either.

We could make a similar argument for other points that are off the *AS* and *AD* curves, always coming to the same conclusion: Unless the economy is on *both* the *AS* and the *AD* curves, the price level and the level of output will change. Only when the economy is at point *E*—on *both* curves—will we have reached a sustainable level of real GDP and the price level.



Find the Equilibrium

Short-run macroeconomic equilibrium A combination of price level and GDP consistent with both the *AD* and *AS* curves.

WHAT HAPPENS WHEN THINGS CHANGE?

Now that we know how the short-run equilibrium is determined, and armed with our knowledge of the *AD* and *AS* curves, we are ready to put the model through its paces. In this section, we'll explore how different types of events cause the short-run equilibrium to change.

Our short-run equilibrium will change when either the *AD* curve, the *AS* curve, or both, *shift*. Since the consequences for the economy are very different for shifts in the *AD* curve as opposed to shifts in the *AS* curve, economists have developed a shorthand language to distinguish between them:

Demand shock Any event that causes the *AD* curve to shift.

Supply shock Any event that causes the *AS* curve to shift.

An event that causes the AD curve to shift is called a demand shock. An event that causes the AS curve to shift is called a supply shock.

In this section, we'll first explore the effects of demand shocks, both in the short run and during the adjustment process to the long run. Then, we'll take up the issue of supply shocks.

What Happens When
Things Change?



DEMAND SHOCKS IN THE SHORT RUN

Demand shocks can be caused by spending shocks or by changes in monetary policy. In both cases, the *AD* curve shifts. Figure 4, which lists the reasons for a shift in the *AD* curve, also serves as a list of demand shocks to the economy. Let's consider some examples.

An Increase in Government Purchases. You've learned that an increase in government purchases shifts the *AD* curve rightward. Now we can see how it affects the economy in the short run. Figure 9 shows the initial equilibrium at point *E*, with the price level equal to 100 and output at \$10 trillion. Now, suppose that government purchases rise by \$2 trillion. Figure 4 (b) tells us that the *AD* curve will shift rightward. What will happen to equilibrium GDP?

In our example in the previous chapter, a \$2 trillion rise in government purchases increased output to \$13.5 trillion, and also raised the interest rate in the money market to 8 percent. (Flip back to Figure 9 in that chapter to refresh your memory.) But nowhere in our previous analysis did we consider any change in the price level. Thus, the rise in GDP to \$13.5 trillion in the previous chapter makes sense *only if the price level does not change*. Here, in Figure 9, this *would* be a movement rightward, from point *E* to point *J*. However, *point J does not describe the economy's short-run equilibrium*. Why not? Because it ignores two facts that you've learned about in this chapter: The rise in output will change the price level, and the change in the price level will, in turn, affect equilibrium GDP.

To see this more clearly, let's first suppose that the price level did *not* rise when output increased, so that the economy actually *did* arrive at point *J* after the *AD* shift. Would we stay there? Absolutely not. Point *J* lies below the *AS* curve, telling us that when GDP is \$13.5 trillion, the price level consistent with firms' unit costs, and average markup is 130, not 100. Firms would soon raise prices, and this would cause a movement upward along AD_2 . The price level would keep rising, and output would keep falling, until we reached point *H*. At that point—with output at \$12 trillion—we would be on both the *AS* and *AD* curves, so there would be no reason for a further rise in the price level and no reason for a further fall in output.

THE EFFECT OF A DEMAND SHOCK

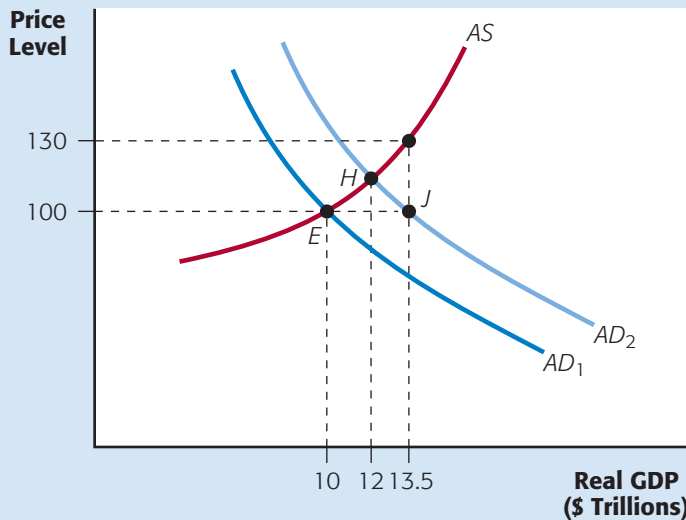
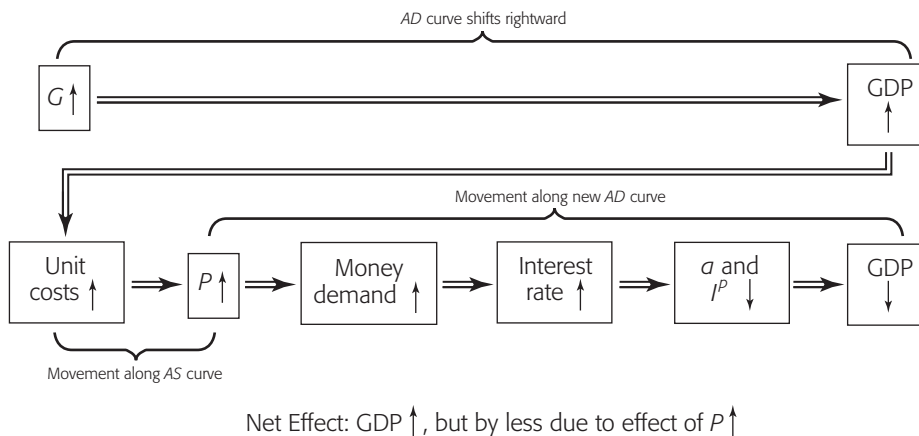


FIGURE 9

Starting at point *E*, an increase in government purchases would shift the *AD* curve rightward to *AD*₂. Point *J* illustrates where the economy would move if the price level remained constant. But as output increases, the price level rises. Thus, the economy moves along the *AS* curve from point *E* to point *H*.

However, the process we’ve just described is not entirely realistic. It assumes that when government purchases rise, *first* output increases (the move to point *J*), and *then* the price level rises (the move to point *H*). In reality, output and the price level tend to rise *together*. Thus, the economy would likely *slide* along the *AS* curve from point *E* to point *H*. As we move along the *AS* curve, output rises, increasing unit costs and the price level. At the same time, the rise in the price level *reduces* equilibrium GDP—the level of output toward which the economy is heading on the *AD* curve—from point *J* to point *H*.

We can summarize the impact of a rise in government purchases this way:



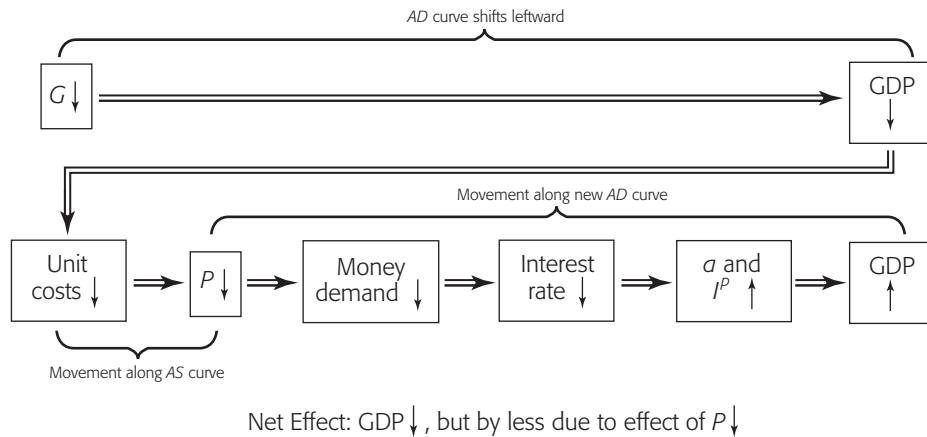
Let’s step back a minute and get some perspective about this example of fiscal policy. This is the third time in this text that we’ve considered fiscal policy in the short run. Each time, the discussion became more realistic, and we’ve seen that the effect of fiscal policy becomes weaker. In our first analysis, we ignored any increase in the interest rate, and found that a rise in government purchases increased

equilibrium GDP according to the simple multiplier formula $1/(1 - MPC)$. In our second analysis, in the chapter before this one, you learned that a rise in government purchases increases the interest rate, crowding out some interest-sensitive spending, thus making the rise in GDP smaller than it would otherwise be. The multiplier, therefore, was smaller than $1/(1 - MPC)$. Now you've learned that the rise in government purchases *also* increases the price level. This leads to a *further* rise in the interest rate, crowding out still *more* interest-sensitive spending, and making the rise in GDP smaller still. The size of the multiplier has been reduced yet again. (In our example, a \$2 trillion increase in government purchases increases equilibrium GDP by \$2.5 trillion, so the multiplier would be $\$2.5 \text{ trillion}/\$2 \text{ trillion} = 1.25$.) However, as you can see in Figure 9, a rise in government purchases—even when we include the rise in the price level—still raises GDP in the short run.

We can summarize the impact of price-level changes this way:

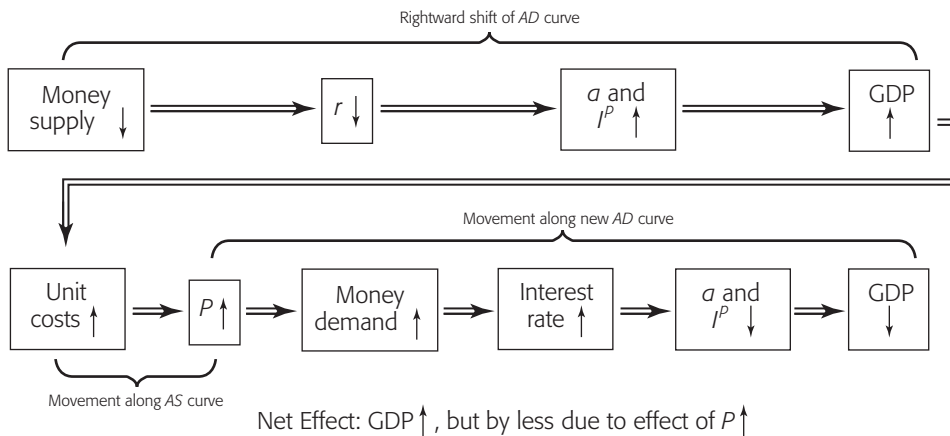
When government purchases increase, the horizontal shift of the AD curve measures how much real GDP would increase if the price level remained constant. But because the price level does rise, real GDP rises by less than the horizontal shift in the AD curve.

Now let's switch gears into reverse: How would we illustrate the effects of a *decrease* in government purchases? In this case, the AD curve would shift *leftward*, causing the following to happen:



As you can see, the same sequence of events occurs in the same order, but each variable moves in the opposite direction. A decrease in government purchases decreases equilibrium GDP, but the multiplier effect is smaller because the price level falls.

An Increase in the Money Supply. Although monetary policy stimulates the economy through a different channel than fiscal policy, once we arrive at the AD and AS diagram, the two look very much alike. For example, an increase in the money supply, which reduces the interest rate, will stimulate interest-sensitive consumption and investment spending. Real GDP then increases, and the AD curve shifts rightward, just as in Figure 9. Once output begins to rise, we have the same sequence of events as in fiscal policy: The price level rises, so the increase in GDP will be smaller. We can represent the situation as follows:



Other Demand Shocks. On your own, try going through examples of different demand shocks (see the list in Figures 4 (b) and (c)) and explain the sequence of events in each case that causes output and the price level to change. This will help you verify the following general conclusion about demand shocks:

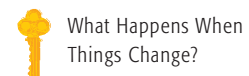
A positive demand shock—one that shifts the AD curve rightward—increases both real GDP and the price level in the short run. A negative demand shock—one that shifts the AD curve leftward—decreases both real GDP and the price level in the short run.

An Example: The Great Depression. As mentioned at the beginning of the chapter, the U.S. economy collapsed far more seriously during the period 1929 through 1933—the onset of the Great Depression—than it did at any other time in the country’s history. Because the price level fell during this time, we know that the contraction was caused by an adverse demand shock. An adverse supply shock would have caused the price level to rise as GDP fell.

What do we know about the demand shocks that caused the depression? This question has been debated by economists almost continuously over the past 70 years. The candidates are numerous, and it appears that a combination of events was responsible. The 1920s were a period of optimism—with high levels of investment by businesses and spending by families on houses and cars. The stock market soared. In the fall of 1929, the bubble of optimism burst. The stock market crashed, and investment and consumption spending plummeted. Similar events occurred in other countries, and the demand for products exported by the United States fell. The Fed—then only 16 years old—reacted by cutting the money supply sharply, which added an adverse monetary shock to all of the cutbacks in spending. Each of these events contributed to a leftward shift of the AD curve, causing both output and the price level to fall.

DEMAND SHOCKS: ADJUSTING TO THE LONG RUN

In Figure 9, point *H* shows the new equilibrium after a positive demand shock *in the short run*—a year or so after the shock. But point *H* is not necessarily where the economy will end up in the long run. For example, suppose full-employment output is \$10 trillion, and point *H*—representing an output of \$12 trillion—is above full-employment output. Then—with employment unusually high and



unemployment unusually low—business firms will have to compete to hire scarce workers, driving up the wage rate. It might take a year or more for the wage rate to rise significantly—recall our earlier list of reasons that wages adjust only slowly. But when we extend our horizon to several years or more, we must recognize that if output is above its potential, the wage rate will rise. Since the *AS* curve is drawn for a *given wage*, a rise in the wage rate will *shift* the curve upward, changing our equilibrium.

Alternatively, we could imagine a situation in which short-run equilibrium GDP was *below* its potential. In this case, with abnormally high unemployment, workers would compete to get scarce jobs, and eventually the wage rate would fall. Then the *AS* curve would shift downward, once again changing our equilibrium GDP.

In the short run, we treat the wage rate as given. But in the long run, the wage rate can change. When output is above full employment, the wage rate will rise, shifting the AS curve upward. When output is below full employment, the wage rate will fall, shifting the AS curve downward.

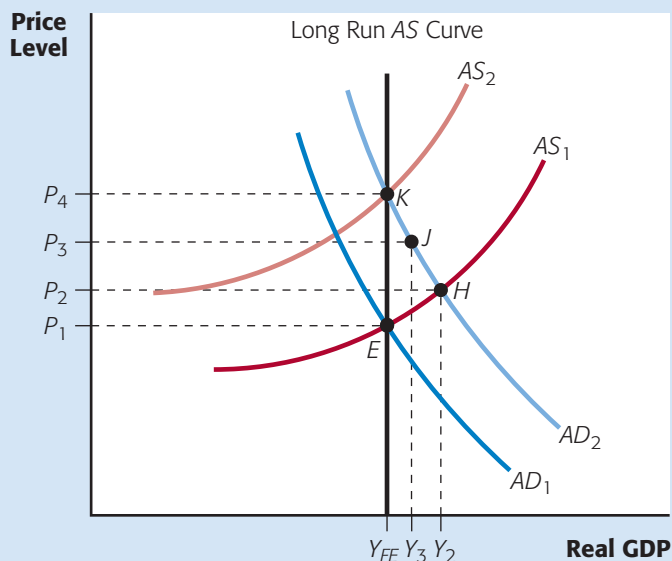
Now we are ready to explore what happens over the long run in the aftermath of a demand shock. Figure 10 shows an economy in equilibrium at point *E*. We assume that the initial equilibrium is at full-employment output (Y_{FE}), since—as you are about to see—this is where the economy always ends up after the long-run adjustment process is complete. To make our results as general as possible, we'll use symbols, rather than numbers, to represent output and price levels.

Now suppose the *AD* curve shifts rightward due to, say, an increase in government purchases. In the short run, the equilibrium moves to point *H*, with a higher price level (P_2) and a higher level of output (Y_2). Point *H* tells us where the economy will be about a year after the increase in government purchases, before the

FIGURE 10

Beginning at point *E*, a positive demand shock would shift the aggregate demand curve to AD_2 , raising both output and the price level. At point *H*, output is above the full-employment level, Y_{FE} . Firms will compete to hire scarce workers, thereby driving up the wage rate. The higher wage rate will shift the *AS* curve to AS_2 . Only when the economy returns to full-employment output at point *K* will there be no further shifts in *AS*.

THE LONG-RUN ADJUSTMENT PROCESS



wage rate has a chance to adjust. (Remember, along any given AS curve, the wage rate is assumed to be constant.)

But now let's extend our analysis beyond a year. Notice that Y_2 is greater than Y_{FE} . The wage will begin to rise, raising unit costs at any given output level and causing firms to raise prices. In the figure, the AS curve would begin shifting upward. Point J shows where the shifting aggregate supply curve might be two years after the shock, after the long-run adjustment process has begun. (You might want to pencil this intermediate AS curve into the figure, so that it intersects AD_2 at point J .) At this point, output would be at Y_3 , and the rise in the price level has moved us along the new aggregate demand curve, AD_2 .

Now, is point J our final, long-run equilibrium? No, it cannot be. At Y_3 , output is *still* greater than Y_{FE} , so the wage rate will continue to rise, and the AS curve will continue to shift upward. At point J , the long-run adjustment process is not yet complete. When will the process end? Only when the wage rate stops rising—that is, only when output has returned to Y_{FE} . This occurs when the AS curve has shifted all the way to AS_2 , moving the economy to point K —our new, long-run equilibrium.

As you can see, the increase in government purchases has no effect on equilibrium GDP in the long run: The economy returns to full employment, which is just where it started. This is why the long-run adjustment process is often called the economy's **self-correcting mechanism**. And this mechanism applies to any demand shock, not just an increase in government purchases:

Self-correcting mechanism The adjustment process through which price and wage changes return the economy to full-employment output in the long run.

If a demand shock pulls the economy away from full employment, changes in the wage rate and the price level will eventually cause the economy to correct itself and return to full-employment output.

For a positive demand shock that shifts the AD curve rightward, the self-correcting mechanism works like this:

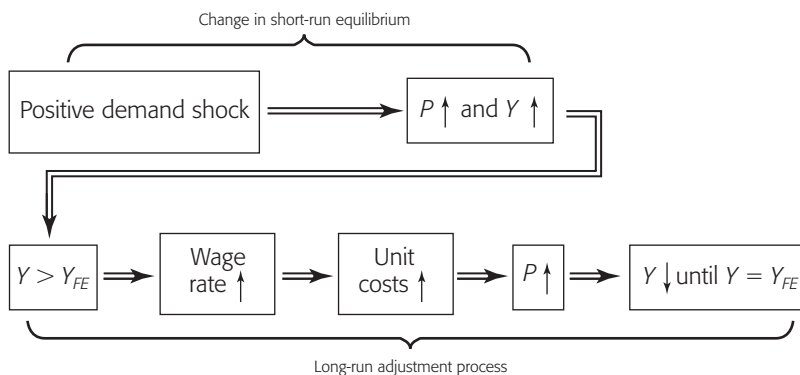


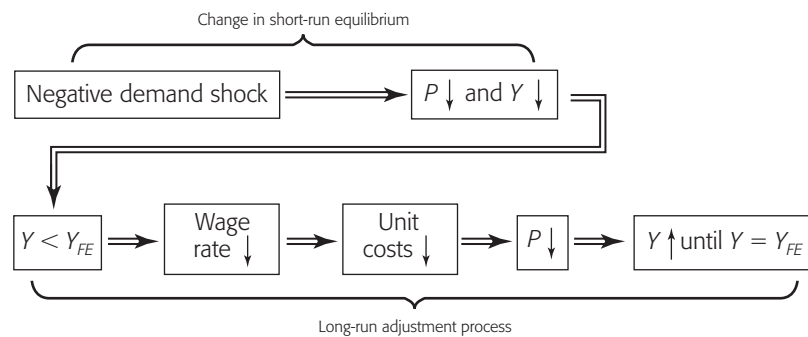
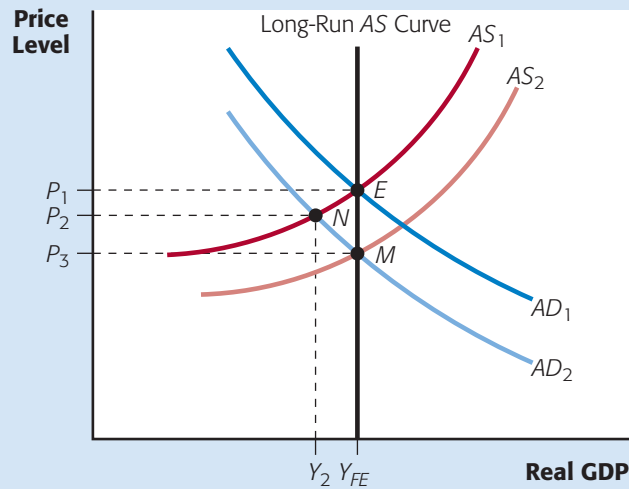
Figure 11 illustrates the case of a negative demand shock, in which the AD curve shifts leftward. In this case, the short-run equilibrium GDP is *below* Y_{FE} —at point N . Over the long run, however, the unusually high unemployment drives the wage rate down, shifting the AS curve down as well. The price level decreases, causing equilibrium GDP to rise along the AD_2 curve. The process comes to a halt only when output returns to Y_{FE} . Thus, in the long run, the economy moves from point E to point M , and the negative demand shock causes no change in equilibrium GDP.

The complete sequence of events after a negative demand shock looks like this:

FIGURE 11

Starting from point E , a negative demand shock shifts the AD curve to AD_2 , lowering GDP and the price level. At point N , output is below the full-employment level. With unemployed labor available, wages will fall, enabling firms to lower their prices. The AS curve shifts downward until full employment is regained at point M , with a lower price level.

LONG-RUN ADJUSTMENT AFTER A NEGATIVE DEMAND SHOCK



Pulling all of our observations together, we can summarize the economy's self-correcting mechanism as follows:

Whenever a demand shock pulls the economy away from full employment, the self-correcting mechanism will eventually bring it back. When output exceeds its full-employment level, wages will eventually rise, causing a rise in the price level and a drop in GDP until full employment is restored. When output is less than its full-employment level, wages will eventually fall, causing a drop in the price level and a rise in GDP until full employment is restored.

THE LONG-RUN AGGREGATE SUPPLY CURVE

The self-correcting mechanism provides an important link between the economy's long-run and short-run behaviors. It helps us understand why deviations from full employment don't last forever. Often, however, we are primarily interested in the long-run effects of a demand shock. In these cases, we may want to skip over the self-correcting mechanism and go straight to its end result. A new version of the AS curve helps us do this.

Look again at Figure 10, which illustrates the impact of a positive demand shock. The economy begins at full employment at point *E*, then moves to point *H* in the short run (before the wage rate rises), and then goes to point *K* in the long run (after the rise in wages). If we skip over the short-run equilibrium, we find that the positive demand shock has moved the economy from *E* to *K*, which is vertically above *E*. That is, in the long run, the price level rises, but output remains unchanged.

Now look at the vertical line in Figure 10, which shows another way of illustrating this long-run result. In the figure, the vertical line is the economy's **long-run aggregate supply curve**. It summarizes all possible output and price-level combinations at which the economy could end up in the long run. It is vertical because, in the long run, GDP will be the same—full-employment output—*regardless* of the position of the *AD* curve. The price level, however, will depend on the position of the *AD* curve. In the long run, a positive demand shock shifts the *AD* curve rightward, moving the economy from *E* to *K*: a higher price level, but the same level of output. Similarly, in Figure 11, a negative demand shock—which shifts the *AD* curve leftward—moves the economy from *E* to *M* in the long run: a lower price level with the same level of output.³

The long-run aggregate supply curve in Figures 10 and 11 tell us something very important about the economy: In the long run, after the self-correcting mechanism has done its job, *the economy behaves as the classical model predicts*. In particular, the classical model tells us that demand shocks cannot change equilibrium GDP in the long run. The figure brings us to the same conclusion: While demand shocks shift the *AD* curve, this only moves the economy up or down along a vertical long-run *AS* curve, leaving output unchanged.

The long-run aggregate supply curve also illustrates another classical conclusion. In the classical model, an increase in government purchases causes *complete crowding out*—the rise in government purchases is precisely matched by a drop in consumption and investment spending, leaving total output and total spending unchanged. In Figure 10, the same result holds in the long run. How do we know? The figures tell us that, in the long run, the rise in government purchases causes no change in GDP. But if GDP is the same, and government purchases are higher, then the other components of GDP—consumption and investment—must decrease by the amount that government purchases increased.

The self-correcting mechanism shows us that, in the long run, the economy will eventually behave as the classical model predicts.

But notice the word *eventually* in the previous statement. It can take several years before the economy returns to full employment after a demand shock. This is why governments around the world are reluctant to rely on the self-correcting mechanism alone to keep the economy on track. Instead, they often use fiscal and monetary policies in an attempt to return the economy to full employment more quickly. We'll explore fiscal and monetary policies in more detail in the next two chapters.

SUPPLY SHOCKS

In recent decades, supply shocks have been important sources of economic fluctuations. The most dramatic supply shocks have resulted from sudden changes in

Long-run aggregate supply curve

A vertical line indicating all possible output and price-level combinations at which the economy could end up in the long run.



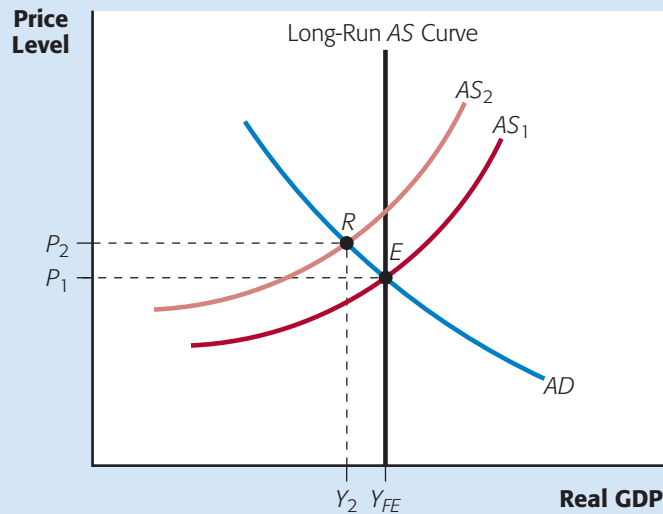
What Happens When Things Change?

³ Of course, full-employment output can increase from year to year, as you learned in the chapter on economic growth. When the economy is growing, the long-run *AS* curve will shift rightward. In that case, the level of output at which the economy will eventually settle increases from year to year.

FIGURE 12

An adverse supply shock would shift the AS curve upward from AS_1 to AS_2 . In the short-run equilibrium at point R , the price level is higher and output is below Y_{FE} . Eventually, wages will fall, causing unit costs to fall, and the AS curve will shift back to its original position. A positive supply shock would have just the opposite effect.

THE EFFECT OF A SUPPLY SHOCK



world oil prices. As you are about to see, supply shocks affect the economy differently from demand shocks.

Short-Run Effects of Supply Shocks. Figure 12 shows an example of a supply shock: an increase in world oil prices that shifts the aggregate supply curve upward, from AS_1 to AS_2 . As rising oil prices increase unit costs, firms will begin raising prices, and the price level will increase. The rise in the price level decreases equilibrium GDP along the AD curve. In the short run, the price level will continue to rise, and the economy will continue to slide upward along its AD curve, until we reach the AS_2 curve at point R . At this point, the price level is consistent with firms' unit costs and average markup (we are on the AS curve), and total output is equal to total spending (we are on the AD curve). As you can see, the short-run impact of higher oil prices is a rise in the price level and a fall in output. We call this a *negative* supply shock, because of the negative effect on output.

In the short run, a negative supply shock shifts the AS curve upward, decreasing output and increasing the price level.

Notice the sharp contrast between the effects of negative supply shocks and negative demand shocks in the short run. After a negative demand shock (see, for example, Figure 11), both output and the price level fall. After a negative supply shock, however, output falls, but the price level rises. Economists and journalists have coined the term **stagflation** to describe a *stagnating* economy experiencing *inflation*.

A negative supply shock causes stagflation in the short run.

Stagflation The combination of falling output and rising prices.

Stagflation caused by increases in oil prices is not just a theoretical possibility. Three of our recessions in the last quarter century—in 1973–74, 1980, and 1990–91—followed increases in world oil prices. And each of these three recessions also saw jumps in the price level.

A *positive supply shock* would increase output by shifting the AS curve downward. As you can see if you draw such a shift on your own,

a positive supply shock shifts the AS curve downward, increasing output and decreasing the price level.

Examples of positive supply shocks include unusually good weather, a drop in oil prices, and a technological change that lowers unit costs. In addition, a positive supply shock can sometimes be caused by government policy. A few chapters ago, we discussed how the government could use tax incentives and other policies to increase the rate of economic growth. These policies work by shifting the AS curve downward, thus increasing output while tending to decrease the price level.

Another type of policy tries to deal directly with negative supply shocks. For example, after the oil price shocks of the 1970s, the federal government built a strategic reserve of oil in huge underground storage areas. The idea was to release oil from the reserve if another oil price shock hit, in order to stabilize the price. The reserve was used in this way, but not enough to make much difference in the world price of oil.

Long-Run Effects of Supply Shocks. What about the effects of supply shocks in the long run? In some cases, we need not concern ourselves with this question, because some supply shocks are temporary. For example, except in unusual cases, periods of rising oil prices are followed by periods of falling oil prices. Similarly, supply shocks caused by unusually good or bad weather, or by natural disasters, are always short lived. A temporary supply shock causes only a temporary shift in the AS curve; over the long run, the curve simply returns to its initial position, and the economy returns to full employment. In Figure 12, the AS curve would shift back from AS_2 to AS_1 , the price level would fall, and the economy would move from point R back to point E.

In other cases, however, a supply shock can last for an extended period. One example was the rise in oil prices during the 1970s, which persisted for several years. In cases like this, is there a self-correcting mechanism that brings the economy back to full employment after a long-lasting supply shock? Indeed, there is, and it is the same mechanism that brings the economy back to full employment after a demand shock.

Look again at Figure 12. At point R, output is below full-employment output. In the long run, as workers compete for scarce jobs, the wage rate will decline. This will cause the AS curve to shift *downward*. The wage will continue to fall until the economy returns to full employment—that is, until we are back at point E.

In the long run, the economy self-corrects after a supply shock, just as it does after a demand shock. When output differs from its full-employment level, the wage rate changes, and the AS curve shifts until full employment is restored.

SOME IMPORTANT PROVISOS ABOUT THE AS CURVE

The upward-sloping aggregate supply curve we've presented in this chapter gives a realistic picture of how the economy behaves after a demand shock. In the short run, positive demand shocks that increase output also raise the price level. Negative demand shocks that decrease output generally put downward pressure on prices.

However, the story we have told about what happens as we move along the AS curve is somewhat incomplete.

First, we made the assumption that prices are completely flexible—that they can change freely over short periods of time. In fact, however, some prices take time to adjust, just as wages take time to adjust. Firms print catalogs containing prices that are good for, say, six months. The public utility commission in your state may set the prices of electricity, gas, water, and basic telephone service in advance for a year or more.

Second, we assumed that wages are completely *inflexible* in the short run. But in *some* industries, wages respond quickly. For example, in the construction industry, contractors hire workers for projects lasting a few months. When they can't find the workers they want, they immediately offer higher wages—they don't wait a year.

Third, there is more to the process of recovering from a shock than the adjustment of prices and wages. During a recession, many workers lose their jobs at the same time. It takes time for those workers to become re-established in new jobs. As time passes, and job losers become job finders, the economy tends to recover. This process, in addition to the changes in wages and prices we've discussed, is part of the long-run adjustment process and helps to bring the economy back to full employment after a shock.

Using the THEORY



THE RECESSION AND RECOVERY OF 1990–92

The aggregate demand and aggregate supply curves are not just graphs; they are tools that help us understand important economic events. In this section, we'll look at how we can use these tools to understand our most recent recession.

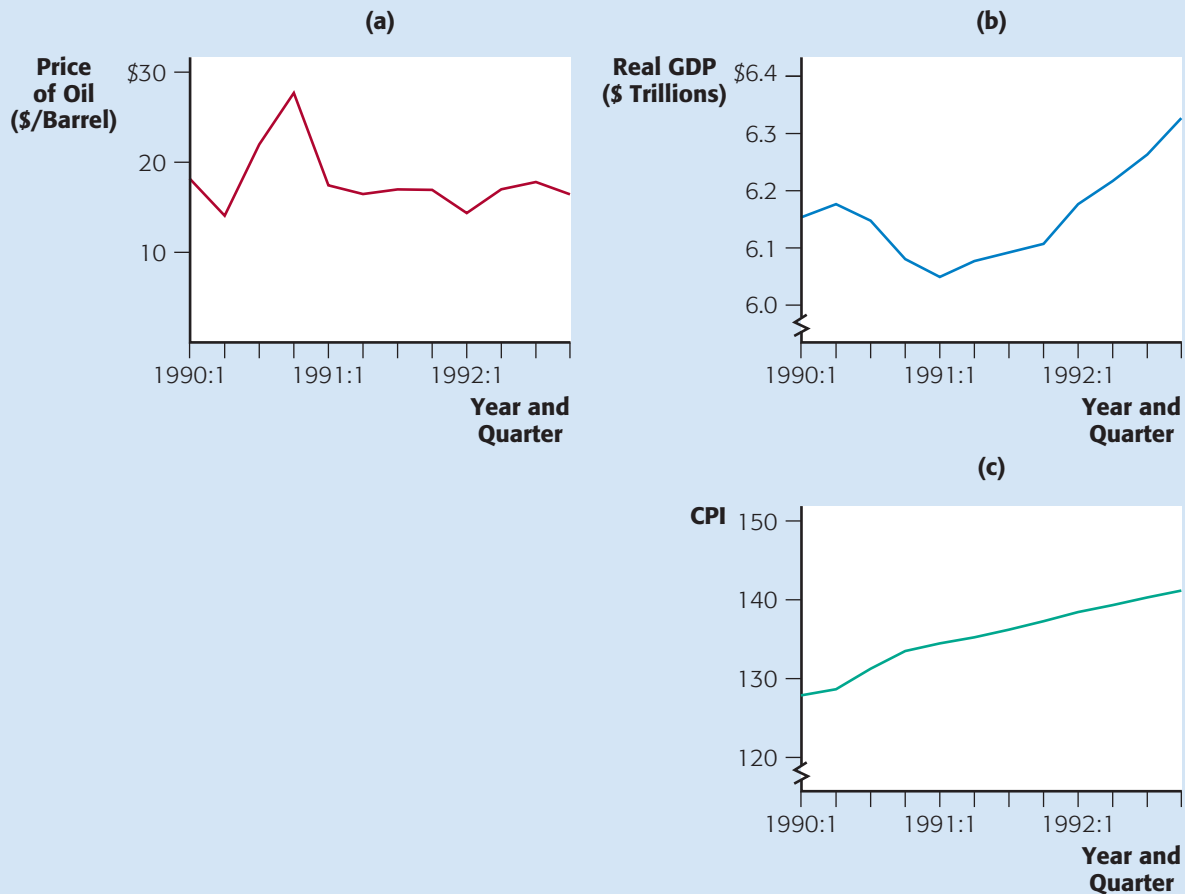
Our story begins in mid-1990, when Iraq invaded Kuwait, a major oil producer. During this conflict, Kuwait's oil was taken off the world market, and so was Iraq's. The reduction in oil supplies resulted in an immediate and substantial increase in the price of oil, a key input to many industries. Panel (a) of Figure 13 shows that the price of oil rose from a low of \$14 to a high of \$27 per barrel in 1990.

Figure 14 shows our AS – AD analysis of the shock. Initially, the economy is on both AD_1 and AS_1 , with equilibrium at point E , and output at its full-employment level. Then, the oil price shock shifts the AS curve upward, to AS_2 . As the short-run equilibrium moves to point R , real GDP falls and the price level rises. Going back to Figure 13, we see that this is indeed what happened. Panel (b) shows that real GDP did fall in the period after the shock, from \$6.7 trillion in mid-1990 to about \$6.6 trillion in early 1991. Although the fall was not large, the economy was well below potential by 1992, because potential continued to grow. In panel (c), you can see that the Consumer Price Index rose especially rapidly during this period. Late 1990 through early 1991 was clearly a period of stagflation.

Now let's return to our AS – AD analysis in Figure 14. At point R , output is below its full-employment level. If the price of oil had remained high, our theory tells us, the self-correcting mechanism would have begun to work: Falling wages would have decreased unit costs. However, the self-correcting mechanism wasn't needed in this case: As you can see in Figure 13, the oil price shock was temporary. Oil prices fell back down in early 1991, shifting the AS curve back to AS_1 without the self-correcting mechanism. In panel (b) of Figure 13, you can see that real GDP began to recover in early 1991, and continued moving back to its full-employment level in the succeeding years.

THE PRICE OF OIL, REAL GDP, AND THE PRICE LEVEL, 1990–92

FIGURE 13



In mid-1990, Kuwaiti and Iraqi oil was taken off the world market, resulting in a substantial increase in the world price of oil, as shown in panel (a). U.S. GDP fell, and the consumer price index rose. When oil prices fell in 1991, GDP recovered.

But something looks fishy here. In our $AS-AD$ analysis, the price level should rise when the negative supply shock hits and then gradually *fall* back to its original level when the shock proves temporary. But panel (c) of Figure 13 shows that this prediction was not borne out by the experience of 1990–92. While the price level did rise rapidly in the year after the shock, it *continued to rise* in the next two years as the economy self-corrected. Have we missed something?

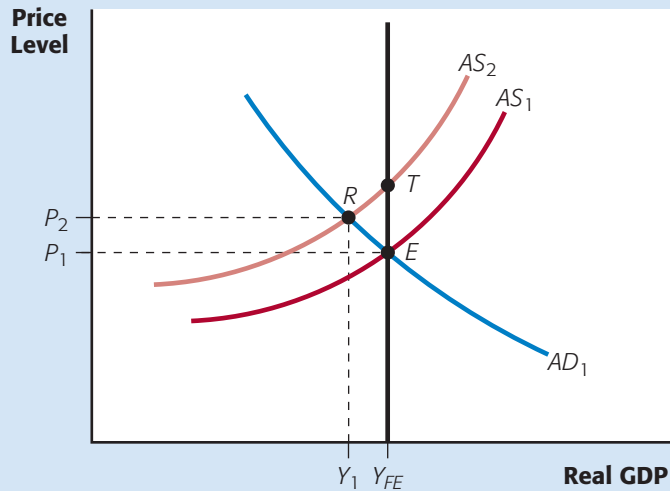
Yes, we have. In our analysis of demand and supply shocks in this chapter, we've been focusing on only one change at a time. And here, too, we've been looking at the events of 1990–92 by considering *only* the shift of the AS curve. In particular, as the AS curve shifts upward and then downward, we've assumed that the AD curve stays put.

But that is not what happened in the early 1990s. Instead, in the period after the shock, the Fed increased the money supply, shifting the AD curve rightward. Thus, instead of moving from point R back to E , the economy moved from R to T . (You can draw in the new AD curve to help you see the move.) Output rose, but the price level rose as well.

FIGURE 14

AN AD-AS ANALYSIS OF THE OIL PRICE SHOCK

Beginning at point E , the increase in the world price of oil shifted the AS curve from AS_1 to AS_2 . Output fell and the price level rose. When oil prices fell in 1991, the AS curve shifted back to AS_1 . Because the Fed simultaneously increased the money supply, the AD curve shifted rightward (not shown). By 1992, output was back to Y_{FE} but with a higher price level at point T .



Why did the Fed increase the money supply, rather than hold it constant and let the economy adjust back to point E ? This is a question about monetary policy and the Fed's motives in conducting it—a subject we will consider in detail in the next chapter.

SUMMARY

The model of aggregate supply and demand explains how the price level and output are determined in the short run—a period of a year or so following a change in the economy—and how the economy adjusts over longer time periods as well.

The aggregate demand (AD) curve shows how changes in the price level affect equilibrium real GDP. A change in the price level shifts the money demand curve and alters the interest rate in the money market. The change in the interest rate, in turn, affects interest-sensitive forms of spending, shifts the aggregate expenditure curve, triggers the multiplier process, and leads to a new level of equilibrium real GDP. A lower price level means a higher equilibrium real GDP, and a higher price level means lower GDP. The downward-sloping AD curve is drawn for given values of government spending, taxes, autonomous consumption spending, investment spending, the money supply, and the public's preferences for holding money and bonds. Changes in any of those factors will cause the AD curve to shift.

The aggregate supply (AS) curve summarizes the way changes in output affect the price level. To draw the AS curve, we assume that firms set the price of individual products as a markup over their costs per unit, and that the economy's average markup is determined by competitive conditions. We also assume that the nominal wage rate is fixed in the short run. As we move upward along the AS curve, a rise in real GDP, by raising unit costs, causes the price level to increase. When any-

thing other than a change in real GDP causes the price level to change, the entire AS curve shifts.

AD and AS together determine real GDP and the price level. The economy must be on the AD curve, or real GDP would not be at its equilibrium level. It must be on the AS curve or firms would not be charging prices dictated by their unit costs and markups. Both conditions are satisfied at the intersection of the two curves.

The AD/AS equilibrium can be disturbed by a demand shock. An increase in government purchases, for example, shifts the AD curve rightward. As a result, the price level rises, and so does real GDP. In the long run, if GDP is above potential, wages will rise. This causes unit costs to rise and shifts the AS curve upward. Eventually, GDP will return to potential and the only long-run result of the demand shock is a higher price level. This implies that the economy's long-run aggregate supply curve is vertical at potential output.

The short-run AD/AS equilibrium can also be disturbed by a supply shock, such as an increase in world oil prices. With unit costs higher at each level of output, the AS curve shifts upward, decreasing real GDP and increasing the price level. Eventually, the shock will be self-correcting: With output below potential, the wage rate will fall, unit costs will decrease, and the AS curve will shift back downward until full employment is restored.

KEY TERMS

aggregate demand (<i>AD</i>) curve	short-run macroeconomic equilibrium	supply shock	long-run aggregate supply curve
aggregate supply (<i>AS</i>) curve	demand shock	self-correcting mechanism	stagflation

REVIEW QUESTIONS

1. What causal relationship does the aggregate demand curve describe? Why does the *AD* curve slope downward? What does each point on the *AD* curve represent?
2. “Only spending shocks can shift the aggregate demand curve.” True or false? Explain.
3. List three reasons why a change in output affects unit costs and subsequently the price level.
4. What causal relationship does the aggregate supply curve describe? Why does the *AS* curve slope upward?
5. Why does equilibrium occur only where the *AD* and *AS* curves intersect?
6. What is the economy’s *self-correcting mechanism*, and how does it work?
7. What is the long-run aggregate supply curve? Why is it vertical?
8. Does the vertical shape of the long-run aggregate supply curve support the predictions of the classical model with regard to the effectiveness of fiscal policy and crowding out? Explain.
9. How does an economy recover from a negative supply shock?

PROBLEMS AND EXERCISES

1. With a three-panel diagram—one panel showing the money market, one showing the aggregate expenditure diagram, and one showing the *AD* curve—show how a *decrease* in the money supply shifts the *AD* curve leftward.
2. Using a diagram showing the aggregate expenditure line, the money market, and the *AD* curve, describe how an increase in taxes affects the interest rate, real aggregate expenditure, and the aggregate demand curve. (Assume that the price level does not change.) What other changes would result in these same effects?
3. Suppose firms become pessimistic about the future and consequently investment spending falls. With an *AD* and *AS* graph, describe the short-run effects on GDP and the price level. If the price level were constant, how would your answer change?
4. With an *AD* and *AS* diagram, explain the short-run effect of a decrease in the money supply on GDP and the price level. What is the effect in the long run? Assume the economy begins at full employment.
5. A new government policy successfully lowers firms’ unit costs. What are the short-run and the long-run effects of such a policy? (Assume that full-employment output does not change.)

CHALLENGE QUESTIONS

1. Suppose that wages are slow to adjust downward but rapidly adjust upward. What would the *AS* curve look like? How would this affect the economy’s adjustment to spending shocks (compared to the analysis given in the chapter)?
2. In recent years, because of technological change, the *AS* curve has been shifting downward, but the price level has not fallen. Why? (*Hint*: What has the Fed been doing?)

EXPERIENTIAL EXERCISE

1. Net exports are an important influence on aggregate demand. Find a story in today's *Wall Street Journal* that describes an event that will affect U.S. imports or exports. A good place to look is in the "International" page in the first section of the *Journal*. Analyze the story you have chosen, and illustrate the event using the aggregate expenditure model and the aggregate demand and supply model.

INFLATION AND MONETARY POLICY

In the late 1970s, the annual inflation rate in the United States reached 13 percent. At the time, polls showed that the public considered inflation the most serious economic problem facing the country. In the nine years following 1991, however, the annual inflation rate never exceeded 3.5 percent, and the problem receded as a matter of public concern. Keeping the inflation rate low has been one of the solid victories of national economic policy.

How did the Fed achieve this victory? Why was it less successful in earlier periods? Are there costs, as well as benefits, to a lower inflation rate? And how should the Fed respond to economic disturbances as it faces the future?

In this chapter, we'll be addressing these and other questions as we take a closer look at the Fed's conduct of monetary policy. Our earlier discussions of monetary policy were somewhat limited, because we lacked the tools—aggregate demand and aggregate supply—to explain changes in the price level. In this chapter, we'll explore monetary policy more fully, making extensive use of the *AD* and *AS* curves.

THE OBJECTIVES OF MONETARY POLICY

The Fed's objectives have changed over the years. When the Fed was first established in 1913, its chief responsibility was to ensure the stability of the banking system. By acting as a *lender of last resort*—injecting reserves into the banking system in times of crisis—the Fed was supposed to alleviate financial panics.

By the 1950s, the stability of the banking system was no longer a major concern, largely because the United States had not had a banking panic in decades. (Deposit insurance programs had effectively eliminated panics.) Accordingly, the Fed's objective in the 1950s and 1960s changed to keeping the interest rate low and stable. In the 1970s, the Fed's objectives shifted once again. As stated in the Federal Reserve Banking Act of 1978, which is still in force, the Fed is now responsible for achieving a low, stable rate of inflation, and full employment of the labor force. Let's consider each of these goals in turn.

CHAPTER OUTLINE

The Objectives of Monetary Policy

Low, Stable Inflation
Full Employment

The Fed's Performance

Federal Reserve Policy: Theory and Practice

Responding to Changes in Money Demand
Responding to Spending Shocks
Responding to Supply Shocks

Expectations and Ongoing Inflation

How Ongoing Inflation Arises
Ongoing Inflation and the Phillips Curve
Why the Fed Allows Ongoing Inflation

Using the Theory: Conducting Monetary Policy in the Real World

Information About the Money Demand Curve
Information About the Sensitivity of Spending to the Interest Rate
Uncertain and Changing Time Lags
The Natural Rate of Unemployment

LOW, STABLE INFLATION

Why is a low rate of inflation important? Several chapters ago, we reviewed the social costs of inflation. When the inflation rate is high, society uses up resources coping with it—resources that could have been used to produce goods and services. Among these resources are the labor needed to update prices at stores and factories, as well as the additional time spent by households and businesses to manage their wealth and protect it from a loss of purchasing power.

In addition to keeping the inflation rate low, the Fed tries to keep it *stable* from year to year. For example, the Fed would prefer a steady yearly inflation rate of 3 percent to an inflation rate of 5 percent half the time, and 1 percent the other half, even though the average inflation rate would be 3 percent in both cases. The reason is that unstable inflation is difficult to predict accurately; it will often turn out higher or lower than people expected. As you learned several chapters ago, an inflation rate higher than expected redistributes real income from lenders to borrowers, while an inflation rate lower than expected has the opposite effect. Thus, unstable inflation adds to the risk of lending and borrowing, and interferes with long-run financial planning.

The Fed, as a public agency, chooses its policies with the costs of inflation in mind. And the Fed has another concern: Inflation is very unpopular with the public. Surveys show that most people associate high rates of inflation with a general breakdown of government and the economy.¹ A Fed chairman who delivers low rates of inflation is seen as popular and competent, while one who tolerates high inflation goes down in history as a failure.

FULL EMPLOYMENT

“Full employment” means that unemployment is at normal levels. But what, exactly, is a *normal* amount of employment?

Recall that there are different types of unemployment. Some of the unemployed in any given month will find jobs after only a short time of searching. This *frictional* unemployment is part of the normal working of the labor market, and is not a serious social problem. Other job seekers will spend many months or years out of work because they lack the skills that employers require, or because they lack information about available jobs. While this *structural* unemployment is a serious social problem, it is best solved with *microeconomic* policies, such as job-training programs or improved information flows.

Cyclical unemployment, by contrast, is a *macroeconomic* problem. It occurs during a recession, in which millions of workers lose their jobs and remain unemployed as they seek new ones. This is why macroeconomists use the term “full employment” to mean *the absence of cyclical employment*. When the economy achieves full employment according to this definition, macroeconomic policy has done all that it can do.

The Fed is concerned about cyclical unemployment for two reasons. First is its *opportunity cost*: the output that the unemployed could have produced if they were working. Part of this opportunity cost is paid by the unemployed themselves, in the form of lost earnings, and part is paid by people who remain employed, but pay higher taxes to provide unemployment benefits to job losers. By maintaining full employment, the Fed can help society avoid this cost.

Second, cyclical unemployment represents a social failure. In a recession, people who have the right skills and who could be working actually *lose* their jobs. Excess

¹ Robert J. Shiller, “Public Resistance to Inflation: A Puzzle,” *Brookings Papers on Economic Activity*, 1997.

unemployment lingers for several years after a recession strikes. Thus, cyclical unemployment caused by a recession is a partial breakdown of the system. The economy is not doing what it should do: provide a job for anyone who wants to work and who has the needed skills.

But why should the Fed try to eliminate only *cyclical* unemployment? Why not go further—pushing output above its full-employment level? After all, at higher levels of output, business firms would be more willing to hire *any* available workers. The frictionally unemployed would find jobs more easily, and some of the structurally unemployed would be hired as well. If unemployment is a bad thing, shouldn't the Fed aim for the lowest possible unemployment rate possible?

The answer is no. If the unemployment rate falls too low, GDP rises beyond its potential, full-employment level. As you learned in the last chapter, this causes the economy's self-correcting mechanism to kick in: The *AS* curve shifts upward, increasing the price level. Thus, unemployment that is too low compromises the Fed's other chief goal by creating inflation. And, as you will see later in the chapter, the Fed could not keep the economy operating above full employment for more than a short time anyway. In the long run, its attempts to push the economy too hard would only create more inflation and would not succeed in lowering unemployment.

The unemployment rate at which GDP is at its full-employment level—that is, with no cyclical unemployment—is sometimes called the **natural rate of unemployment**.

When the unemployment rate is below the natural rate, GDP is greater than potential output. The economy's self-correcting mechanism will then create inflation. When the unemployment rate is above the natural rate, GDP is below potential output. The self-correcting mechanism will then put downward pressure on the price level.

The word *natural* must be interpreted with care. The natural unemployment rate is not etched in stone, nor is it the outcome of purely natural forces that can't be influenced by public policy. But it is determined by rather slow-moving forces in the economy: how frequently workers move from job to job, how efficiently the unemployed can search for jobs and firms can search for new workers, and how well the skills of the unemployed match the skills needed by employers. The natural rate can also be influenced by government policies that provide incentives or disincentives for workers to find jobs quickly, or for employers to hire them. The natural rate can change when any of these underlying conditions change. Indeed, economists generally believe that over the past decade, the natural rate has decreased in the United States—from 6 or 6.5 percent in the mid-1980s to 4 or 4.5 percent today. Meanwhile, in many European countries, the natural rate of unemployment has increased in recent years—exceeding 10 percent in France and close to 20 percent in Spain. The causes of these changes in the natural rate, as well as the *extent* of the changes, are hotly debated by economists. But there is general agreement about the direction: down in the United States, up in Europe.

Why use the term *natural* for such a changeable feature of the economy? The term makes sense only from the perspective of *macroeconomic* policy. Simply put, there isn't much that macroeconomic policy can do about the natural rate. Stimulating the economy with fiscal or monetary policy may bring the *actual* unemployment rate down for a time, but it will not change the natural rate itself. And pushing unemployment below the natural rate would cause inflation. Thus, the natural rate of unemployment can be seen as a kind of goalpost for the Fed. The location of the goalpost may change over the years, but during any given year, it tells us where the Fed is aiming.

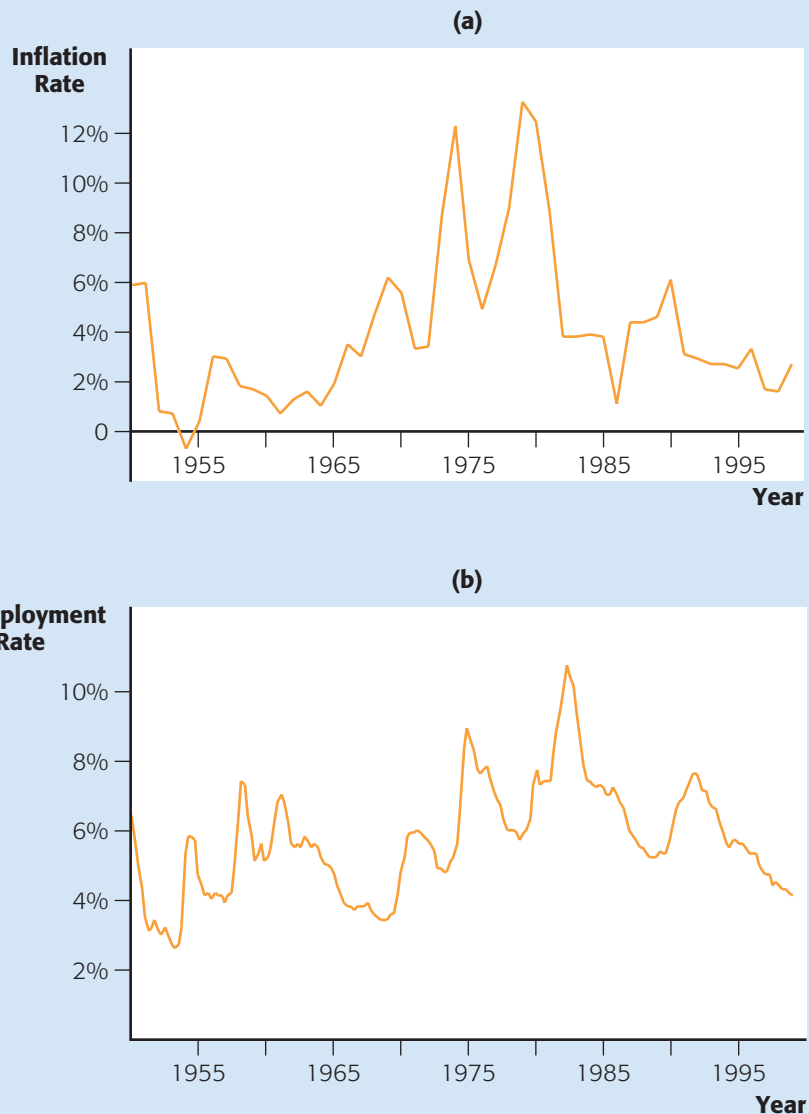
Natural rate of unemployment

The unemployment rate when there is no cyclical unemployment.

FIGURE 1

Panel (a) shows the annual inflation rate since 1950. The United States suffered periods of high inflation in the 1970s and early 1980s. Since then, the inflation rate has been much lower. Panel (b) shows the quarterly unemployment rate. Unemployment was particularly high during the early 1980s, and dropped dramatically during the 1990s.

THE FED'S PERFORMANCE SINCE 1950



THE FED'S PERFORMANCE

How well has the Fed achieved its goals? Panel (a) of Figure 1 shows the annual inflation rate since 1950, as measured by the Consumer Price Index. You can see that monetary policy permitted extended periods of high inflation in the 1970s and early 1980s. You can also see, as noted at the beginning of the chapter, that the Fed has achieved great success in controlling inflation since then. Indeed, in the 16 years from 1984 to 1999, the annual inflation rate exceeded 4.6 percent only once—in 1990, during the supply shock caused by higher oil prices. And in recent years, inflation at or below 3 percent has become the norm.

Panel (b) shows the quarterly rate of unemployment since 1950. Over the last decade or so, the Fed's performance on unemployment has been somewhat mixed. From 1984 to 1999, the unemployment rate was 7 percent or greater—significantly above its natural rate—slightly more than one-fourth of the time. The most recent period of high unemployment was during the recession of the early 1990s, when the unemployment rate stayed above 7.5 percent for half a year. But notice the remarkable improvement in unemployment from mid-1992 and after. Through 1997, the Fed kept the unemployment rate hovering very close to 5 percent, and after 1997, it slowly inched the unemployment rate down to 4 percent, which it finally reached in January 2000. And this reduction in unemployment was accomplished *without* heating up inflation.

As you can see, the Fed has had a good—and improving—record in recent years. The inflation rate has been kept low and relatively stable, and—especially in the last few years—unemployment has been near and even below most estimates of the natural rate. How has the Fed done it? Are there any general conclusions we can reach about how a central bank should operate to achieve the twin goals of full employment and a stable, low inflation rate? Indeed there are, as you'll see in the next section.

FEDERAL RESERVE POLICY: THEORY AND PRACTICE

So far in this text, we've assumed that the Fed's response to spending shocks is a **passive monetary policy**. That is, in the face of spending shocks, the Fed conducts neither open market purchases nor open market sales of bonds, and just keeps the money supply constant. While this was useful for understanding how different events can affect the economy, it is not a realistic description of the Fed's actions. In recent years, the Fed has tried to maintain a stable level of real GDP, rather than a stable money supply. Ideally, the Fed would like to keep the economy operating as close to its potential output as possible. If output falls below potential, there is painful and wasteful unemployment; if output rises above potential, there is a danger of inflation.

In order to keep real GDP as close as possible to its potential, the Fed must pursue an **active monetary policy**, in which it responds to events in the economy by *changing* the money supply. As you'll see, the required change in the money supply depends on what type of event the Fed is responding to.

In some cases, the proper response is easy to determine, because the same action that maintains full employment also helps maintain low inflation. But in other cases, the Fed must trade off one goal for another: Responses that maintain full employment will worsen inflation, and responses that alleviate inflation will create more unemployment.

We'll make a temporary simplifying assumption in this section: that the Fed's goal for the inflation rate is *zero*. In reality, the Fed's goal is *low*, but not zero, inflation. Later, we'll discuss why the Fed prefers a low inflation rate to a zero rate, and how this modifies our analysis.

RESPONDING TO CHANGES IN MONEY DEMAND

Potential disturbances to the economy sometimes arise from a shift in the money demand curve. For example, two chapters ago, you learned about the effects of expectations on money demand. If people expect the interest rate to rise (the price of bonds to fall) in the near future, they will want to hold less wealth in the form of

Passive monetary policy When the Fed keeps the money supply constant regardless of shocks to the economy.

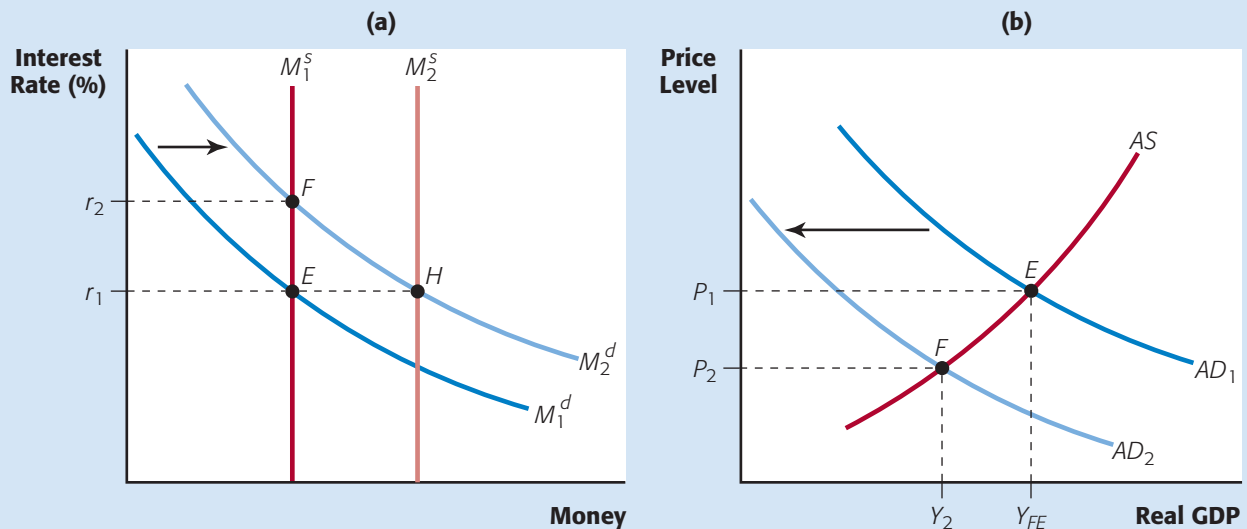
Active monetary policy When the Fed changes the money supply to achieve some objective.



What Happens When Things Change?

FIGURE 2

RESPONDING TO SHIFTS IN MONEY DEMAND



Beginning at point E in panel (a), an increase in money demand drives the interest rate up to r_2 (point F). Under a passive monetary policy, this would cause interest-sensitive spending to decrease. In panel (b), the aggregate demand curve would shift to AD_2 , decreasing GDP from Y_{FE} (at point E) to Y_2 (at point F). The economy would suffer a recession. To maintain full employment, the Fed could increase the money supply to M_2^s , preventing any change in the interest rate and any shift in AD .

bonds and more in money, so the money demand curve will shift rightward. Larger and longer-lasting shifts in the money demand curve may occur for reasons that are not well understood, although leading suspects are the development of new types of financial assets and new methods of making payments.

How should the Fed respond to shifts in the money demand curve? Figure 2 shows the effect of a rightward shift of the money demand curve. Look first at panel (a). Initially, the money market is in equilibrium at point E , with the interest rate equal to r_1 . When the money demand curve shifts rightward, to M_2^d , the equilibrium moves to point F , with the higher interest rate r_2 . With a passive monetary policy—leaving the money supply unchanged—the rise in the interest rate would cause interest-sensitive spending to fall. This, in turn, would decrease equilibrium GDP at any given price level.

Panel (b) shows another way to view the effect of the change in money demand: the AD curve shifts leftward, from AD_1 to AD_2 . With a passive monetary policy, the economy would slide down the AS curve from point E to point F , causing a recession. Since the economy began at full-employment output (Y_{FE}), the passive monetary policy would cause unemployment to rise above the natural rate, and the price level would decrease.

If the Fed wants to maintain full employment with zero inflation—an unchanged price level—then a passive monetary policy is clearly the wrong response. Is there a better policy?

Indeed there is—an *active* monetary policy. By increasing the money stock—shifting the money supply curve from M_1^s to M_2^s —the Fed moves the money market to a new equilibrium at point H , *preventing any rise in the interest rate*. If the Fed acts quickly enough, there will be no decrease in interest-sensitive spending and no

shift in the AD curve. In panel (b), the economy remains at point E , and the Fed maintains full employment with zero inflation.

As you can see, shifts in the money demand curve present the Fed with a no-lose situation: By adjusting the money supply to prevent changes in the interest rate, the Fed can achieve both price stability and full employment. During most periods, when the economy is not affected by any shocks other than money demand shifts, the constant interest rate policy will keep the economy on an even keel. This is why, in its day-to-day operations, the Fed sets and maintains an **interest rate target** and then adjusts the money supply to achieve that target.

To deal with money demand shocks, the Fed sets an interest rate target and changes the money supply as needed to maintain the target. In this way, the Fed can achieve its goals of price stability and full employment simultaneously.

Interest rate target The interest rate the Federal Reserve aims to achieve by adjusting the money supply.

How the Fed Keeps the Interest Rate on Target. A quick review of the day-to-day mechanics of Fed policy making shows how it sets and maintains its interest rate target in practice. Fed officials meet each morning to determine that day's monetary policy, based on information gathered the previous afternoon and earlier that morning. A key piece of information is what actually happened to the interest rate since the morning before. A rise in the interest rate means that the money demand curve has shifted rightward; a drop in the interest rate means the curve has shifted leftward.

Using this and other information about the banking system and the economy, the Fed decides what to do. At 11:30 A.M., if the interest rate is above target, the Fed buys government bonds. This increases the money supply and brings the interest rate back down to its target level, as in Figure 2. If, instead, the interest rate is below target, the Fed sells government bonds, decreasing the money supply and raising the interest rate back up to its target level.

RESPONDING TO SPENDING SHOCKS

The Fed has a somewhat more difficult job responding to spending shocks than to shifts in money demand. Figure 3 illustrates why. In panel (a) the money market is initially in equilibrium at point E , with the interest rate at its initial target level of r_1 . In panel (b) the economy's short-run equilibrium is at point E , with output at full employment.

Now suppose that there is a positive spending shock. The shock might originate with the government—an increase in government purchases or a decrease in taxes—or in the private sector—an increase in investment or autonomous consumption or net exports. Whatever the source, the impact in panel (b) is the same: The AD curve will shift rightward—from AD_1 to AD_2 —and output will rise. Back in panel (a), the rise in output will shift the money demand curve rightward to M_2^d , raising the interest rate. Now let's consider three possible responses by the Fed.

First, the Fed could follow a *passive* monetary policy, leaving the money supply unchanged. In this case, the interest rate would be allowed to rise above its target. In panel (b), the economy would slide upward along the AS curve, moving to point F . Both output and the price level would rise.

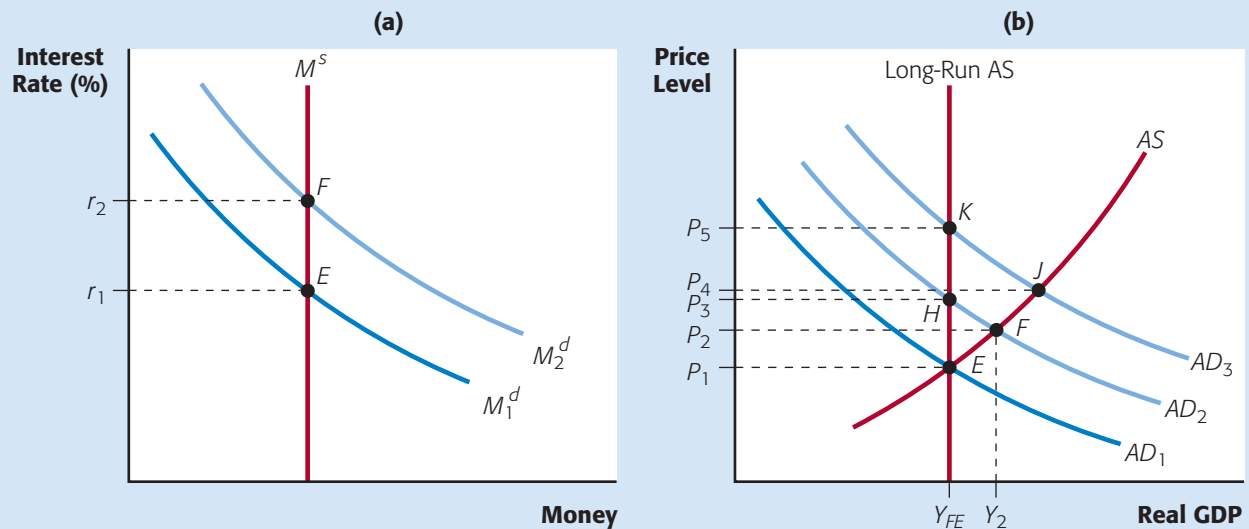
As you can see, the Fed would not want to respond to a spending shock with a passive monetary policy. Output would rise, bringing the unemployment rate below the natural rate. The price level would rise as well—to P_2 . And in the long run, the price level would rise further—to P_3 —as the self-correcting mechanism returned the economy to full employment at point H .



What Happens When Things Change?

FIGURE 3

RESPONDING TO SPENDING SHOCKS



A positive spending shock would shift the AD curve rightward to AD_2 in panel (b), causing both the price level and output to rise. Under a passive monetary policy, that rise in income would cause the money demand curve to shift to M_2^d in panel (a), driving the interest rate upward from r_1 to r_2 .

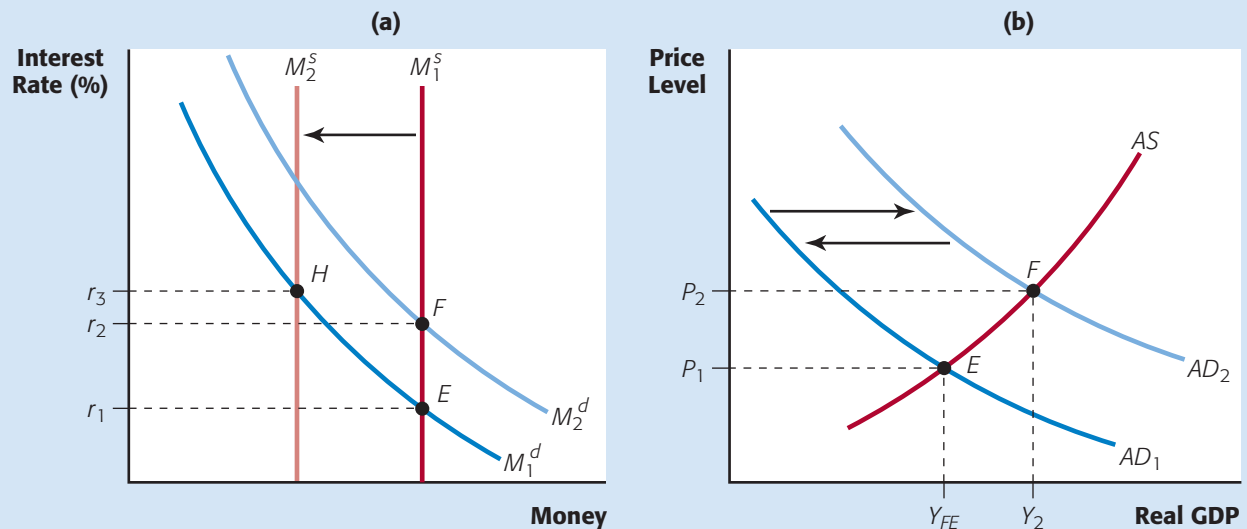
An active policy of maintaining the interest rate at r_1 would make matters worse. To maintain the interest rate target, the Fed would have to increase the money supply, causing an additional rightward shift of the AD curve to AD_3 and pushing the economy even further above full employment. The price level would increase to P_4 in the short run and P_5 in the long run.

Would the active policy described earlier—maintaining an interest rate target—be an improvement? Actually, no—it would be even worse. To maintain the interest rate at r_1 , the Fed would have to *increase* the money supply (not shown). But with no rise in the interest rate to crowd out some consumption and investment spending, the spending shock would shift the AD curve rightward even further—say, to AD_3 . The new short-run equilibrium would then be at point J . As you can see, maintaining the interest rate target would push the economy even further beyond its potential output, and increase the price level even more—both in the short run (to P_4) and in the long run (P_5).

How, then, should the Fed respond to the spending shock? To maintain full employment and a stable price level, the Fed must pursue an active policy, but one that shifts the AD curve back to AD_1 . And it can, indeed, do so. Look at Figure 4. Once again, the figure shows a spending shock that shifts the AD curve rightward to AD_2 , increasing both output and the price level. In the money market, the higher price level and higher income shift the money demand curve rightward, raising the interest rate to r_2 . But, as you saw in Figure 3, the rise to r_2 is not enough to choke off the increase in spending; it causes *some* crowding out of consumption and investment, but not *complete* crowding out. In order to shift the AD curve back to AD_1 , the Fed must raise the interest rate *further*, enough to cause *complete* crowding out. That is, it must raise the interest rate by just enough so that consumption and investment spending decline by an amount equal to the initial spending shock. In the figure, we assume that an interest rate of r_3 will do the trick (point H). The Fed must decrease the money supply to M_2^s . If the Fed acts

THE BEST RESPONSE TO A SPENDING SHOCK

FIGURE 4



A spending shock that shifts the AD curve to AD_2 threatens to raise output beyond its full-employment level, and increase the price level as well. The Fed can neutralize that shift by decreasing the money supply to M_2^s . The resulting rise in the interest rate (to r_3) would reduce interest-sensitive spending and return the AD curve to AD_1 .

quickly enough, it can prevent the spending shock from shifting the AD curve at all.²

To maintain full employment and price stability after a spending shock, the Fed must change its interest rate target. A positive spending shock requires an increase in the target; a negative spending shock requires a decrease in the target.

In recent years, the Fed has changed its interest rate target as frequently as needed to keep the economy on track. If the Fed observes that the economy is overheating—and that the unemployment rate has fallen below its natural rate—it will raise its target. The Fed—believing that the AD curve was shifting rightward too rapidly—reacted this way in 1999 and early 2000, raising its interest rate target four times in just nine months. When the Fed raises its target, it responds to forces that shift the AD curve rightward by creating an opposing force—a higher interest rate—to shift it leftward again.

When the Fed observes that the economy is sluggish—and the unemployment rate has risen above its natural rate—the Fed will lower its target. This tends to neutralize leftward shifts of the AD curve.

As you can see, spending shocks present the Fed with another no-lose situation: The same policy that helps to keep unemployment at its natural rate also helps to maintain a stable price level. However, spending shocks present a challenge to the Fed that it doesn't face during other, less-eventful periods. To change the interest rate target by just the right amount, the Fed needs accurate information about how

² Notice that the new money market equilibrium is along the original money demand curve M_1^d , since the policy will return both the price level and income to their original values.

the economy operates. We'll return to this and other problems in conducting monetary policy in the "Using the Theory" section of this chapter.

The Interest Rate Target and the Financial Markets. The members of the Open Market Committee think very hard before they vote to change the interest rate target. In addition to its effects on the level of output and the price level, changes in the interest rate target can create turmoil in the stock and bond markets.

Why? Recall that the interest rate and the price of bonds are negatively related. Thus, when the Fed moves the interest rate to a higher target level, the price of bonds drops. Because the public holds trillions of dollars in government and corporate bonds, even a small rise in the interest rate—say, a quarter of a percentage point—causes the value of the public's bond holdings to drop by billions of dollars.

The stock market is often affected in a similar way. People hold stocks because they entitle the owner to a share of a firm's profits, and because stock prices are usually expected to rise as the economy grows and firms become more profitable. But stocks must remain competitive with bonds, or else no one would hold them. The lower the price of a stock, the more attractive the stock is to a potential buyer.

When the Fed raises the interest rate, the rate of return on bonds increases, so bonds become more attractive. As a result, stock prices must fall, so that stocks, too, will become more attractive. And that is typically what happens. Unless other changes are affecting the stock market, a rise in the interest rate causes people to try to sell their stocks in order to acquire the suddenly-more-attractive bonds. This causes stock prices to fall, until stocks are once again as attractive as bonds. Thus, a rise in the interest rate causes stock prices, as well as bond prices, to fall:

The stock and bond markets move in the opposite direction to the Fed's interest rate target: When the Fed raises its target, stock and bond prices fall; when it lowers its target, stock and bond prices rise.

The destabilizing effect on stock and bond markets is one reason the Fed prefers not to change its interest rate target very often. Frequent changes in the target would make financial markets less stable, and the public more hesitant to supply funds to business firms by buying stocks and bonds.

Importantly, financial markets are also affected by *expected* changes in the interest rate target—whether or not they occur. If you expect the Fed to raise its target, you also expect stock and bond prices to fall. Therefore, you would want to dump these assets *now*, before their price drops. Similarly, an expectation of a drop in the interest rate target would make you want to buy stocks and bonds now, before their prices rise. Thus, *changes in expectations* about the Fed's future actions can be as destabilizing as the actions themselves.

This is why the financial press speculates constantly about the likelihood of changes in the interest rate target. Most of the time, the news is of the dog that didn't bark—the Federal Open Market Committee meets and decides to keep the target unchanged. Still, interest rates and stock prices often jump around in the days leading up to meetings of the Open Market Committee.

Once you understand the Fed's logic in changing its interest rate target, you can understand a phenomenon that—at first glance—appears mystifying: Stock and bond prices often fall when good news about the economy is released, and rise when bad news is released. For example, if the Bureau of Labor Statistics announces that jobs are plentiful and the unemployment rate has dropped, or the Commerce Department announces that real GDP has grown rapidly in the previous quarter, the stock and bond markets may plummet. Why? Because owners of stocks and bonds



Financial Markets react when people expect—as they did in early May, 2000—that the Fed may change its interest rate target.

RESPONDING TO SUPPLY SHOCKS

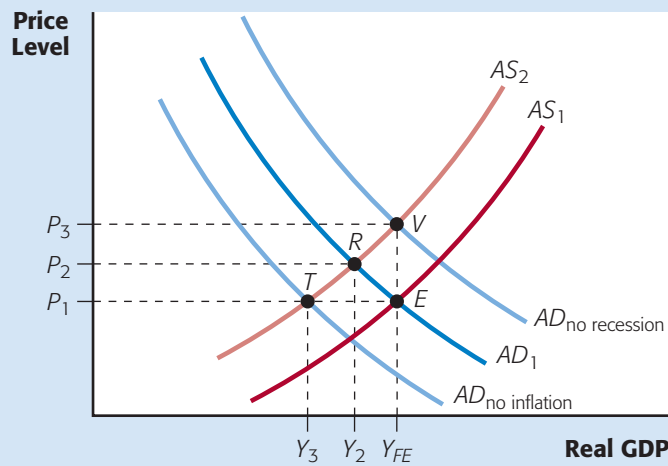


FIGURE 5

Starting at point E , a negative supply shock shifts the AS curve upward to AS_2 . Under a passive monetary policy, a new short-run equilibrium would be established at point R , with a higher price level (P_2) and a lower level of output (Y_2). The Fed could prevent inflation by decreasing the money supply and shifting AD to $AD_{\text{no inflation}'}$, but output would fall to Y_3 . At the other extreme, it could increase the money supply and shift the AD curve to $AD_{\text{no recession}'}$. This would keep output at the full-employment level, but at the cost of a higher price level, P_3 .

know that the Open Market Committee might interpret the good news as evidence that the economy is overheating. They expect the Committee to raise its interest rate target, so they try to sell their stocks and bonds before the committee even meets.


Good news about the economy sometimes leads to expectations that the Fed—fearing inflation—will raise its interest rate target. This is why good economic news sometimes causes stock and bond prices to fall. Similarly, bad news about the economy sometimes leads to expectations that the Fed—fearing recession—will lower its interest rate target. This is why bad economic news sometimes causes stock and bond prices to rise.³

RESPONDING TO SUPPLY SHOCKS

So far in this chapter, you've seen that demand shocks, in general, present the Fed with easy policy choices. By sticking to its interest rate target, it can neutralize any demand shocks that arise from shifts in money demand. And by changing its interest rate target from time to time, it can deal with demand shocks caused by changes in spending. In each of these cases, the very policy that maintains a stable price level also helps to maintain full employment.

But adverse or negative *supply* shocks present the Fed with a true dilemma: If the Fed tries to preserve price stability, it will worsen unemployment; if it tries to maintain high employment, it will worsen inflation. And even though supply shocks are usually temporary, the shocks themselves—and the Fed's response—can affect the economy for several quarters or even years.

Figure 5 illustrates the Fed's dilemma when confronting an adverse supply shock. Initially, the economy is at point E (full employment). Then, a supply shock—say, a rise in world oil prices—shifts the AS curve up to AS_2 . Under a passive monetary policy, the Fed would not change the money stock, keeping the AD curve at AD_1 . The short-run equilibrium would then move from point E to point R ,

 What Happens When Things Change?

³ For a more complete discussion of the stock market, see the Using All the Theory chapter at the end of this book.



For more information about supply shocks, download Bharat Trehan's "Supply Shocks and the Conduct of Monetary Policy," available at <http://www.frbsf.org/econsrch/wklytr/wklytr99/el99-21.html>.

and the economy would experience stagflation—both inflation and a recession—with output falling to Y_2 and the price level rising to P_2 .

But the Fed can instead respond with an active monetary policy, changing the money stock in order to alter the short-run equilibrium. Which policy should it choose? The answer will depend on whether it is mostly concerned about rising prices or rising unemployment. Let's start by imagining two extreme positions.

First, the Fed could prevent inflation entirely by decreasing the money stock, shifting the AD curve leftward to the curve labeled $AD_{\text{no inflation}}$. This would move the short-run equilibrium to point T . Notice, though, that while the price level remains at P_1 , output decreases to Y_3 —even lower than under the passive policy.

At the other extreme, the Fed could prevent any fall in output. To accomplish this, the Fed would *increase* the money stock and shift the AD curve rightward, to $AD_{\text{no recession}}$. The equilibrium would then move to point V , keeping output at its full-employment level. But this policy causes more inflation, raising the price level all the way to P_3 .

In practice, the Fed is unlikely to choose either of these two extremes to deal with a supply shock, preferring instead some intermediate policy. But the extreme positions help illustrate the Fed's dilemma:

An adverse supply shock presents the Fed with a short-run trade-off: It can limit the recession, but only at the cost of more inflation; and it can limit inflation, but only at the cost of a deeper recession.

The choice between the two policies is a hard one. After supply shocks, there are often debates within the Fed—and in the public arena—about how best to respond. Inflation *hawks* lean in the direction of price stability, and are willing to tolerate more unemployment in order to achieve it. In the face of an adverse supply shock, hawks would prefer a response that shifts the AD curve closer to $AD_{\text{no inflation}}$, even though it means higher unemployment. Inflation *doves* lean in the direction of a milder recession, and are more willing to tolerate the cost of higher inflation. They would prefer a response that brings the AD curve closer to $AD_{\text{no recession}}$.

Choosing Between Hawk and Dove Policies. When a supply shock hits, should the Fed use a hawk policy, should it employ a dove policy, or should it keep the AD curve unchanged? That depends. Over time, as the economy is hit by supply shocks, the hawk policy maintains more stability in the price level, but less stability in output and employment. The dove policy gives the opposite result: more stability in output and less stability in the price level. The Fed should choose a hawkish policy if it cares more about price stability, and a dovish policy if it cares more about the stability of output and employment. Or it can pick an intermediate policy—one that balances price and employment stability more evenly.

The proper choice depends on how the Fed weights the harm caused by unemployment against the harm caused by inflation. And since the Fed is a public institution, its views should reflect the assessment of society as a whole. This is why supply shocks present such a challenge to the Fed: The public itself is divided between hawks and doves. Both inflation and unemployment cause harm, but of very different kinds. Inflation imposes a more general cost on society—the resources used up to cope with it. If the inflation is unexpected, it will also redistribute income between borrowers and lenders. The costs of unemployment are borne largely by the unemployed themselves—who suffer the harm of job loss—but partly by taxpayers, who provide funds for unemployment insurance. Balancing the gains and losses from hawk and dove policies is no easy task.

In recent years, some officials at the Fed have argued that having two objectives—stable prices *and* full employment—is unrealistic when there are supply

shocks. The current chair of the Board of Governors, Alan Greenspan, has asked Congress to change the Fed's mandate to one of controlling inflation, period. But it would be difficult for the Fed to ignore the costs of higher unemployment, even if it was legally permitted to do so. Regardless of any future change in the Fed's mandate, the debate between hawks and doves is destined to continue.

EXPECTATIONS AND ONGOING INFLATION

So far in this chapter, we've assumed that the Fed strives to maintain *zero* inflation, and that the price level remains constant when the economy reaches its long-run, full-employment equilibrium. But as we discussed earlier, this is not entirely realistic. Look again at panel (a) of Figure 1. There you can see that the U.S. economy has been characterized by *ongoing inflation*. Even in the 1990s—with unemployment at its natural rate—the annual inflation rate has hovered around 2 to 3 percent. That means that, even though the economy is at full employment, prices are *continually rising*.

Why should the price level continue to rise when unemployment is at its natural rate? And how does ongoing inflation change our analysis of the effects of monetary policy, or the guidelines that the Fed should follow? We'll consider these questions next.

HOW ONGOING INFLATION ARISES

The best way to begin our analysis of ongoing inflation is to explore how it arises in an economy. We can do this by revisiting the 1960s, when the inflation rate rose steadily, and ongoing inflation first became a public concern.

What was special about the economy in the 1960s? First, it was a period of exuberance and optimism, for both businesses and households. Business spending on plant and equipment rose, and household spending on new homes and automobiles rose as well. At the same time, government spending rose—both military spending for the war in Vietnam and social spending on programs to help alleviate poverty. These increases in spending all contributed to rightward shifts of the *AD* curve—they were positive demand shocks. The unemployment rate fell below the natural rate—hovering around 3 percent in the late 1960s. And, as expected, the economy's self-correcting mechanism kicked in: Higher wages shifted the *AS* curve upward, causing the price level to rise.

As you've learned in this chapter, the Fed could have neutralized the positive demand shocks by raising its interest rate target (as in Figure 4), shifting the *AD* curve back to its original position. Alternatively, the Fed could have done nothing, allowing the self-correcting mechanism to bring the economy back to full employment with a higher—but stable—price level (as in the move from point *F* to point *H* in Figure 3). But in the late 1960s, the Fed made a different choice: It maintained its low interest rate target. This required the Fed to increase the money supply, thus adding its *own* positive demand shock to the spending shocks already hitting the economy. In Figure 3, this was the equivalent of moving the *AD* curve all the way out to *AD*₃, preventing any rise in interest rates but overheating the economy even more.

Why did the Fed act in this way? No one knows for sure, but one likely reason is that, in the 1960s, the Fed saw its job differently than it does today. The Fed's goal was to keep the interest rate stable and low, both to maintain high investment spending and to avoid instability in the financial markets. This is what it had been doing for years, with good effect: Americans had prospered in the previous decade, the 1950s, and financial markets were, indeed, stable.

But while this policy worked well in the 1950s, it did not serve the economy well during and after the demand shocks of the 1960s. That's because the Fed's



Some income-distributional aspects of monetary policy are explored by Christina and David Romer in "Monetary Policy and the Well-Being of the Poor." It is available at <http://www.kc.frb.org/publicat/econrev/PDF/lq99romr.pdf>.

policy—year after year—prevented the self-correcting mechanism from bringing the economy back to full employment. Instead, each time the price level began rising, and the economy began to self-correct, the Fed would increase the money supply *again*, causing output to remain *continually* above its potential output. And that, in turn, meant that the price level would continue to rise, year after year.

Now comes a crucial part of the story: As the price level continued to rise in the 1960s, the public began to *expect* it to rise at a similar rate in the future. This illustrates a more general principle:

When inflation continues for some time, the public develops expectations that the inflation rate in the future will be similar to the inflation rates of the recent past.

Why are expectations of inflation so important? Because when managers and workers expect inflation, it gets built into their decision-making process. Union contracts that set wages for the next three years will include automatic increases to compensate for the anticipated loss of purchasing power caused by future inflation. Non-union wages will tend to rise each year as well, to match the wages in the unionized sector. And contracts for future delivery of inputs—like lumber, cement, and unfinished goods—will incorporate the higher prices everyone expects by the date of delivery.

A continuing, stable rate of inflation gets built into the economy. The built-in rate is usually the rate that has existed for the past few years.

Once there is built-in inflation, the economy continues to generate continual inflation even *after* the self-correcting mechanism has finally been allowed to do its job and bring us back to potential output. To see why, look at Figure 6. It shows what might happen over three years in an economy with built-in inflation. In the figure, output is at its full-employment level. Each year, the AS curve shifts upward, and the AD curve shifts rightward, so the price level rises from P_1 to P_2 to P_3 . Why does all this happen when there is built-in inflation?

Let's start with the reason for the upward shift of the AS curve. Unemployment is at its natural rate, so the self-correction mechanism is no longer contributing to any rise in wages or unit costs. But something else *is* causing unit costs to increase: inflationary expectations. Based on recent experience, the public expects the price level to rise as it has been rising in the past, so wages and input prices will continue to increase, *even though output remains unchanged at full employment*. Thus,

in an economy with built-in inflation, the AS curve will shift upward each year, even when output is at full employment and unemployment is at its natural rate. The upward shift of the AS curve will equal the built-in rate of inflation.

For example, if the public expects inflation of 3 percent per year, then contracts will call for wages and input prices to rise by 3 percent per year. This means that unit costs will increase by 3 percent. Firms—marking up prices over unit costs—will raise their prices by 3 percent as well, and the AS curve will shift upward by 3 percent each year.

Explaining why the AS curve shifts upward is only half the story of the long-run equilibrium in Figure 6. We must also explain why the AD curve continues to shift rightward. The simple answer is: The AD curve shifts rightward because the Fed continues to increase the money supply. But *why* does the Fed shift the AD curve rightward, when it knows that doing so only prolongs inflation? One reason is that reducing inflation is *costly* to the economy.

LONG-RUN EQUILIBRIUM WITH BUILT-IN INFLATION

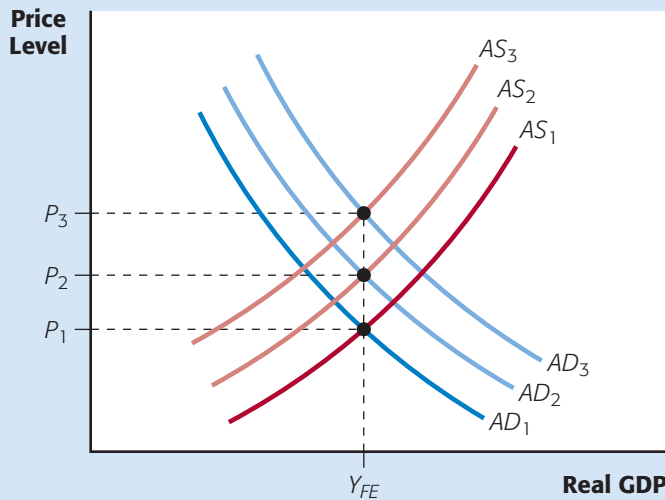


FIGURE 6

Each year, the aggregate supply curve shifts upward by the built-in rate of inflation. To keep the economy at full employment, the Fed shifts the AD curve rightward each year by increasing the money supply.

Imagine what would happen if, one year, the Fed decided *not* to shift the AD curve rightward as it had done in the past. During the year, the AS curve will shift upward anyway, by a percentage shift equal to the built-in rate of inflation. This will happen *no matter what the Fed does*, because the shift is based on wage and price decisions that, in turn, are based on past experiences of inflation. There is nothing the Fed can do today to affect what has happened in the past, so each year, it must accept the upward shift of the AS curve as a given.

But now suppose the Fed decides to reduce inflation by *not* shifting the AD curve as it has in the past. Instead, it will just leave the AD curve where it was the year before. For example, as the AS curve shifts from AS_2 to AS_3 , the Fed might keep the AD curve at AD_2 . See if you can draw the new, temporary equilibrium that the Fed will achieve for the economy. (*Hint*: It's at the intersection of AD_2 and AS_3). If you've identified the point correctly, you'll see that the Fed would achieve its goal of bringing down inflation this year. The price level would rise from P_2 to something less than P_3 , instead of all the way to P_3 . But the reduction in inflation is not without cost: The economy's output will decline—a recession.

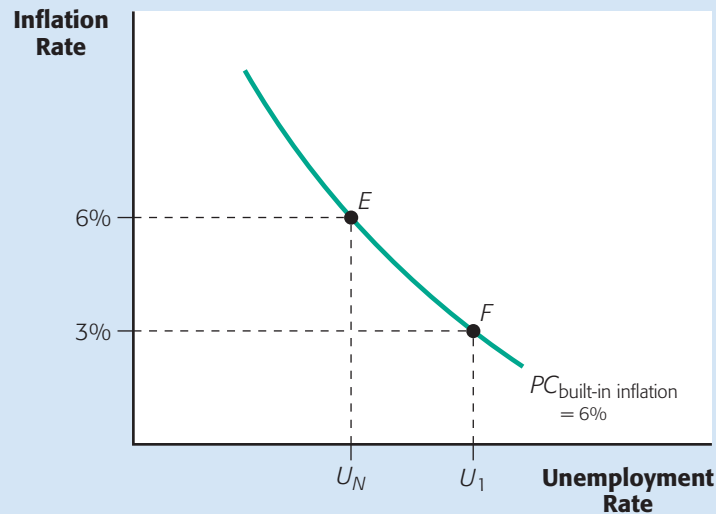
In the short run, the Fed can bring down the rate of inflation by reducing the rightward shift of the AD curve, but only at the cost of creating a recession.

Would the Fed ever purposely create a recession to reduce inflation? Indeed it would, and it has—more than once. By far the most important episode occurred during the early 1980s. As Figure 1 shows, inflation reached the extraordinary level of 14.8 percent in early 1980. Soon after, with the support of the newly elected President Reagan, the Fed embarked on an aggressive campaign to bring inflation down. The Fed stopped increasing the money supply, stopped shifting the AD curve rightward, and a recession began in July of 1981. Unemployment peaked, as shown in Figure 1, at 10.7 percent at the end of 1982. With tremendous slack in the economy, the inflation rate fell rapidly, to below 4 percent in 1982. The Fed deliberately created a serious recession, but it brought down the rate of inflation.

FIGURE 7

The Phillips curve illustrates possible combinations of inflation and unemployment for the economy in the short run, with a given built-in inflation rate. Point E represents a long-run equilibrium, with the economy at the natural rate of unemployment, U_N , and inflation at the built-in rate of 6%. If the Fed wishes to decrease the inflation rate to 3%, it must accept a higher short-run unemployment rate— U_1 at point F .

THE PHILLIPS CURVE



Creating a recession is not a decision that the Fed takes lightly. Recessions are costly to the economy and painful to those who lose their jobs. The desire to avoid a recession is one reason that the Fed tolerates ongoing inflation and continues to play its role by shifting the AD curve rightward. We'll discuss other reasons for the Fed's tolerance of ongoing inflation a bit later.

ONGOING INFLATION AND THE PHILLIPS CURVE

Ongoing inflation changes our analysis of monetary policy. For one thing, it forces us to recognize a subtle, but important, change in the Fed's objectives: While the Fed still desires full employment, its other goal—price stability—is not zero inflation, but rather a *low and stable inflation rate*.

Another difference is in the graphs we use to illustrate the Fed's policy choices. Instead of continuing to analyze the economy with AS and AD graphs, when there is ongoing inflation, we usually use another powerful tool.

This tool is the *Phillips curve*—named after the late economist A. W. Phillips, who did early research on the relationship between inflation and unemployment. The **Phillips curve** illustrates the Fed's choices between inflation and unemployment in the short run, for a given built-in inflation rate.

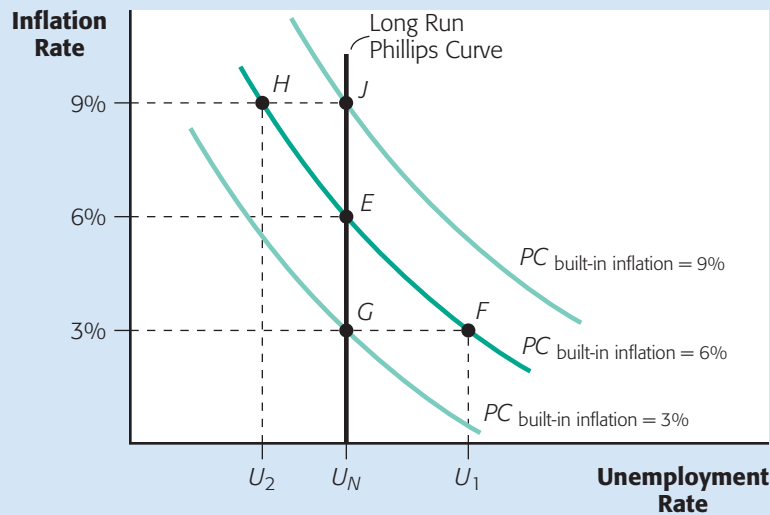
Figure 7 shows a Phillips curve for the U.S. economy. The inflation rate is measured on the vertical axis, the unemployment rate on the horizontal. Point E shows the long-run equilibrium in the economy when the built-in inflation rate is 6 percent. At point E , unemployment is at its natural rate— U_N —and inflation remains constant from year to year at the built-in rate of 6 percent.

Notice that the Phillips curve is downward sloping. Why? Because it tells the same story we told earlier—with AD and AS curves—about the Fed's options in the short run. If the Fed wants to decrease the rate of inflation from 6 percent to 3 percent, it must slow the rightward shifts of the AD curve. This would cause a movement *along* the Phillips curve from point E to point F . As you can see, in moving to point F , the economy experiences a recession: Output falls, and unemployment rises above the natural rate.

Phillips curve A curve indicating the Fed's choice between inflation and unemployment in the short run.

THE SHIFTING PHILLIPS CURVE

FIGURE 8



Initially, the economy is at point E , with inflation equal to the built-in rate of 6%. If the Fed moves the economy to point F and keeps it there, the public will eventually come to expect 3% inflation in the future. At that point, the built-in inflation rate will fall and the curve will shift down to $PC_{\text{built-in inflation} = 3\%}$. The economy will move to point G in the long run, with unemployment at the natural rate and an actual inflation rate equal to the built-in rate of 3%.

Starting again at point E , a demand shock that was not neutralized by the Fed would move the economy to point H ; the inflation rate would rise to 9 percent, and the unemployment rate would fall to U_2 . If the Fed then held the economy at point H , the built-in inflation rate would rise to 9%, and the Phillips curve would shift up to $PC_{\text{built-in inflation} = 9\%}$. Eventually, the economy would move to point J . The vertical line connecting points E , G , and J is the long-run Phillips curve.

In the short run, the Fed can move along the Phillips curve by adjusting the rate at which the AD curve shifts rightward. When the Fed moves the economy downward and rightward along the Phillips curve, the unemployment rate increases, and the inflation rate decreases.

Now suppose the Fed keeps the economy at point F . In the long run, the public—observing a 3-percent inflation rate—will come to expect 3-percent inflation in the future. Thus, in the long run, 3 percent will become the economy's built-in rate of inflation. Figure 8 shows the effect on the Phillips curve: It shifts downward, to the lower curve. At any unemployment rate, the inflation rate will be lower, now that the public expects inflation of only 3 percent, rather than 6 percent.

In the long run, a decrease in the actual inflation rate leads to a lower built-in inflation rate, and the Phillips curve shifts downward.

Once the Fed has reduced the built-in inflation rate, it can locate anywhere on the new Phillips curve by adjusting how rapidly it lets the money supply grow (and therefore, how rapidly the AD curve shifts rightward each year). Therefore, the Fed can choose to bring the economy back to full employment (point G), with a new, lower inflation rate of 3 percent, rather than the previous 6 percent.

Riding Up the Phillips Curve. The process we've described—moving down the Phillips curve and thereby causing it to shift downward—also works in reverse: Moving *up* the Phillips curve will cause it to shift *upward*. Figure 8 also illustrates this case. Once again, the economy begins at point *E*, with a built-in inflation rate of 6 percent and unemployment at its natural rate. Now suppose the Fed begins to increase the money supply *more rapidly* than in the past, and—in Figure 6—begins shifting the *AD* curve further rightward than before. In the short run, the economy would move *along* the Phillips curve from point *E* to point *H* in Figure 8. The inflation rate would rise to 9 percent, and the unemployment rate would fall below its natural rate—in the short run.

But suppose the Fed keeps the economy at point *H* for some time—continuing to shift the *AD* curve rightward at a faster rate than before. Then, in the long run, the public will begin to expect 9-percent inflation, and that will become the new built-in rate of inflation. The Phillips curve will then shift upward. At this point, if the Fed returns the economy to full employment, we end up at point *J*. The economy will be back in long-run equilibrium—but with a higher built-in inflation rate.

The Long-Run Phillips Curve. In Figure 8, you can see that the Fed's policy choices are different in the short run and in the long run. In the short run, the Fed can move along the Phillips curve, exploiting the trade-off between unemployment and inflation. But in the long run—once the public expectations of inflation adjust to the new reality—the built-in inflation rate will change, and the Phillips curve will shift. Indeed, the Phillips curve will *keep* shifting whenever the unemployment rate is kept above or below the natural rate. (To see why, ask yourself what would happen in the future if the Fed tried to keep the unemployment rate *permanently* at a level like U_2 —below the natural rate?) Thus, in the long run, the unemployment rate must eventually return to the natural rate, and output must go back to its potential level. In the long run, the Fed can only choose *which* Phillips curve the economy will be on at that time. That is,

in the short run, there is a trade-off between inflation and unemployment: The Fed can choose lower unemployment at the cost of higher inflation, or lower inflation at the cost of higher unemployment. But in the long run, since unemployment always returns to its natural rate, there is no such trade-off.

Now let's reconsider what we've learned about the Fed's options in the long run. Figure 8 shows us that, when the Fed slows the rightward shifts of the *AD* curve, unemployment returns to the natural rate, but the inflation rate is lower. The figure also shows us that, when the Fed allows the *AD* curve to shift rightward more rapidly than in the past, unemployment returns once again to the natural rate, but the inflation rate is higher. As you can see,

in the long run, monetary policy can change the rate of inflation, but not the rate of unemployment.

Now look at the vertical line in Figure 8. It tells us how monetary policy affects the economy in the long run, without the distractions of the short-run story. The vertical line is the economy's **long-run Phillips curve**, which tells us the combinations of unemployment and inflation that the Fed can choose in the long run. No matter what the Fed does, unemployment will always return to the natural rate, U_N , in the long run. However, the Fed can use monetary policy to select any rate of inflation it wants:

Long-run Phillips curve A vertical line indicating that in the long run, unemployment must equal its natural rate, regardless of the rate of inflation.

The long-run Phillips curve is a vertical line at the natural rate of unemployment. The Fed can select any point along this line in the long run, by using monetary policy to speed or slow the rate at which the AD curve shifts rightward.

WHY THE FED ALLOWS ONGOING INFLATION

Since the Fed can choose any rate of inflation it wants, and since inflation is costly to society, we might think that the Fed would aim for an inflation rate of zero. But a look back at panel (a) of Figure 1 shows that this is not what the Fed has chosen to do. In recent years, with unemployment very close to its natural rate, the Fed has maintained annual inflation at around 2 or 3 percent. Why doesn't the Fed eliminate inflation from the economy entirely?

One reason is a widespread belief that the Consumer Price Index (CPI) and other measures of inflation actually *overstate* the true rate of inflation in the economy. As you've learned, many economists believe that the CPI has overstated the true inflation rate by 1 to 2 percent per year. Although the Bureau of Labor Statistics has been working hard to correct the problem, some significant upward bias remains. If the Fed forced the *measured* rate of inflation down to zero, the result would be a true rate of inflation that was negative—prices would actually *fall* each year. But negative rates of inflation can be as costly to society as positive rates: People are as likely to make errors in financial planning when the price level is falling at 2 percent per year as they are when the price level is rising at the same rate. And if the price level drops by more or less than expected, real income will be shifted between borrowers and lenders.

Some economists have offered another explanation for the Fed's behavior: Low, stable inflation makes the labor market work more smoothly. The argument goes as follows: While no one wants a cut in their *real* wage rate, people seem to react differently, depending on *how* the real wage is decreased. For example, suppose there is an excess supply of manufacturing workers, and a wage cut of 3 percent is needed to bring that labor market back to equilibrium. Workers would strongly resist a 3-percent cut in the nominal wage. But they would more easily tolerate a freeze in the nominal wage while the price level rises by 3 percent, even though in both scenarios, the real wage falls by 3 percent. If this argument is correct, then a low or modest inflation rate would help wages adjust in different markets. In some labor markets, real wages can be raised by increasing nominal wages faster than prices. In other labor markets, real wages can be cut by increasing nominal wages more slowly than prices, or not at all.

But the strongest reason for the Fed's tolerance of low inflation is one we've already discussed: Once inflation is built into the economy, it is costly to reduce it. For example, to reduce the built-in inflation rate from its current 2 or 3 percent, the Fed would have to engineer a recession. Even if the Fed believed that the economy would be better off with lower inflation, it would not necessarily choose to pursue this goal. In fact, as a result of the Fed's success in controlling inflation for the past several years, popular concern about inflation has practically disappeared. Since a further reduction in inflation is not valued highly by the public, it is not politically worthwhile to pay the costs of achieving it.

The Fed has tolerated measured inflation at 2 to 3 percent per year because it knows that the true rate of inflation is lower, because low rates of inflation may help labor markets adjust more easily, and because there is not much pay-off to lowering inflation further.

Using the THEORY

CONDUCTING MONETARY POLICY IN THE REAL WORLD

So far in this chapter, we've described some clear-cut guidelines the Fed *can* and *does* follow in conducting monetary policy. We've seen that the proper policy for dealing with day-to-day changes in money demand is to set and maintain an interest rate target. The proper response to a spending shock is a change in the interest rate target. Dealing with a supply shock is more problematic, since it requires the Fed to balance its goal of low, stable inflation with its goal of full employment. But even here, once the Fed decides on the proper balance, its policy choice is straightforward: Shift the *AD* curve to achieve the desired combination of inflation and unemployment in the short run, and then bring the economy back to full employment in the long run.

In most of our discussion, we've assumed that the Fed has all of the information it needs to determine where the economy *is* operating, where it *should* be operating, and what change in monetary policy will get it there. Unfortunately, the real world is not that simple: The information available to the Federal Open Market Committee is far from perfect. As a result, the Fed's selection and execution of policy are sometimes more complicated than we've suggested so far. In this section, we'll consider some of the problems of monetary policy, and how the Fed has adapted to them.

INFORMATION ABOUT THE MONEY DEMAND CURVE

The easiest job facing the Fed is responding to shifts in the money demand curve. As you saw in Figure 2, the Fed can stop money demand shocks from affecting output or the price level by adjusting the money stock to keep the interest rate unchanged. If the money demand curve shifts to M_2^d , the interest rate rises, so the Fed knows it must increase the money supply to keep the interest rate at r_1 . The Fed maintains the interest rate by moving *along* the new money demand curve.

But Figure 2 also reveals a potential problem: The Fed cannot know how much to increase the money stock unless it knows the *slope* of the new money demand curve. For example, if the money demand curve has become flatter, the Fed will have to increase the money supply beyond M_2^s in order to maintain its interest rate target.

How does the Fed deal with this problem? In two ways. First, the Fed's research staff tries to estimate the changes in the position and slope of the money demand curve from available data. While the techniques are not perfect, they enable the Fed to make reasonable guesses about the required change in the money supply on any given day.

Second, the Fed uses the trial-and-error procedure that we discussed earlier. For example, suppose the interest rate rises and Fed officials underestimate the required change in the money supply. Then the interest rate will remain above its target rate, and the Fed can try again the next day, increasing the money supply further. In recent years, using a combination of research on the one hand and trial and error on the other, the Fed has been quite successful in reaching and maintaining its interest rate target.

INFORMATION ABOUT THE SENSITIVITY OF SPENDING TO THE INTEREST RATE

The proper response to spending shocks presents the Fed with a more significant problem. Look back at Figure 3, in which a positive spending shock shifts the *AD* curve rightward to AD_2 . The Fed will want to neutralize the shock by shifting the

AD curve back to AD_1 . To do so, it will raise its interest rate target. But by how much? That depends on the sensitivity of consumption and investment spending to the interest rate. If spending is *very* sensitive to interest rate changes, only a small rise in the target is needed; if spending is less sensitive, the Fed will need to raise its target rate by more.

Once again, the Fed addresses this problem with both research and trial-and-error methods. The research in this case focuses on how households and businesses change their spending plans when the interest rate rises and falls. This enables the Fed to make reasonable guesses about the required change in the interest rate target.

Trial and error helps the Fed get even closer. Suppose, for example, that there is a positive spending shock that the Fed wants to neutralize. Suppose, too, that the Fed makes an error, and selects an interest rate target that is too low. Then the economy will begin to overheat. As output rises beyond full employment, the price level (or the inflation rate) will rise. The Fed then observes the changes in output and prices, and adjusts its interest rate target again.

There is one major drawback to this procedure, however: It may take many months for the Fed's error to show up. GDP is measured only once each quarter. The Consumer Price Index is released each month, but prices may be slow to adjust to the increase in output. Thus, in contrast to the case of money-demand shifts—where the Fed can correct its errors within days—spending shocks often require the Fed to “wing it” for many months.



UNCERTAIN AND CHANGING TIME LAGS

We've just seen that it can take many months before the Fed observes how a change in its interest rate target is affecting output and the price level. More importantly, the Fed does not know precisely *how many* months it will take. This presents a serious challenge for monetary policy. Suppose Fed officials believe that the economy is beginning to overheat, and they raise the interest rate target. The new, higher interest rate might not reduce spending for some time. Business firms will finish building the new plants and new homes that they've already started, even at higher interest rates. Investment spending will finally come down only at the point when canceled investment projects *would* have entered the pipeline, many months later. By the time the higher interest rate target has its maximum effect, the economy may be returning to full employment on its own, or it may be hit by a negative demand shock. In this case, the Fed—by raising its interest rate target—will be reining in the economy at just the wrong time, causing a recession.

Economists often use an analogy to describe this problem. Imagine that you are trying to drive a car with a special problem: When you step on the gas, the car will go forward . . . but not until five minutes later. Similarly, when you step on the brake, the car will slow, but also with a five-minute lag. It would be very difficult to maintain an even speed with this car: You'd step on the gas, and when nothing happened, you'd be tempted to step on it harder. By the time the car begins to move, you will have given too much gas and find yourself speeding down the road. So you try to slow down, but once again, hitting the brakes makes nothing happen. So you brake harder, and when the car finally responds, you come to a dead halt.

The Fed can make—and, in the past, has made—a similar mistake. When it tries to cool off an overheated economy, it may find that nothing is happening. Is it just

a long time lag, or has the Fed not hit the brakes hard enough? If it hits the brakes harder, it runs the risk of braking the economy too much; if it doesn't, it runs the risk of continuing to allow the economy to overheat. Even worse, the time lag before monetary policy affects prices and output can change over the years: Just when the Fed may think it has mastered the rules of the game, the rules change.

THE NATURAL RATE OF UNEMPLOYMENT

Finally, we come to the most controversial information problem facing the Fed: uncertainty over the natural rate of unemployment. While there is wide agreement that the natural rate rose in the 1970s and has fallen since the late 1980s, economists remain uncertain about its value during any given period. Many economists believe that today the natural rate is between 4 and 4.5 percent, but no one is really sure.

Why is this a problem? It's very much like the two mountain climbers who become lost. One of them pulls out a map. "Do you see that big mountain over there," he says, pointing off into the distance. "Yes," says the other. "Well," says the first, "according to the map, we're standing on top of it." In order to achieve its twin goals of full employment and a stable, low rate of inflation, the Fed tries to maintain the unemployment rate as close to the natural rate as possible. If its estimate of the natural rate is wrong, it may believe it has succeeded when, in fact, it has not.

For example, suppose the Fed believes the natural rate of unemployment is 4.5 percent, but the rate is really 4 percent. Then—at least for a time—the Fed will be steering the economy toward an unemployment rate that is unnecessarily high, and an output level that is unnecessarily low. We've already discussed the costs of cyclical unemployment; and an overestimate of the natural rate makes society bear these costs needlessly. On the other hand, if the Fed believes the natural rate is 4 percent when it is really 4.5 percent, it will overheat the economy. This will raise the inflation rate—and a costly recession may be needed later in order to reduce it.

Trial and error can help the Fed determine the true natural rate. If the Fed raises unemployment above the true natural rate, the inflation rate will drop. If unemployment falls below the true natural rate, the inflation rate will rise. But—as we discussed earlier—trial and error works best when there is continual and rapid feedback. It can take some time for the inflation rate to change—six months, a year, or even longer. In the meantime, the Fed might believe it has been successful, even while causing avoidable unemployment, or planting the seeds for a future rise in the inflation rate.

Estimating the natural rate of unemployment is made even more difficult because the economy is constantly buffeted by shocks of one kind or another. If the Fed observes that the inflation rate is rising, does that mean that unemployment is below the natural rate? Or is the higher inflation being caused by a negative supply shock? Or by the Fed's response to an earlier, negative demand shock? This information is difficult to sort out, although the Fed has become increasingly sophisticated in its efforts to do so.

As you can see, conducting monetary policy is not easy. The Fed has hundreds of economists carrying out research and gathering data to improve its information about the status of the economy and its understanding of how the economy works. And the effort seems to have paid off, especially over the last decade. But years from now, this period may be seen as the golden age of successful monetary policy. After all, the 1950s also seemed to be a period of good policy, but then the 1960s and especially the 1970s turned into disasters for monetary policy. Because we don't know

what kinds of shocks will hit the economy in the future (oil price shocks came out of the blue in the 1970s) or how the Fed will respond to them, we cannot say that monetary policy will necessarily continue to work well in the future.

S U M M A R Y

As the nation's central bank, the Federal Reserve bears primary responsibility for maintaining a low, stable rate of inflation and for maintaining full employment of the labor force as the economy is buffeted by a variety of shocks. The money demand curve, for example, may shift, causing a change in the interest rate, a shift in the *AD* curve, and a change in output and employment. The Fed can neutralize such money demand shocks by setting an interest rate target. To maintain the target, it increases the money supply whenever money demand increases, and decreases the money supply when money demand decreases. This policy enables the Fed to stabilize both inflation and unemployment.

Spending shocks—spontaneous shifts in aggregate expenditures—can also shift the *AD* curve, causing output to deviate from its full-employment level. The Fed can neutralize spending shocks by adjusting its interest rate target—changing the money supply to shift the *AD* curve back to its original position.

The Fed's most difficult problem is responding to supply shocks. A negative supply shock—an upward shift of the *AS* curve—presents the Fed with a dilemma. In the short run, it must choose a point along that new *AS* curve. If it wishes to maintain price stability, it must shift the *AD* curve to the left and accept higher unemployment. If the Fed wishes to maintain full employment, it must shift the *AD* curve to the right and accept a higher rate of inflation. A “hawk” policy puts

greater emphasis on price stability, while a “dove” policy emphasizes lower unemployment.

If Fed policy leads to ongoing inflation, then businesses and households come to expect the prevailing inflation rate to continue. As a result, the *AS* curve continues to shift at that built-in expected inflation rate. To maintain full employment, the Fed must shift the *AD* curve rightward, creating an inflation rate equal to the expected rate.

If the Fed wishes to change the built-in inflation rate, it must first change the expected inflation rate. For example, to lower the expected inflation rate, the Fed will slow down the rightward shifts of the *AD* curve. The actual inflation rate will fall, and expectations will eventually adjust downward. While they do so, however, the economy will experience a recession. The Fed's short-run choices between inflation and unemployment can be illustrated with the Phillips curve. In the short run, the Fed can move the economy along the downward-sloping Phillips curve by adjusting the rate at which the *AD* curve shifts. If the Fed moves the economy to a new point on the Phillips curve and holds it there, the built-in inflation rate will eventually adjust and the Phillips curve will shift. In the long run, the economy will return to the natural rate of unemployment with a different inflation rate. This is why we draw the long-run Phillips curve as a vertical line at the natural rate of unemployment.

K E Y T E R M S

natural rate of unemployment

passive monetary policy
active monetary policy

interest rate target
Phillips curve

long-run Phillips curve

R E V I E W Q U E S T I O N S

1. “The Fed should aim for the lowest possible unemployment rate.” True or false? Explain.
2. What effect does a change in the Fed's interest rate target have on financial markets? How do changes in expectations regarding the Fed's position on the interest rate target affect financial markets?
3. “The Fed should respond to any shift in the *AD* curve by maintaining its interest rate target.” True or false? Explain.
4. Explain the trade-off that the Fed faces with regard to negative supply shocks. What do “hawks” and “doves” have to do with this trade-off?
5. Why do expectations of inflation have a significant impact on the economy? What is the impact?
6. What relationship does the Phillips curve illustrate? How does the Fed control movements along the Phillips curve? Why is the long-run Phillips curve vertical?
7. List and explain three reasons why the Fed tolerates some ongoing inflation.

P R O B L E M S A N D E X E R C I S E S

1. Suppose that a law required the Fed to do everything possible to keep the inflation rate equal to zero. Using *AD* and *AS* curves, illustrate and explain how the Fed would deal with (a) a positive money demand shock, (b) a spending shock, and (c) an aggregate supply shock. What would the costs and benefits of such a law be?
2. Suppose that, in a world with *no* ongoing inflation, the government raises taxes. Using *AD* and *AS* curves, describe the effects on the economy if the Fed decides to practice a passive monetary policy. Alternatively, how could the Fed use active policy to neutralize the spending shock?
3. Suppose the economy has been experiencing a low inflation rate. A new chair of the Federal Reserve is named, and she is known to be highly sympathetic to dove policies. Explain the possible effects on the Phillips curve.
4. Using a graph, illustrate why the Fed, if it practices interest rate targeting, is concerned about the slope of the money demand curve. What are the implications of incorrectly estimating the slope?
5. Suppose that initially the price level is P_1 and GDP is Y_1 , with no built-in inflation. The Fed reacts to a negative spending shock by shifting the aggregate demand curve. The next time the Fed receives data on GDP and the price level, it finds that the price level is above P_1 and GDP is above Y_1 . Give two possible explanations for this finding.

C H A L L E N G E Q U E S T I O N S

1. Suppose the economy is experiencing ongoing inflation. The Fed wants to reduce expected inflation, so it *announces* that in the future it will tolerate less inflation. How does the Fed's credibility affect the success of the reduction? How can the Fed build its credibility? Are there costs to building credibility? If so, what are they?
2. This chapter mentioned what would happen if the Fed over- or underestimated the natural rate of unemployment. Using the *AD-AS* model, suppose the economy is at the true natural rate of unemployment, so that GDP is at its potential level. Suppose, too, that the Fed wrongly believes that the natural rate of unemployment is higher (potential GDP is lower), and acts to bring the economy back to its supposed potential. What will the Fed do? What will happen in the short run? If the Fed continues to maintain output below potential, what will happen over the long run?

E X P E R I E N T I A L E X E R C I S E

The Federal Reserve Bank of Cleveland's monthly publication *Economic Trends* is available online at <http://www.clev.frb.org/research>. Choose a recent issue and click



on "Monetary Policy." What are some current developments in U.S. monetary policy? See if you can illustrate them using the *AD-AS* model. A good source for the latest information regarding monetary policy is the Economy column that appears daily in *The Wall Street Journal*. Take a look at today's issue. What is the Fed's current policy stance? Is it focusing more on controlling inflation, or does it seem to be more concerned with the unemployment rate?

FISCAL POLICY: TAXES, SPENDING, AND THE FEDERAL BUDGET

Almost every year throughout the 1970s, 1980s, and early 1990s, a best-selling book would be published that predicted economic disaster for the United States and the world. In most of these books, the U.S. federal government played a central role. Arguments and statistics were offered to show that federal government spending—which was growing by leaps and bounds—was out of control, causing us to run budget deficits year after year. As a result, the United States was facing a growing debt burden that would soon swallow up all of our incomes, sink the United States economy, and bring about a worldwide depression.

During the late 1990s, as the federal budget picture improved, these disaster books quietly disappeared. In their place came news articles and public statements describing an economic future so bright that it would have been unimaginable just a few years earlier. And the situation faced by the federal government seemed to have flipped on its head. Instead of trying to bring down the ever-growing budget deficit, the key question became: What shall we do with our mounting budget *surpluses*?

What should we make of this flip-flop of public sentiment? Is it realistic? Were we really headed toward disaster until just a few years ago? And have all of our budget problems really been solved so suddenly? In this chapter, we'll take a close look at the government's role in the macroeconomy. You'll learn how to interpret trends in the government's budget, and how to identify the causes and effects of those trends. You'll see that while the United States did, indeed, face a growing budget problem over the last few decades—one that justified some concern—we were *not* on the brink of a disaster. You'll also see that while the U.S. fiscal picture has improved in the early 2000s, it is not as secure as is often suggested.

THINKING ABOUT SPENDING, TAXES, AND THE BUDGET

Let's start with some simple numbers. In 1959, the federal government's total spending—its outlays for goods and services, transfer payments, and interest on its debt—was \$81 billion. By 1999, the total had grown to \$1,806 billion, an increase of 2,130 percent. Government spending is out of control—right?

CHAPTER OUTLINE

Thinking About Spending, Taxes, and the Budget

Spending, Taxes, and the Budget: Some Background

- Government Spending
- Federal Tax Revenues
- The Federal Budget and the National Debt

The Effects of Fiscal Changes in the Short Run

- How Economic Fluctuations Affect Spending, Taxes, and the Federal Budget
- Countercyclical Fiscal Policy?

The Effects of Fiscal Changes in the Long Run

Were We Headed for a Debt Disaster?

Using the Theory: Understanding the New Budget Surpluses

- Measuring the Budget Surplus From Deficit to Surplus: Why?
- Future Surpluses: How Large?
- The Bright Budgetary Future: How Certain?

Or consider the national debt—the total amount that the government owes to the public from past borrowing in years in which it ran a budget deficit. In 1959, the national debt was \$235 billion; by the end of 1997—at its peak—it had grown to \$3,771 billion.¹ That amounted to about \$14,500 for every man, woman, and child in the United States—a sum that would have been very painful for each of us, individually, to repay. Wasn't this a crushing burden on the economy?

Actually, these figures are highly misleading. The first problem is that they are *nominal* values, and between 1959 and 1999, the price level rose. Even if the government had continued to spend the same amount or owe the same amount in *purchasing power* terms, the nominal figures would still have more than quintupled over the period. Thus, increases in nominal figures tell us very little.

But if we translate from nominal values into *real values*, we find much smaller increases: From 1959 to the late 1990s, *real* government spending and the *real* national debt each roughly tripled (compare this to the nominal values, which increased more than fifteenfold). Thus,

when examining budget-related figures over time, it is grossly misleading to use nominal figures, since the price level rises over time.

But even if we use real values to make our comparisons, we are still making a serious mistake. From 1959 to 1999, the U.S. population grew, the labor force grew, and the average worker became more productive. As a result, real GDP and real income tripled during this period. Why is that important? Because *spending and debt should be viewed in relation to income*.

We automatically recognize this principle when we think about an individual family or business. Suppose you are told that a family is spending \$50,000 each year on goods and services, and has a total debt—a combination of mortgage debt, car loans, student loans, and credit card balances—of \$200,000. Is this family acting responsibly? Or is its spending and borrowing out of control? That depends. If the income of the household is \$40,000 per year and is expected to remain roughly constant, there is serious trouble. This family would be spending more than it is earning, and its debt would grow each year until it could not handle the monthly interest payments.

But what if the family's income is \$800,000 per year? Then our conclusion would change dramatically: We'd wonder, why does this family spend so *little*? And if it owed \$200,000, we would not think it irresponsible at all. After all, the family could pay the interest on its debt—or even many times that interest—with a small fraction of its income.

What is true for an individual family is also true for the nation. Spending and debt are important only as *relative* concepts. As a country's total income grows, it will want more of the things that government can provide—education, high environmental standards, police protection, programs to help the needy, and more. Therefore, we expect government spending to rise as a nation becomes richer.

¹ There are many ways to measure the national debt. Some measures include amounts that the U.S. Treasury owes to other government agencies. But this part of the debt, since it is owed by one branch of government to another, could be canceled out at the stroke of a pen. Other measures include unfunded liabilities of U.S. government for Social Security, Medicare, and other benefits in future years. (Unfunded liabilities are the extent to which promises that the government has made to pay out benefits exceed expected revenue sources for those payments.) While unfunded liabilities are a concern for policy makers, they are not yet actual government debt. In this chapter, the national debt is defined as U.S. government bonds currently held outside of U.S. government agencies.

Moreover, as its income grows, a country can *handle* higher interest payments on its debt. Government spending and the total national debt, considered in isolation, tell us nothing about how responsibly or irresponsibly the government is behaving.

Budget-related figures such as government spending or the national debt should be considered relative to a nation's total income. This is why we should always look at these figures as percentages of GDP.

When we take this last step in adjusting our figures, we discover that both government spending and the national debt were not dramatically higher in the late 1990s than in the late 1950s. In 1959, government spending as a fraction of GDP was 16 percent. Over the next four decades, government spending fluctuated between 16 and 23 percent of GDP, but—by 1999—it had returned to 19 percent. Similarly, the national debt was 46 percent of GDP in 1959, and reached a peak of 50 percent in 1993. It ended the century at 40 percent of GDP—lower than in 1959. Thus, the story about federal spending and the federal debt is *not* the story suggested by the unadjusted figures.

In the rest of this chapter, as we explore recent trends in fiscal behavior and their effects on the economy, we'll do so with these lessons in mind. Accordingly, we'll look at fiscal variables as *percentages of GDP*.²

SPENDING, TAXES, AND THE BUDGET: SOME BACKGROUND

Our ultimate goal in this chapter is to understand how fiscal changes have affected, and continue to affect, the macroeconomy. But before we do this, some background will help. What has happened to the *composition* of government spending in recent decades? How does the U.S. tax system work, and what has happened to the government's tax revenues? Why has the national debt risen slowly in some periods, and more rapidly in other periods? And why—in recent years—has the national debt been falling? This section provides answers to these and other questions about the government's finances. Although state and local spending also play an important role in the macroeconomy, most of the significant macroeconomic changes in recent decades have involved the *federal* government. This is why we'll focus on spending, taxing, and borrowing at the federal level.

GOVERNMENT SPENDING

The federal government's *spending*—the total amount spent or disbursed by the federal government in all of its activities—can be divided into three categories:

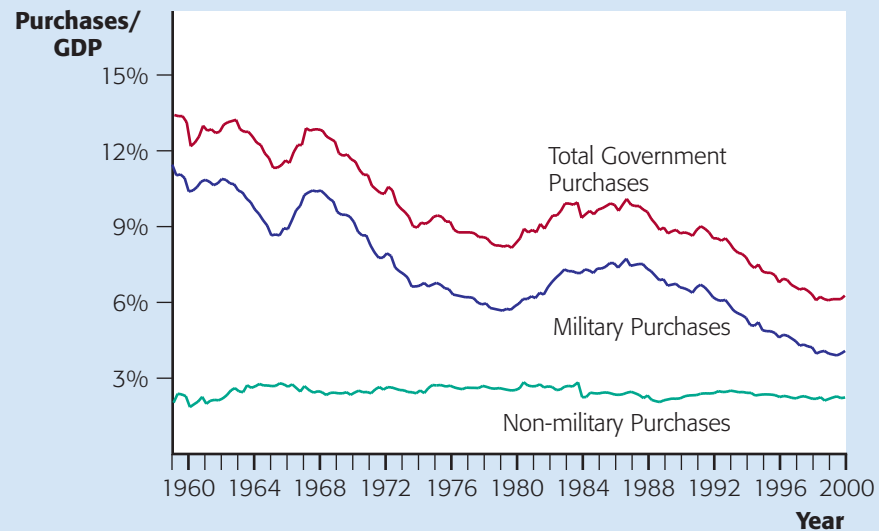
- *government purchases*—the total value of the goods and services that the government buys
- *transfer payments*—income supplements the government provides to people, such as Social Security benefits, unemployment compensation, and welfare payments
- *interest on the national debt*—the interest payments the government must make to those who hold government bonds

² It makes no difference whether we use nominal or real figures when dividing by GDP, as long as we're consistent. For example, we get the same fraction whether we divide nominal government spending by nominal GDP, or real government spending by real GDP.

FIGURE 1

The federal government's purchases have declined dramatically (relative to GDP) over the past 40 years. Non-military purchases have always been a stable, small percentage of GDP. Military purchases have declined, except for temporary buildups during the Vietnam War in the late 1960s and during the Reagan administration in the early 1980s.

FEDERAL GOVERNMENT PURCHASES AS A PERCENTAGE OF GDP



Government Purchases. Until the 1980s, government purchases of goods and services were the largest component of government spending. To understand how these purchases have changed over time, it's essential to divide them into two categories: military and non-military. Figure 1 shows total federal purchases, as well as federal military and non-military purchases, from 1959 to 1999.

One fact stands out from the figure: The federal government uses up only a tiny fraction of our national resources for non-military purposes. These non-military purchases include the salaries paid to all government workers outside the Defense Department (for example, federal judges, legislators, and the people who run federal agencies), as well as purchases of buildings, equipment, and supplies. Added together, all the different kinds of non-military government purchases account for a stable, low fraction of GDP—about 2 percent.

This strongly contradicts a commonly held notion: that government spending is growing by leaps and bounds because of bloated federal bureaucracies. Those who believe that government spending has become a growing concern must look somewhere besides non-military purchases for the reason.

As a percentage of GDP, non-military government purchases have remained very low and stable. They have not contributed to growth in total government spending.

What about military purchases? Here, we come to an even stronger conclusion:

As a percentage of GDP, military purchases have declined dramatically over the past several decades. Like non-military purchases, they have not contributed to any growth in government spending.

The decline in military purchases is shown by the middle line in Figure 1. They were around 11 percent of GDP in 1959 and fell almost continuously to about

MAJOR FEDERAL TRANSFER PROGRAMS, 1999

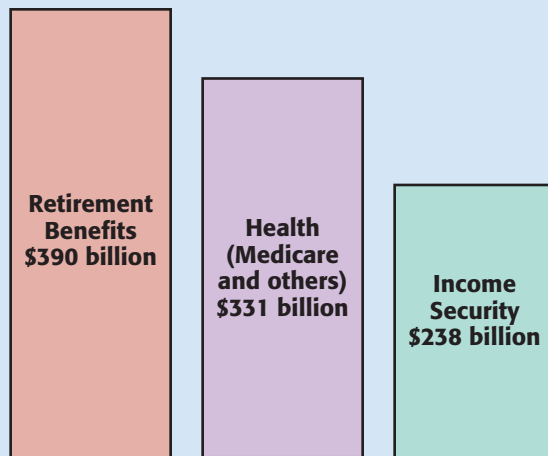


FIGURE 2

Approximately equal amounts of transfer payments were made for Social Security retirement benefits, health programs such as Medicare and Medicaid, and income security programs such as food stamps and welfare.

4 percent recently. Two buildups interrupted the decline, one associated with the Vietnam War in the late 1960s and the other during the Reagan administration in the 1980s. But both of these buildups were temporary.

The decline of military spending freed up resources amounting to 7 percent of GDP over the span shown in Figure 1. There are debates about whether U.S. defense spending can be cut even more, but given the current U.S. role in global politics, it is unlikely that any future cuts would be substantial. The implications are tremendously important for thinking about the recent past and the future of the federal government's role in the economy:

The decline in military spending in relation to GDP since the early 1960s has made huge amounts of resources available for other purposes. Because military spending is now only 4 percent of GDP and probably cannot drop much further, there cannot be any similar freeing up of resources in coming decades.

The resources released from military spending eased many otherwise tough decisions about resource allocation in the economy. In particular, they made it easy for the federal government to provide huge increases in resources to some parts of the population, through transfer payments.

Social Security and Other Transfers. Transfer programs provide cash and in-kind benefits to people whom the federal government designates as needing help. Figure 2 shows the three major categories of transfers. As you can see, they are roughly equal in size.

The largest category is retirement benefits—the payments made by the Social Security system to retired people. Although the benefits are loosely related to past contributions to the Social Security

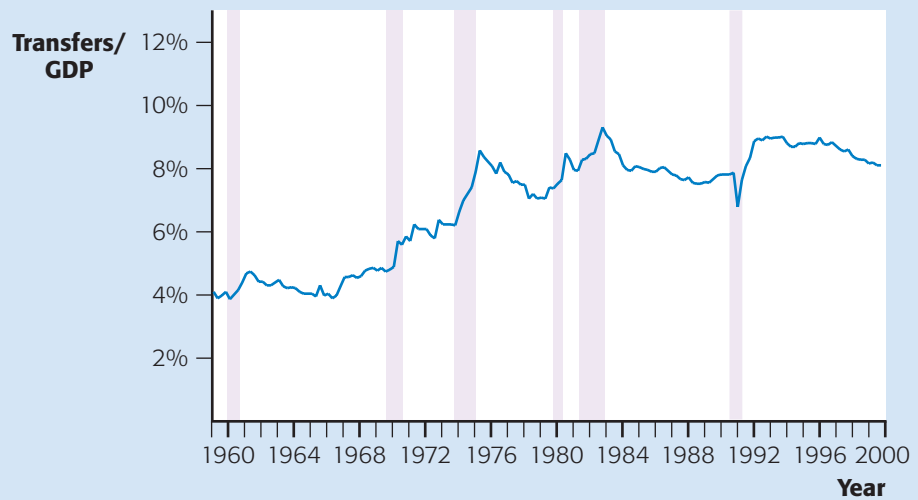


Don't confuse government spending with government purchases, which are just one component of the government's spending. The other components are transfer payments and interest on the debt.

FIGURE 3

Until about 1980, transfers grew rapidly as a fraction of GDP; thereafter, growth slowed. Transfers jumped upward during recessions (shaded), as in 1974, 1981, and 1991.

FEDERAL TRANSFER PAYMENTS AS A PERCENTAGE OF GDP



system, workers whose earnings are low receive benefits that are worth far more than their contributions. And after age 72, even someone with no history of contributions receives the minimum benefit.

The second-largest category of transfers occurs in health programs. The Social Security system provides health-related benefits to everyone aged 62 and over through Medicare. This is a health insurance plan in which people can go to any doctor they choose, as often as they want, and Medicare will pay 80 percent of the bills. Reform of Medicare to reduce its rapidly growing cost has been proposed, but little reduction in growth has been achieved so far. In addition to funding Medicare, the federal government helps finance state-operated health plans for the poor, through a program called Medicaid. The costs of these programs have been rising rapidly as well.

The third and smallest of the three categories of transfers is *income security*—programs to help poor families. Within this category, the largest component is the food stamp program, which gives coupons or special credit cards—good only for buying food—to qualified families. Welfare payments to poor families are also in this category, but these payments are much smaller than outlays on food stamps.

Have transfer payments been growing as a fraction of GDP? Indeed, they have. All three categories of transfer programs have grown rapidly in recent decades. And Figure 3 shows that total transfer payments as a percentage of GDP have trended upward as well.

In recent decades, transfers have been the fastest-growing part of federal government spending and are currently equal to about 8 percent of GDP.

Growth in transfers relative to GDP was most rapid in the 1970s during the Nixon administration. During this period, government-financed retirement benefits became much more generous, food stamps were introduced, and Medicare expanded. Since then, transfers have remained high, but they have not shown any long-term growth in relation to GDP.

FEDERAL GOVERNMENT INTEREST PAYMENTS AS A PERCENTAGE OF GDP

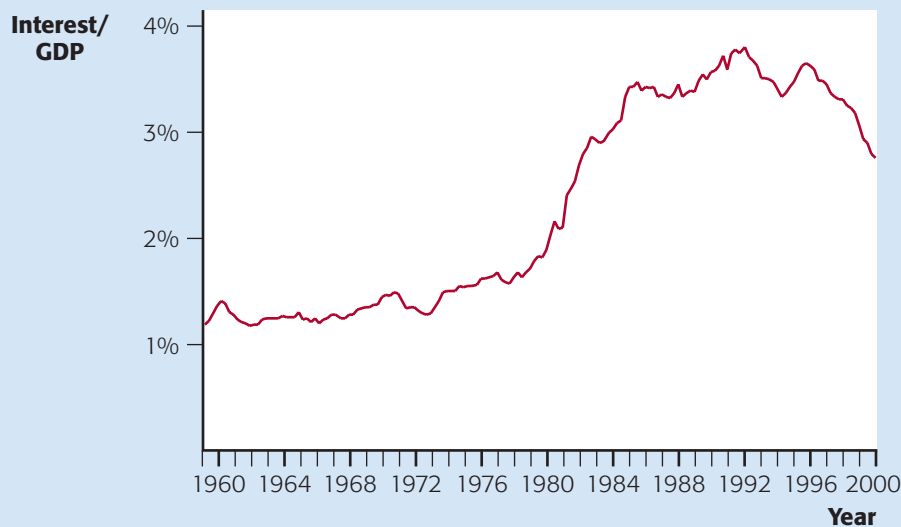


FIGURE 4

Interest payments on the national debt grew rapidly during the early 1980s.

Notice, however, that transfers are sensitive to the ups and downs of the economy. Transfers as a fraction of GDP rise during recessions, as in 1974, 1981, and 1991. This is for two reasons. First, the number of needy recipients rises in a recession, so transfer payments—the numerator of the fraction—increases. Second, GDP—the denominator—falls in a recession. Similarly, transfers as a fraction of GDP fall during expansions, such as our most recent, long expansion that began in 1991. During expansions, the numerator of this fraction falls (why?), and the denominator rises. We will come back to these movements in transfers toward the end of the chapter.

Interest on the National Debt. Figure 4 shows the behavior of the third and smallest category of government spending: interest on the national debt. As you can see, interest as a percentage of GDP grew rapidly in the early 1980s, when the debt was growing and interest rates were rising. We'll discuss the reasons for the rise in debt a bit later.

Total Government Spending. Figure 5 shows total spending in relation to GDP over the past several decades. There are two important things to notice in the figure. The first is the *fluctuations* in government spending over the period. There was a sharp increase in spending in each recession (shaded) due to the jump in transfers that we saw in Figure 3. The recession of 1981–82 is a striking example. Also visible is the increase in military spending for the Vietnam War in the late 1960s.

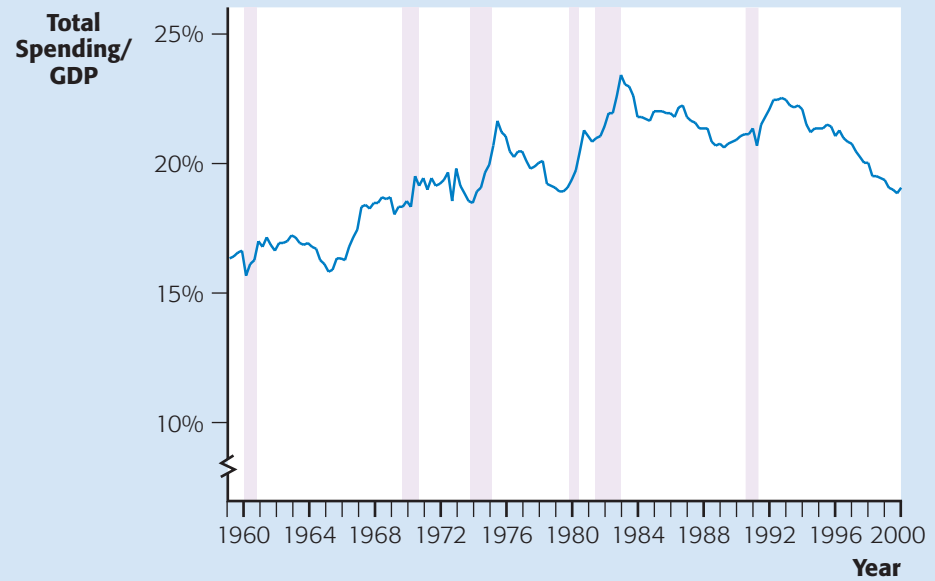
The second thing to notice is the *upward trend* of federal spending as a percentage of GDP that lasted until recently:

Over the past several decades, and until the early 1990s, federal government spending as a percentage of GDP rose steadily. The main causes of increases in transfer payments and increases in interest on the national debt that exceeded the decreases in military spending.

FIGURE 5

Federal spending tends to rise relative to GDP during wartime and during recessions; it falls during good times. Over the past several decades, the ratio of federal spending to GDP has trended upward.

TOTAL FEDERAL SPENDING AS A PERCENTAGE OF GDP



Finally, notice the important *downward* trend in the mid and late 1990s:

From 1992 to 1999, federal government spending as a percentage of GDP fell steadily, although it remained a higher percentage of GDP than in 1959. The main causes of the decline have been the continued sharp decreases in military spending, and more modest decreases in transfer payments relative to GDP.

The rise and recent fall in government spending relative to GDP have been important long-run trends. But in order to understand their impact on the macroeconomy, we must look at the other side of the budget: tax revenue.

FEDERAL TAX REVENUES

The federal government obtains most of its revenue from two sources: the personal income tax and the social security tax. Table 1 breaks down the revenue from these and the other less-important sources.

TABLE 1

SOURCES OF FEDERAL REVENUE, 1999

Source	Revenue (Billions of Dollars)
Personal income taxes	879
Corporate income taxes	185
Social Security taxes	612
Excise taxes	70
Other sources	81
Total	1,827

Source: *Economic Report of the President, 1999*, Table B-79.

TABLE 2

**THE 1999 PERSONAL
INCOME TAX FOR A
MARRIED COUPLE WITH
TWO CHILDREN**

Income	Tax	Average Tax Rate	Marginal Tax Rate
\$ 10,000	\$ 0	0%	0%
20,000	272	1.4	15
30,000	1,766	5.9	15
50,000	4,766	9.5	15
75,000	10,315	13.8	28
150,000	32,140	21.4	31
250,000	66,802	26.7	36
400,000	128,710	32.2	39.6

Source: Calculated from the 1999 Form 1040 tax table with the standard deduction of \$6,700.

The Personal Income Tax. The personal income tax is the most important source of revenue for the federal government and also the most conspicuous and painful. Almost every adult has to file Form 1040 or one of its shorter cousins. One of the signs of success as an American is seeing your federal tax return swell to the size of a magazine. Proposals to reduce both the amount of taxes people pay and the complexity of the tax forms are immensely popular.

The personal income tax is designed to be **progressive**—to tax those at the higher end of the income scale at higher rates than those at the lower end of the scale, and to excuse the poorest families from paying any tax at all. Table 2 shows how the income tax works, in theory, by computing the amount of tax a family of four should have paid in 1999 if it took the standard deduction.³ The table also shows the **average tax rate**—the fraction of total income a family pays in taxes—and the **marginal tax rate**—the tax rate paid on *each additional dollar* of income.

We can see from Table 2 that the income tax is designed to be quite progressive. In principle, a family in the middle of the income distribution, earning \$50,000 per year, should have paid 9.5 percent of its income in taxes, while a family at the top should have paid 32.2 percent of its income in taxes. The table also shows that marginal tax rates on families with the highest income are in the range of 28 to 40 percent.

But the tax system shown in the table does not reflect the ways that people can avoid tax. Many people have deductions far above the standard deduction. Some people earn income that they never report to the government, thereby evading taxes entirely. And people can shelter income in their employer's retirement plan or in a plan of their own. Studies have shown that higher-income households avoid more taxes than poorer families and that the tax system—while still progressive—is much less progressive than suggested by Table 2.

In addition to making the tax system less progressive, tax avoidance reduces the total tax revenues of the federal government. If we use Table 2—along with the incomes people actually earn—to estimate tax revenue, we'd predict that the government would collect between 15 and 20 percent of total personal income. But in reality, income tax revenues amount to only about 10 percent of total personal income.

Progressive tax A tax whose rate increases as income increases.

Average tax rate The fraction of a given income paid in taxes.

Marginal tax rate The fraction of an additional dollar of income paid in taxes.

³ The federal government allows households to deduct certain expenses (like medical care or the costs of moving to a new job) from their income before calculating the tax that they owe. Alternatively, they may deduct a standard amount (the *standard deduction*) from their income, regardless of their spending patterns.



The Congressional Budget Office maintains historical data on the U.S. federal budget. You can find it at <http://www.cbo.gov/showdoc.cfm?index=1821&sequence=0&from=7#1>.

The Social Security Tax. The Social Security tax applies to wage and salary income only. It was put in place in 1936, to finance the Social Security system created in that year. Whereas the personal income tax is a nightmare of complex forms and rules, the Social Security tax is remarkably simple. The current tax rate is a flat 15.3 percent,⁴ except for one complication: The tax is applied only on earnings below a certain amount (\$76,200 per year in 2000, although that salary cap rises each year).

The Social Security tax is actually the largest tax paid by many Americans, especially those with lower incomes. These families pay little or no income tax, but pay the Social Security tax on all of their wage earnings. For example, a family with \$30,000 of earnings in Table 2 would pay \$1,766 in federal income tax, but Social Security taxes on those earnings would be \$4,590.

Other Federal Taxes. Table 1 shows that the federal government also collects a little more than \$336 billion annually from other taxes. The most important of these is the *corporate profits tax*, which raises \$185 billion by taxing the profits earned by corporations at a rate of 35 percent.

The corporate profits tax is widely criticized by economists because of two important problems. First, it only applies to corporations. Thus, a business owner can avoid it completely by setting up a sole proprietorship or partnership instead of a corporation. As a result, the tax causes many businesses to forego the benefits of being corporations because of the extra tax they would have to pay.

Second, the corporation tax results in *double taxation* on the portion of corporate profits that corporations pay to their owners. This portion of profits is taxed once when the corporation is taxed and again when the profits are included as part of personal income. The corporation tax is thus a prime target for tax reform. Almost all reform proposals put forward by economists involve integrating the taxation of corporations into the tax system in a way that avoids these two distortions.

The federal government also taxes the consumption of certain products, such as gasoline, alcohol, tobacco, and air travel. These are called *excise taxes*. Excise taxes raise additional revenue for the government, but they are usually put in place for other, nonrevenue reasons as well. The excise tax on gasoline is seen, in part, as a fee on drivers for the use of federal highways. The taxes on alcohol and tobacco are intended to discourage consumption of these harmful products.

Trends in Federal Tax Revenue. The top line in Figure 6 shows total federal government revenue, as a percentage of GDP, from all of the taxes we've discussed. Over the 37 years shown in the figure,

federal revenue has trended upward from around 17 percent of GDP in 1959 to around 20 percent in 1999.

While the upward trend in total federal revenue as a fraction of GDP has been rather mild, its *composition* has changed dramatically. The lower two lines in Figure 6 show the part of federal revenue that comes from Social Security taxes and all other taxes. Notice the steady upward trend in Social Security tax revenue. Also notice that all other sources of revenue have remained roughly constant over the same period.

Why have Social Security taxes grown in importance? First, a little background. The Social Security system operates on a pay-as-you-go principle—it taxes people

⁴ If you look at your own paycheck, it may seem that the Social Security tax is only 7.65 percent instead of the 15.3 percent we've just mentioned. The reason is that your employer pays half the tax and you pay the other half. But the amount paid on your earnings is the sum, 15.3 percent.

FEDERAL GOVERNMENT REVENUE AS A PERCENTAGE OF GDP

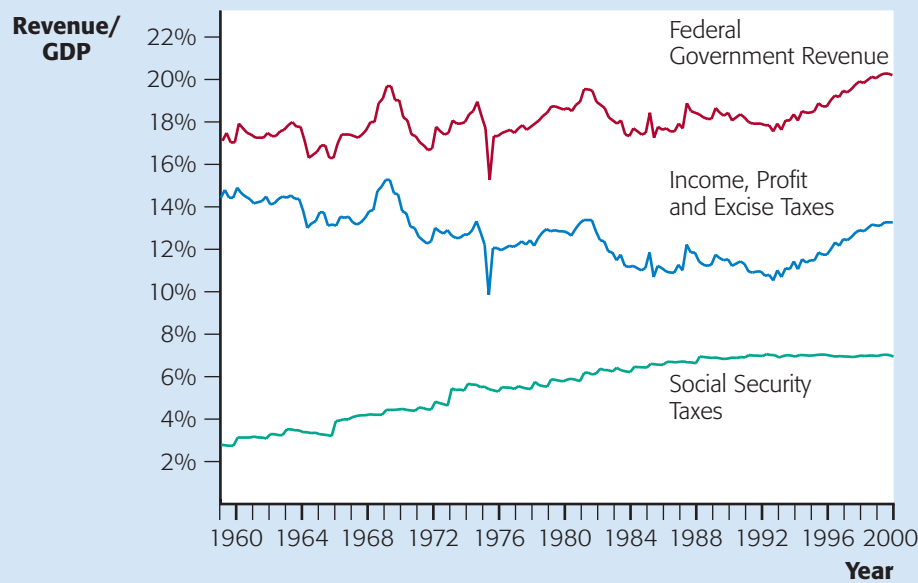


FIGURE 6

In recent decades, federal government revenue has trended upward to about 20 percent of GDP. This is because total revenue from the personal income tax, corporate profits tax, and excise taxes has remained stable relative to GDP, while Social Security tax revenue has increased.

who are working now in order to pay benefits to those who worked earlier and are now retired. But it also pays more benefits to those who have worked longer. Over the years, the system has benefited from a number of favorable circumstances. In its early decades, most retirees who received benefits had started working before the system began, so their benefits were small in relation to the earnings of those at work. Then, for the past several decades, the system benefited from two favorable demographic factors: first, a relatively small number of retirees (due to very low birth rates during the 1930s); and second, a large number of taxpayers (due to the baby boomers of the 1950s entering and remaining in the labor force).

But now, some demographic trends are working against the system. First, improved health is allowing people to spend a larger fraction of their lives in retirement. That is good from a human perspective, but from an accounting point of view, it means that the average retiree is drawing more benefits. At the same time, the baby boomers will soon begin retiring en masse, which means greater *numbers* of people drawing benefits. Finally, these increased benefits will be funded by a smaller number of working taxpayers. As a result of these trends, *the government has been raising Social Security tax rates* to keep the system solvent. Moreover, the government has been thinking about the future: It has increased the tax rate above the pay-as-you-go level in order to build up reserves, to cover higher expected benefit payouts.

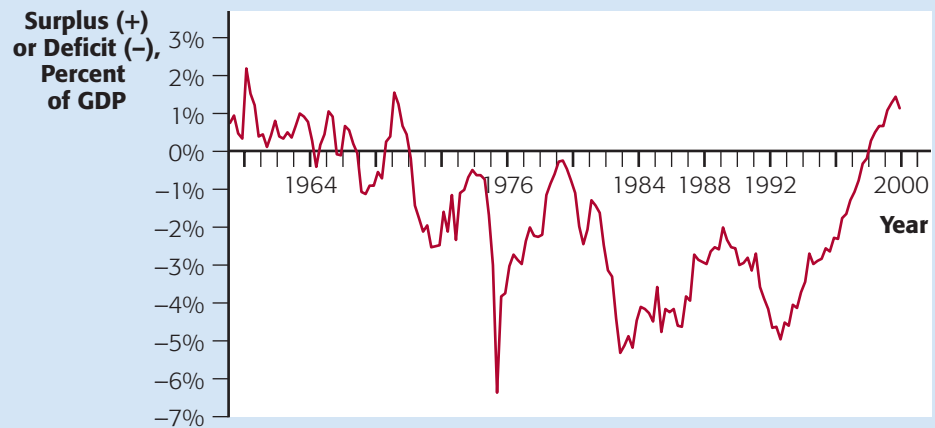
THE FEDERAL BUDGET AND THE NATIONAL DEBT

Finally, we can bring together what we've learned about the government's tax revenue (from the Social Security tax, personal income tax, corporate profits tax, and other sources) with what we've learned about the government's spending (on purchases, transfers, and net interest). And our first step is straightforward: When total tax revenue exceeds total government spending in any year, the government runs a budget surplus in that year. When the reverse occurs, and total government spending is greater than total tax revenue, the government runs a budget deficit.

FIGURE 7

Until about 1970, the federal government's budget deficit averaged around zero. Since then, and until very recently, there were deficits in most years—sometimes as high as 6 percent of GDP. The deficit increases (or the surplus decreases) during recessions as transfers rise and tax revenue falls.

THE FEDERAL BUDGET DEFICIT OR SURPLUS AS A PERCENTAGE OF GDP



Recent History of the Federal Budget. Figure 7 shows the history of the budget in recent decades. This line in the figure is actually just the difference between the federal spending line in Figure 5 and the federal revenue line in Figure 6 (the top line).⁵ The budget graph looks much choppier because the scale of the diagram is different here. But you can see that there was a dramatic change in the behavior of the budget around 1975. Until that year, the government mostly ran deficits, but rarely more than 2 percent of GDP. But from 1975 until 1993, the deficit grew significantly. During that period, it was usually greater than 3 percent of GDP, and often more than 4 percent. Notice, for example, the especially large rise in the deficit that occurred in the early 1980s. This was the combined result of a severe recession, which caused transfers to rise as shown in Figure 3, the buildup in military spending shown in Figure 1, and a large cut in income taxes during President Reagan's first term in office.

But then, in the mid-1990s, the deficit began to come down, and finally, in the late 1990s, the federal government began running budget surpluses for the first time in 30 years. Why did the budget shift from large deficits to surpluses during the 1990s? We'll answer this question in the "Using the Theory" section at the end of this chapter.

The National Debt. Before we consider the government's budget further, we need to address some common confusion among three related, but very different, terms: the federal *deficit*, the federal *surplus*, and the national *debt*. The federal deficit and surplus are *flow* variables—they measure the difference between government spending and tax revenue *over a given period*, usually a year. The national debt, by contrast, is a *stock* variable—it measures the total amount that the federal government owes *at a given point in time*. (See the second macroeconomics chapter if you need a refresher on stocks and flows.)

The relationship between these terms is this: Each year that the government runs a deficit, it must borrow funds to finance it, *adding to the national debt*. For

⁵ To measure the deficit or surplus, we have included all sources of revenue, and all types of federal spending, whether they are part of the official federal budget or not. In particular, Social Security taxes and social security payments are *officially* considered "off budget" in U.S. government statistics. In this chapter, however, we include the Social Security system in our budget calculations.

FEDERAL DEBT AS A PERCENTAGE OF GDP

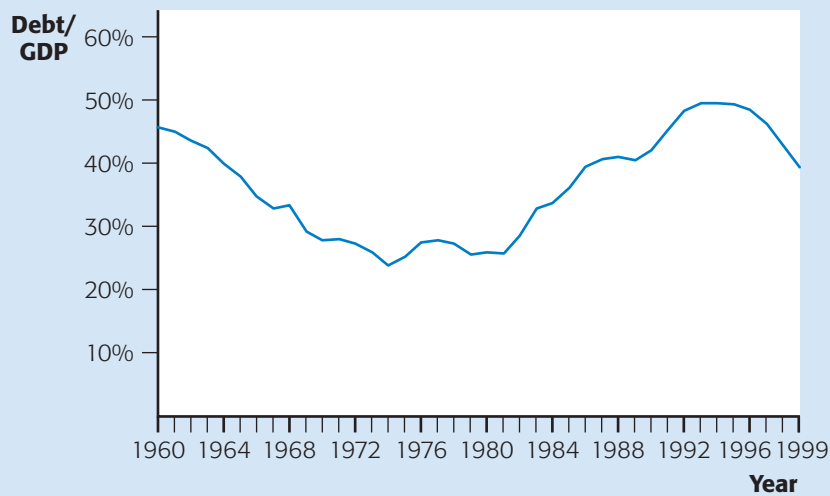


FIGURE 8

Until about 1980, the debt was shrinking as a fraction of GDP. For the next 15 years, it increased relative to GDP, until the ratio turned down again recently.

example, in 1996, the federal government ran a deficit of \$108 billion. During that year, it issued about \$108 billion in new government bonds, adding that much to the national debt. On the other hand, each year the government runs a surplus, it uses the surplus to *pay back* some of the national debt. For example, in 1999, the federal government ran a surplus of about \$125 billion. That year, it purchased about that much in government bonds it had issued in the past, thus reducing the national debt.⁶

We can measure the national debt as the total value of government bonds held by the public. Thus,

deficits—which add to the public’s holdings of government bonds—add to the national debt. Surpluses—which decrease the public’s bond holdings—subtract from the national debt.

Since the cumulative total of the government’s deficits has been greater than its surpluses, the national debt has grown over the past several decades. For most of this period, it has also grown relative to GDP, as shown in Figure 8.

The rise in the national debt also explains another trend we discussed earlier: the rise in *interest payments* the government must make to those who hold government bonds. The larger the national debt, the greater will be the government’s yearly interest payments on the debt. As you saw in Figure 4, total interest payments rose rapidly during the 1980s—the same period in which the national debt zoomed upward. In the 1990s, as the national debt decreased relative to GDP, so did interest payments on the debt.

Now that we’ve outlined the recent history of federal government spending, taxes, and debt, we can turn our attention to how fiscal changes affect the economy.

⁶ The increase or decrease in the national debt is never exactly the same as the annual deficit or surplus, because of accounting details.

What Happens When
Things Change?



THE EFFECTS OF FISCAL CHANGES IN THE SHORT RUN

In the short run, there is a two-way relationship between the government's budget and the macroeconomy. On the one hand, changes in the economy affect the government's spending and taxes; on the other hand, changes in spending and taxes affect the economy. Let's begin by considering how economic fluctuations affect the government's budget.

HOW ECONOMIC FLUCTUATIONS AFFECT SPENDING, TAXES, AND THE FEDERAL BUDGET

Economic fluctuations affect both transfer payments and tax revenues. In a recession, in which many people lose their jobs, the federal government contributes larger amounts to state-run unemployment insurance systems and pays more in transfers to the poor, since more families qualify for these types of assistance. Thus, a recession causes transfer payments to rise. Recessions also cause a drop in tax revenue, because household income and corporate profits—two important sources of tax revenue—decrease during recessions.

In a recession, because transfers rise and tax revenue falls, the federal budget deficit increases (or the surplus decreases).

An expansion has the opposite effects on the federal deficit: With lower unemployment and higher levels of output and income, federal transfers decrease and tax revenues increase. Thus,

in an expansion, because transfers decrease and tax revenue rises, the budget deficit decreases (or the surplus increases).

Because the business cycle has systematic effects on spending and revenue, economists find it useful to divide the deficit into two components. The **cyclical deficit** is the part that can be attributed to the current state of the economy. It turns positive (a cyclical deficit) when output is below potential GDP, and negative (a cyclical surplus) when output is above potential. When the economy is operating just at full employment, the cyclical deficit is, by definition, zero.

The **structural deficit** is the part of the deficit that is not caused by economic fluctuations. As the economy recovers from a recession, for example, the cyclical deficit goes away, but any structural deficit in the budget will remain.

Cyclical changes in the budget are not a cause for concern, because they average out to about zero, as output fluctuates above and below potential output. Thus, the cyclical deficit should not contribute to a long-run rise in the national debt.

Moreover, changes in the cyclical deficit are actually a good thing for the economy: They help to make economic fluctuations milder than they would otherwise be. Recall that spending shocks have a multiplier effect on output. The larger the multiplier, the greater will be the fluctuations in output caused by any given spending shock. But changes in the cyclical deficit make the multiplier *smaller*, and thus act as an *automatic stabilizer*. How?

Let's use unemployment insurance as an example. In normal times, with the unemployment rate at around, say, 4.5 percent or lower, federal transfers for unemployment insurance are modest. But when a negative spending shock hits the economy, and output and income begin to fall, the unemployment rate rises. Federal

Cyclical deficit The part of the federal budget deficit that varies with the business cycle.

Structural deficit The part of the federal budget deficit that is independent of the business cycle.

transfers for unemployment insurance rise *automatically*. Without assistance from the government, many of the newly unemployed would have to cut back their consumption spending substantially. But unemployment insurance cushions the blow for many such families, allowing them to make smaller cutbacks in consumption. As a result, the total decline in consumption is smaller, and GDP declines by less. Unemployment insurance thus reduces the multiplier.

Other transfer programs have a similar stabilizing effect on output. More people receive food stamps during recessions. Consequently, their consumption falls by less than it would if they did not have this help. And the tax system contributes to economic stability in a similar way. Income tax payments, for example, fall during a recession. With the government siphoning off a smaller amount of income from the household sector, the drop in consumption is smaller than it would be if tax revenues remained constant.

The same principle applies when a positive spending shock hits the economy. Transfer payments automatically decline, as the unemployed find jobs and fewer families qualify for government assistance. And tax revenues automatically rise, since income rises. As a result, the spending shock causes a smaller rise in GDP than would otherwise occur.

Many features of the federal tax and transfer systems act as automatic stabilizers. As the economy goes into a recession, these features help to reduce the decline in consumption spending, and they also cause the cyclical deficit to rise. As the economy goes into an expansion, these features help to reduce the rise in consumption spending, and they also cause the cyclical deficit to fall.

COUNTERCYCLICAL FISCAL POLICY?

In the previous section, you learned that changes in government spending and taxes that occur automatically during expansions and recessions help to stabilize the economy. This immediately raises a question: Can the government *purposely* change its spending or tax policy to make the economy even more stable? For example, suppose the *AD* curve shifts leftward, and the economy enters a recession. Perhaps the government could increase its purchases of goods and services, or cut income tax rates, thereby shifting the *AD* curve rightward again. If a government changes its spending or taxes in this way, specifically to prevent output from rising above or falling below its potential, it is engaging in **countercyclical fiscal policy**.

In the 1960s and early 1970s, many economists and government officials believed that countercyclical fiscal policy could be an effective tool to counteract the business cycle. Today, however, very few economists hold this position. Instead, they would put the Fed in charge of stabilizing the economy and reserve fiscal policy for addressing long-run issues of resource allocation. Indeed, the last clear use of countercyclical fiscal policy occurred in 1975, when the government gave tax rebates in the depths of a serious recession in order to stimulate consumption. (In Figure 7, you can see the especially large downward spike in tax revenue relative to GDP in that year.)

Why do economists recommend against using countercyclical fiscal policy, and why does Washington follow their advice? There are several reasons.

Timing Problems. It takes many months or even longer for a fiscal change to be enacted. Consider, for example, a decision to change taxes in the United States. A tax bill originates in the House of Representatives and then goes to the Senate,

Countercyclical fiscal policy

Changes in taxes or government spending designed to counteract economic fluctuations.



Before tax laws or tax rates can be changed, both the U.S. Senate and the U.S. House of Representatives must approve. This can cause long delays.

where it is usually modified. Then a conference committee irons out the differences between the House and Senate versions, and the tax bill goes back to each chamber for a vote. Once legislation is passed, the president must sign it. Even if all goes smoothly, this process can take many months.

But in most cases, it will *not* go smoothly: The inevitable political conflicts will cause further delays. First, there is the thorny question of distributing the cost of a tax hike, or the benefits of a tax cut, among different groups within the country. Each party may argue for changes in the tax bill in order to please its constituents. And some senators and representatives will see the bill as an opportunity to improve the tax system in more fundamental ways, causing further political debate.

All of these problems create the danger that the tax change will take effect long after it is needed. And changes in transfer payments or government purchases would suffer from similar delays. As a result, a fiscal stimulus might take effect after the economy has recovered from a recession and is beginning to overheat; or a fiscal contraction might take effect just as the economy is entering a recession. Fiscal changes would then be a *destabilizing* force in the economy—stepping on the gas when we should be hitting the brakes, or vice versa.

The Fed, by contrast, can increase or decrease the money supply *on the very day it decides that the change is necessary*. While there are time lags in the *effectiveness* of monetary policy (see the “Using the Theory” section in the previous chapter), the ability to execute the policy in short order gives monetary policy an important advantage over fiscal policy for stabilizing the economy.

Irreversibility. A second reason for favoring monetary rather than fiscal policy to stabilize the economy is the difficulty of reversing changes in government spending or taxes. Spending programs that create new government departments or expand existing ones tend to become permanent, or at least difficult to terminate. Many temporary tax changes become permanent as well—the public is never happy to see a tax cut reversed, and the government is often reluctant to reverse a tax hike that has provided additional revenue for government programs.

Reversing monetary policy, while not always painless, is easier to do. For one thing, the Fed makes its decisions secretly—neither government officials nor the public knows for sure what course the Fed has set until six weeks after the Federal Open Market Committee meets. Thus, the Fed is somewhat insulated from the political process in making its decisions. While there are limits to the Fed’s independence (Congress could change the Fed’s charter, or even eliminate the Fed entirely if it became too unhappy with its performance), these limits do not affect the Fed’s ability to act quickly when it sees the need.

The Fed’s Reaction. Even if the government attempted to stabilize the economy with fiscal policy, it could not do so very effectively, because—to put it simply—the Fed will not allow it. The Fed views a change in fiscal policy just as it views other spending shocks: as a shift in the *AD* curve that needs to be neutralized. For example, suppose the Fed believes that the *AD* curve is shifting leftward and the economy is entering a recession. Then the Fed will increase the money supply to shift the *AD* curve rightward by the amount it thinks necessary, long before any fiscal change takes effect. The fiscal change, when it is finally enacted, will simply be counteracted with an offsetting change in the money supply. As long as the Fed is free to set its own course, and as long as it continues to see its goal as stabilizing the economy at the natural rate of unemployment and low inflation, there is simply no opportunity—and no need—for countercyclical fiscal policy.

THE EFFECTS OF FISCAL CHANGES IN THE LONG RUN



What Happens When Things Change?

Because the Fed acts to neutralize them, fiscal changes have little short-run effect on the macroeconomy. But fiscal changes do have important long-run effects. And to analyze them, we use the model best suited for long-run analysis: the classical model.

We've already considered some of the long-run effects of fiscal policy in this book. In the chapter titled "The Classical, Long-Run Model," we discussed the long-run impact of changes in government purchases. In the chapter titled "Economic Growth and Rising Living Standards," we discussed how government tax and transfer policies can affect the incentives of workers and firms, in turn affecting the economy's long-run growth rate.

Here, we just reiterate the main conclusions of fiscal policy for the long run. First, we can summarize the impact of large and continuing budget deficits—such as occurred during the 1970s and 1980s—as follows:

- Large and continuing budget deficits cause the government to continually demand loanable funds, resulting in higher interest rates and lower investment spending than with a balanced budget.
- Lower investment spending causes the capital stock to grow more slowly. In this way, large budget deficits may contribute to slower growth in the average standard of living.
- Large and continuing budget deficits can harm living standards in another way: they cause the national debt—and annual interest payments on the national debt—to grow. Unless some other form of government spending decreases, tax rates will ultimately have to be raised to pay the higher interest. But higher tax rates, in turn, reduce incentives to work, to invest, and to save.

We can also summarize the impact of large and continuing budget surpluses—such as occurred in the later 1990s and early 2000s—as follows:

- Continual budget surpluses cause the government to *supply* loanable funds (to repay some of the national debt), resulting in lower interest rates and higher investment spending than with a balanced budget.
- Higher investment spending causes the capital stock to grow more rapidly. In this way, budget surpluses may contribute to faster growth in the average standard of living.
- Continuing budget surpluses can benefit living standards in another way: they cause the national debt—and annual interest payments on the national debt—to shrink. This drop in interest payments allows other components of government spending to rise, or else enables the government to lower tax rates. Lower tax rates, in turn, can increase incentives to work, to invest, and to save.

In the rest of this chapter, we will get more specific about the long-run effects of recent fiscal policies. But before you read on, this might be a good time to take out a pencil and paper, and do some active studying. See if you can use the graphs you've learned in this text to illustrate the impact of fiscal policy in the long



This section has argued that rising deficits come at the cost of lower investment spending and therefore reduce the rate of economic growth. But this is not *always* the case. It depends on what *causes* the deficit. In particular, suppose the deficit arises from an increase in government spending to improve the legal, financial, and physical infrastructure of the economy, or to improve education. All of these types of spending contribute to economic growth themselves. Thus, even if the deficits caused by this higher government spending crowd out private investment, their net effect on growth could be favorable.

run. In particular, see if you can use the graphs of the classical model—the loanable funds market, the labor market, and the production function—to illustrate each of the conclusions in the bulleted list on the previous page.

WERE WE HEADED FOR A DEBT DISASTER?

On a billboard in midtown Manhattan, a giant clock-like digital display tracks the U.S. national debt and how it changes each minute. Through the mid-1990s, as the publicly held debt soared beyond \$3 trillion and headed toward \$4 trillion, the clock showed the debt growing by about \$240,000 per minute. The last four digits on the display changed so rapidly that they appeared as a blur.

The national debt clock was one of several public relations campaigns that spread fear among the American public. How could we ever hope to repay all of this debt? Surely, we were speeding toward a debt disaster, right?

Actually, this was not quite right. True, many economists were *concerned* about budget deficits and growing debt—because of their effects on resource allocation and growth that we discussed in the previous section. But there is a big difference between deficits that are costly to society, and deficits that will bring us rapidly toward crises. In fact, most economists believed that—even when deficits were at their worst—we were *not* on the brink of a debt disaster, and only small budgetary adjustments were needed to change our course and avoid a disaster entirely.

Why?

First, it's important to realize that although we might *choose* to repay the national debt, we do not have to. *Ever*. Moreover, there is nothing automatically wrong with a national debt that *grows* every year. That may sound surprising. How could a government keep borrowing funds without every paying them back? Surely, no business could behave that way.

But actually, many successful businesses *do* behave that way, and continue to prosper. For example, the debt of many major corporations—like AT&T and General Motors—continues to grow, year after year. While they continue to pay interest on their debt, they have no plans to pay back the amount originally borrowed in the foreseeable future. As these companies' bonds become due, they simply *roll them over*—they issue new bonds to pay back the old ones.

Why don't these firms pay back their debt? Because they believe they have a better use for their funds: investing in new capital equipment and research and development to expand their businesses. This will lead to higher future profits. And as long as their profits continue to grow, they can continue to increase their debt.

Of course, this does not mean that *any* size debt would be prudent. Recall the important principle we discussed earlier in the chapter: *Debt and interest payments have meaning only in relation to income*. If a firm's income is growing by 5 percent each year, but its interest payments are growing by 10 percent per year, it would eventually find itself in trouble. Each year, its interest payments would take a larger and larger fraction of its income, and at some point interest payments would exceed total income. But even *before* this occurred, the firm would find itself in trouble. Lenders, anticipating the firm's eventual inability to pay interest, would cut the firm off. At that point, the firm would reach its *credit limit*—the maximum amount it can borrow based on lenders' willingness to lend. Since it could no longer roll over its existing debt with further borrowing, it would have to pay back any bonds coming due, until its debt was comfortably below its credit limit.

All of these observations apply to the federal government as well. As long as the nation's total income is rising, the government can safely take on more debt. More

specifically, if the nation's income is growing at least as fast as total interest payments, the debt can continue to grow indefinitely, without putting the government in danger.

The federal government *could* pay back the national debt—by running budget surpluses for many years. But if the government chooses *not* to pay back its debt, it would be acting just like corporations, which behave in similar fashion: It believes it has better uses for its revenue than debt repayment.

But how fast could the government continue to accumulate debt? Or, equivalently, how large could the federal deficit be without making the national debt a greater and greater burden for our citizens to bear?

Let's see. As long as total national income grows at least as fast as interest payments on the debt, the ratio of interest payments to income will not grow. In that case, we could continue to pay interest without increasing the average tax rate on U.S. citizens. Let's use some round numbers to make this clearer. Suppose that the nominal GDP is \$10 trillion and the national debt is \$5 trillion. Suppose, too, that interest payments average out to 10 percent of the national debt, or \$500 billion. Then the ratio of interest payments to nominal GDP would be \$500 billion/\$10 trillion = 0.05. Now suppose that, over some period of time, both nominal GDP and the national debt double, to \$20 trillion and \$10 trillion, respectively. Then interest payments would double as well, to \$1 trillion. But the ratio of interest payments to nominal GDP would remain constant, at \$1 trillion/\$20 trillion = 0.05.

More generally,

as long as the debt grows by the same percentage as nominal GDP, the ratios of debt to GDP and interest payments to GDP will remain constant. In this case, the government can continue to pay interest on its rising debt without increasing the average tax rate in the economy.

This establishes an important *minimal guideline for responsible government*: The debt should grow no faster than nominal GDP. Was the U.S. government within these guidelines when many commentators feared a debt disaster? Not quite . . . but almost. During the 1970s and 1980s, nominal GDP was growing at about 9 percent per year—about 3.2-percent growth in real GDP, plus a little less than 6-percent increase in the price level. So the debt could have grown by an average of about 9 percent per year without any rise in the average tax rate. In fact, over these two decades, the debt grew by an average of 11 percent per year—higher than the guideline. If the debt had continued to grow that much faster than GDP indefinitely, interest payments on the debt would have gradually taken a greater and greater share of our national income, requiring gradually higher tax rates or cuts in other government programs.

However, to prevent a long-term disaster, we didn't have to run surpluses. In fact, we didn't have to stop running deficits. Rather, we had to *decrease the growth rate of the debt* back to, or below, the growth rate of nominal GDP. At the time, this required that we shrink annual deficits by about one percent of GDP. And as we entered the 1990s, we accomplished this and more. True, concern about the mounting debt was instrumental in helping lawmakers shrink the deficit—by creating a political climate in which tax rates could be raised and the growth of government spending slowed down. But while there was certainly cause for concern, and that concern served us well, we were, in truth, far from a debt disaster.

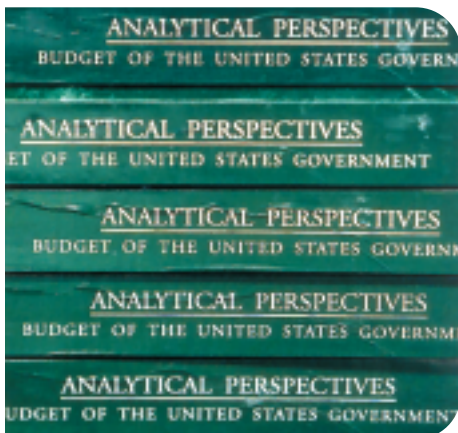
But what about the ratio of debt to GDP—which hit around 0.50 at its peak in 1993? Could the United States have been dangerously close to its credit limit—the amount of debt that would make lenders worry about the government's ability to continue paying interest? If so, we would indeed have been flirting with disaster—a tiny increase in the ratio would have led to a cutoff of further lending and required

the budget to be balanced immediately. It might also have caused a financial panic, if everyone tried to sell their U.S. government bonds at the same time, causing bond prices to fall and household wealth to plummet. This was a common scenario in the disaster books discussed at the beginning of the chapter. Were we facing this danger?

Not really. There is, indeed, some credit limit for the U.S. government, but we were probably far from it in 1993. At the conclusion of World War II, the ratio of federal debt to GDP was 1.08—more than twice as high as the recent peak. And at that time, there was little concern that the government would not honor its debt obligations, and in fact, the debt–GDP ratio was brought down dramatically. Thirty years after the end of the war, in 1975, the debt was down to about 23 percent of GDP. From this experience, we might guess that ratio of debt to GDP could exceed 1.0 before the federal government would reach its credit limit. And in recent decades, we did not even come close to this.

UNDERSTANDING THE NEW BUDGET SURPLUSES

Using the THEORY



Beginning in 1998, as the U.S. federal government ran its first budget surplus in 30 years, government officials, politicians, and the media began speaking of “surpluses as far as the eye can see.” The reason for the optimism was long-term predictions by government agencies that the 1998 surplus was not a one-time affair. Instead, both the President’s Office of Management and Budget and the non-partisan Congressional Budget Office (CBO) projected that the budget would continue to be in surplus—and that the surpluses would grow—at least through 2010, and most likely even further into the future. Over the next two years, each time the projections were updated, it seemed that the future surpluses were getting even larger. And in January 2000, newspaper headlines blared that, over the next 10 years, the government would generate a total of \$4.2 trillion in surpluses.

Then began the debate: what to *do* with this startlingly large sum. The choices were all pleasant. Some political leaders advocated *spending* the funds: to improve education, to repair our aging roads and bridges, or to build up our defenses against terrorist threats. Others wanted to set the surpluses aside and reserve them for future Social Security benefits, to ensure that the Social Security system would remain solvent forever. Still another option was to give a tax cut to U.S. households, increasing their incentives to work and save, and create an even wealthier economy. Or we could use the surpluses to pay back the national debt—all of it—so as to finally free taxpayers from the interest burden they had shouldered for so many years. Even comedians got into the act: In early 2000, Dennis Miller suggested that we could set aside the next 10 years’ worth of surpluses as prize money for a new TV show called *Who Wants to Be a Trillionaire?*

Whatever choice or combination of choices we would ultimately make, one thing seemed certain as we entered the 2000s: the U.S. economy—and the federal budget—was in for a pleasant ride. Was this a valid expectation?

In large part, yes. In early 2000, a close look at the budget predictions showed a mostly rosy future. But it also showed a highly *uncertain* future. In this section, we’ll explore the new budget surpluses in more detail. But first, let’s address an important issue surrounding the budget surplus: how it is measured.

MEASURING THE BUDGET SURPLUS

There are two ways to measure the budget surplus or deficit, and—confusingly enough—the media sometimes focuses on one, and sometimes on the other. For

years, the standard measure was the *on-budget surplus or deficit*. This is the difference between the government's total tax revenues and its total spending—with one important exception: *It excludes tax revenue and spending associated with the Social Security system.*⁷

Why have a budget measure that excludes Social Security?

Largely because the Social Security system was set up in the 1930s as a separate trust fund. True, the government was to *administer* the collection of Social Security taxes and the payment of benefits, but the system was thought of as separate from the government's other functions. Moreover, it was understood—from the beginning—that keeping the Social Security system solvent might require it to run deficits in some periods and surpluses in other periods. Government accountants felt that these temporary imbalances should not reflect on how the rest of the government was doing. In particular, a Social Security surplus or deficit should not influence our view of whether the government was living within its means, or beyond them. Thus, the on-budget surplus or deficit—which excludes the Social Security system—was the right measure for judging the fiscal behavior of the government.

To gauge the macroeconomic impact of the budget, however, we need to *include* the Social Security system. This is why macroeconomists prefer to look at the *unified budget surplus or deficit*—the difference between the government's total tax revenues (including Social Security taxes) and its total spending (including Social Security benefits).

Why is the unified budget a better measure of the macroeconomic impact of government? First, in the short run, we know that the government's budget affects the macroeconomy primarily through its effect on spending. It makes little difference whether a dollar in taxes is collected from households as income tax or Social Security tax: Either way, disposable income is reduced by the same dollar amount. Similarly, it makes little difference whether a dollar in transfer payments is paid out as welfare payments, educational assistance, Medicare, or Social Security benefits: It still puts a dollar of disposable income into a household's hands. Thus, when measuring the short-run impact of the government's fiscal policy on spending, it makes no sense to isolate Social Security from other government programs.

Second, the unified budget is the best measure to tell us about changes in the national debt over time, and changes in the interest burden of that debt. For example, if the unified budget is in deficit by \$10 billion in some year, then the government must borrow \$10 billion that year by issuing new bonds, adding \$10 billion to the national debt. Indeed, the debt grows by the same \$10 billion whether the government has to borrow for the Social Security system or any other reason: Borrowing is borrowing. Similarly, if the unified budget is in surplus by \$10 billion, the government will use the surplus to buy back \$10 billion of the government bonds it has issued in the past, thus reducing the debt by \$10 billion. Once again, it makes no



Several times in this text, you've read that whenever the government runs a budget surplus, it pays down the national debt by the amount of the surplus. And this is true. But in the media, you will often hear someone speak about "spending the surplus" or "using the surplus for tax cuts." This seems to contradict the view that surpluses automatically lead to debt reduction. But actually, the speakers are misusing the word "surplus." What they are really speaking about is different ways of using a *potential surplus*—funds that *would* lead to a surplus if not spent and not given back to taxpayers. Always remember that a surplus is the extent to which tax revenue exceeds government spending. If the government raises its spending or reduces taxes, then, by definition, it is reducing the surplus it will run that year, and reducing the funds it will use that year to pay back the national debt.

⁷ The U.S. postal service is also excluded from the official budget. However, the difference between the postal services revenue and its spending is so small that excluding it hardly makes a difference.

difference whether the surplus comes from Social Security or other parts of government: Paying back debt is paying back debt.

Throughout this chapter, whenever we have referred to the budget deficit or surplus, we've been using the *unified* budget or surplus. We'll continue to do so here.

FROM DEFICIT TO SURPLUS: WHY?

The new budget surpluses—those that we experienced in 1998 and 1999, as well as those projected through 2010—have arisen for two separate reasons. The first is the economic expansion that began 1991 and was still going strong in early 2000—the longest expansion on record. Part of the expansion is associated with a rapid rise in *potential GDP*. Technological changes—particularly the use of computers and, more recently, the Internet—have increased the productivity of the U.S. labor force much faster than previously, leading to more rapid growth in our capacity to produce goods and services. But there has also been a cyclical change: From 1991, the economy has moved from recession to expansion. With each passing year, the economy has operated closer and closer to its potential and—in the late 1990s—output may even have exceeded its potential level.

What has this expansion got to do with the budget? As you learned earlier in this chapter, in any expansion, transfers decrease as a fraction of GDP, and tax revenue grows as a fraction of GDP. The growth in tax revenue relative to GDP has been particularly strong in the most recent expansion for two reasons: (1) the rise in stock market values, which increased capital gains tax revenues—a part of the personal income tax; and (2) especially rapid income growth among high income taxpayers, who pay higher tax rates to begin with, and who were pushed into still higher marginal tax brackets as their incomes grew.

But changes in the economy explain only a part of the current and projected surpluses. A second and very important force has been caps on the government's *discretionary* spending—basically, all government spending except for interest on the debt, and transfer programs that are mandated by law (Social Security, Medicare, Medicaid, etc.). Thus, discretionary spending—which in 1999 amounted to \$575 billion or 6.3 percent of GDP—includes all spending on national defense, law enforcement, the environment, and the general operations of government. Congress first legislated *caps* on discretionary spending—and devised an effective system to enforce the caps—when it passed the Budget Enforcement Act of 1990. The Omnibus Budget Reconciliation Act of 1993 extended the caps through 1998, and the Balanced Budget Act of 1997 extended them again through 2002. These caps have been very effective in controlling growth in discretionary spending, and actually shrinking this component of spending as a fraction of GDP.

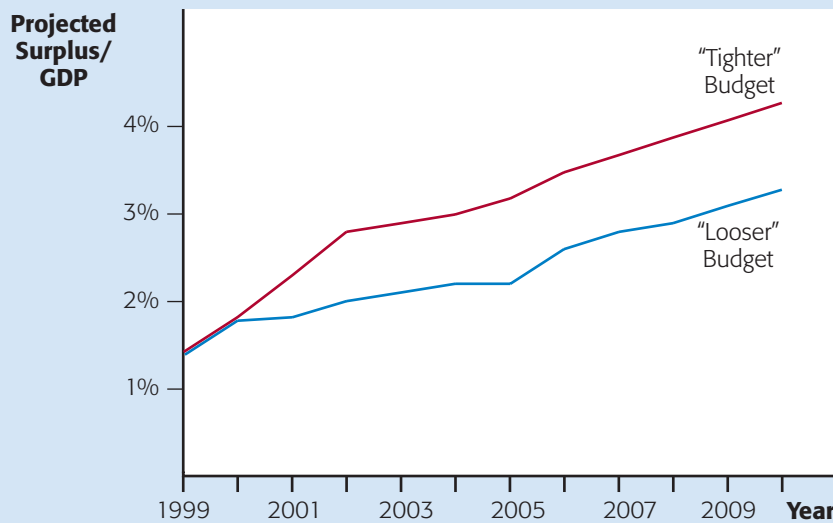
FUTURE SURPLUSES: HOW LARGE?

One of the startling things about the surpluses projected over the next 10 years is their size. For example, under the CBO's official projections, the cumulative surpluses through 2010 would be between \$3.2 and \$4.2 trillion. Either of these totals would be large enough to pay back the entire national debt in the hands of the public. (Whether we *do* pay back the national debt depends on whether we choose to actually *run* those surpluses, rather than cut taxes or increase spending instead. See the Dangerous Curves box earlier in this section.)

Figure 9 shows two projections by the CBO. Both are based on common assumptions about the behavior of the economy over the next 10 years (discussed below), but on two different assumptions about discretionary government spending. The upper line, labeled “tighter budget,” tracks the surplus as a percentage of GDP

CBO PROJECTED ANNUAL SURPLUS AS PERCENTAGE OF GDP

FIGURE 9



Source: Congressional Budget Office, Statement of Dan L. Crippen, Director, before the Committee on the Budget, U.S. House of Representatives, February 16, 2000; Congressional Budget Office, *The Budget and Economic Outlook: Fiscal Years 2001–2010*, January 2000, especially Chapter 5, “The Uncertainties of Budget Projections,” and Appendix C: “How the Economy Affects the Budget.”

assuming that discretionary government spending stays within the official budget caps until 2002, and thereafter grows at the rate of inflation. Under this scenario, the annual budget surplus would reach \$633 billion, or 4.3 percent of GDP in 2010. The lower line, labeled “looser budget,” shows what would happen if discretionary spending is freed from the cap and permitted to grow at the rate of inflation after 2000. Under this scenario, the budget surplus would not be as great—reaching about \$489 billion, or 3.3 percent of GDP in 2010. This difference—1 percent of GDP—might not seem like much, but it makes a big difference. Remember that that is the difference in just *one year’s* surplus. Looked at cumulatively, the tighter budget assumptions yield a total surplus through 2010 that is \$1 trillion greater.

THE BRIGHT BUDGETARY FUTURE: HOW CERTAIN?

How much faith can we have in the CBO’s projections of a bright budgetary future, ones in which we rack up huge potential surpluses year after year? The answer to this question is important. Suppose we decide to use up all or most of the potential surpluses by giving tax cuts, or setting up ambitious new government programs that raise discretionary spending above currently projected levels. And what if we are wrong about the future, and the potential surpluses never materialize. In that case, we will have inadvertently put the budget back into deficit, adding to our national debt. Further, it is difficult to cut government programs once they are put in place, and even more difficult to raise taxes after they have been cut. So a mistake that leads to deficits might not be corrected for years. The economy would then experience all of the effects of deficits and a growing national debt outlined earlier in this chapter (see “The Effects of Fiscal Changes in the Long Run”).

So how much faith can we have in the CBO’s projections? One thing we do know: The projections are *not* politically motivated. The CBO is a very well respected organization whose purpose is to give background information to members



Try your hand at managing the budget by using the National Budget Simulation (<http://socrates.berkeley.edu:3333/budget/budget.html>).

of Congress, not to take positions in political debates. Although some of the CBO's studies have been controversial, its research methods and conclusions are—on the whole—widely used and widely respected by both Democrats and Republicans.

However, just because the CBO's projections are honest does not mean they are reliable. Projections—especially those made over long periods—require assumptions about the economy and the budget that may or may not be realistic. And in early 2000, as politicians and pundits debated how to *use* the potential surpluses, the uncertain assumptions behind them were rarely discussed. Let's go through the assumptions behind the CBO's projections.

- *Discretionary spending:* All of the CBO's projections assume that discretionary spending will continue to fall as a fraction of GDP—from 6.3 percent in 1999 to either 4.7 percent (under the tighter budget scenario), or 5.3 percent (under the looser budget scenario). But both of these assumptions may be unrealistic. One reason that discretionary spending has been so easy to control recently will be hard to repeat: the decrease in defense spending that accompanied the end of the cold war. In fact, since 1990, almost all of the decrease in discretionary spending as a fraction of GDP (from 8.7 percent to 6.3 percent) has been due to the decrease in military spending. During this period, non-military discretionary spending has remained a constant fraction of GDP.
- *Mandatory transfer payments:* The CBO projections assume that this category of spending will rise, from 9.9 percent of GDP in 1999 to 10.9 percent of GDP in 2010. Most of the growth is attributed to Medicare and Medicaid. But these projections assume a slight *deceleration* of growth in these programs, compared to their rates of growth from 1962 to 1999. Whether this proves to be accurate depends on trends in health care costs, as well as the overall state of the economy. (Only the poor are eligible for Medicaid, so slower economic growth could mean higher than projected eligibility, and greater than projected Medicaid spending.)
- *Tax revenues:* The CBO expects total tax revenue to grow by roughly 4 percent annually from 2001 to 2004, then 4.5 percent per year from 2005 to 2010. But—as you've learned—the behavior of tax revenues depends on the behavior of the economy. Thus, the tax revenue assumptions are only as realistic as are the assumptions about the economy, which we'll address now.
- *Macroeconomic Assumptions:* The three assumptions just discussed refer to the budget, but the behavior of the budget depends on the behavior of the macroeconomy. Among the CBO's critical macroeconomic assumptions over the period 2000 to 2010 are the following:
 1. Real GDP growth of 2.8 percent per year
 2. Growth in the GDP price index of 1.6 percent per year
 3. Growth in the CPI of 2.5 percent per year
 4. Unemployment rising steadily from 4.0 percent to 5.2 percent (the CBO's very conservative estimate of the natural rate of unemployment)
 5. A roughly constant interest rate on ten-year treasury bonds of 5.7 percent (which was their average interest rate in 1999)

This is a long list of assumptions, and there are others not even listed here. But as you can imagine, we already have a picture that plenty could go wrong with. For example, suppose GDP grows just 0.1 percent slower each year than projected. Then transfers as a fraction of GDP would be higher than projected, and tax revenue lower than projected, leading to a \$46 billion shrinkage in the annual surplus by 2010—about half a percentage point of GDP.

Or, for another example, suppose that inflation heats up, because unemployment has fallen below its natural rate. (As you've learned, no one—not even the Fed—knows what the natural rate of unemployment is.) Then the Fed, to slow growth and bring the inflation rate down, would raise its interest rate target. But higher interest rates mean higher interest costs on the debt. And, according to CBO calculations, if the interest rate on 10-year Treasury bonds ends up at 8.7 percent—three percentage points higher than the CBO projects—that would wipe out about a fifth of the potential surplus.

As you can see, plenty can go wrong with the projected surpluses. So it seems natural to ask: How well have such projections done in the past? Unfortunately, the CBO has been making 10-year projections only since 1992, which is not enough time to gauge their accuracy. However, the CBO has been making shorter projections since 1986, and two things stand out about them: (1) There are significant deviations of actual budget numbers from projected numbers; and (2) these deviations grow worse as projections are made further out into the future. For example, since 1986, looking at all of the CBO's projections for just *one* year ahead, the average deviation of the actual deficit or surplus from the projected amount was 1.6 percent of GDP. Going *four* years out, the average deviation from actual was 2.4 percent of GDP. You might think that the deviations were caused by unexpected changes in legislation, like changes in tax laws or changes in government programs. But in fact, almost all of the deviations were caused by unexpected macroeconomic changes. No doubt, if the CBO had been making 10-year projections throughout this period, we would by now have discovered an average deviation substantially larger than 2.4 percentage points of GDP. After all, just a few years ago, the CBO was projecting *deficits* as far as the eye can see.

The CBO is fully aware of the possibility of error. So it has also come up with what it calls a “pessimistic projection”—a not unlikely scenario that could change the budget picture substantially. The pessimistic projection assumes that the pleasant changes in the economy from 1996–1999—faster growth and a deceleration in spending on health care—were temporary, and that the economy will return to more normal patterns from 2000 to 2010. Under this scenario, the surpluses disappear entirely in 2003, turning into growing deficits that approach 3 percent of GDP by 2010.

If the CBO's pessimistic scenario turns out to be accurate, then future spending hikes or tax cuts would have profound implications for the economy. Instead of spending future surpluses, we would be pushing the deficits even *beyond 3 percent* of GDP by 2010, and we'd once again have a rapidly rising national debt. This—as well as concerns over the future of social security—explain why many economists urged caution in early 2000.

S U M M A R Y

The U.S. federal government finances its spending through a combination of taxes and borrowing. When government spending exceeds tax revenue, the government runs a budget deficit. It finances that deficit by selling bonds, thereby adding to the national debt. When government spending is less than tax revenue, the government runs a budget surplus. It uses that surplus to buy back bonds it has issued in the past, thus shrinking the national debt.

Federal government spending consists of three broad categories: government purchases of goods and services, transfer

payments, and interest on the national debt. Non-military government purchases have traditionally accounted for a stable, low 2 percent of real GDP. Military purchases vary according to global politics; in recent years, they have declined dramatically relative to GDP. Transfer programs—such as Social Security, Medicare, and welfare—have been the fastest-growing part of government spending. They currently equal about 8 percent of GDP.

On the revenue side, the government relies on personal and corporate income taxes, Social Security taxes, and some

smaller excise taxes and user fees. In recent decades, federal revenue has been trending upward, and is currently about 20 percent of GDP.

From 1970 through the mid-1990s, federal spending exceeded federal revenues every year, so that the government ran budget deficits. Particularly large deficits occurred in the early 1980s. But in the 1990s, the deficit declined, and in 1998 the government began running yearly budget surpluses.

In the short run, there is a two-way relationship between government spending and taxes on the one hand, and the level of output on the other. First, changes in output affect government spending and taxes. In recessions, for example, government tax revenues fall and transfer payments rise. In this way, the tax and transfer system acts as an automatic stabilizer, helping to smooth out fluctuations.

Second, changes in government spending and taxes affect output. In principle, the government could use countercyclical

fiscal policy—changing taxes and spending in order to offset economic fluctuations. However, because of practical problems, countercyclical fiscal policy is seldom used.

In the long run, fiscal changes do have important effects. All else equal, we can expect larger budget deficits to slow growth in living standards, and smaller budget deficits or surpluses to speed the growth of living standards.

Over the 1970s and especially the 1980s, the average federal budget deficit was so large, and the national debt was growing so rapidly, that interest payments on the debt were rising relative to GDP. However, we were not on the brink of a debt disaster, and public concern helped to shrink deficits to more stable levels. By early 2000, the budget picture has turned upside down: Official projections showed growing surpluses through 2010, although these projections were fraught with uncertainty.

KEY TERMS

progressive tax
average tax rate

marginal tax rate
cyclical deficit

structural deficit

countercyclical fiscal policy

REVIEW QUESTIONS

- Why is it misleading to compare the national debt of \$235 billion in 1959 with the national debt of \$3,771 billion in 1999?
- List the three broad categories of federal government spending. According to the most recent data in the chapter, which is the largest category? Have any of the categories decreased relative to GDP over the past 8 years? If so, which ones?
- What is a *progressive* income tax?
- List the main sources of federal revenue. How and why has the composition changed recently?
- Explain the difference between the federal deficit and the national debt. Explain the *relationship* between the federal budget surplus and the national debt.
- Define the cyclical deficit and the structural deficit. Why are changes in the cyclical deficit not a major long-run concern?
- What is countercyclical fiscal policy? Is it an effective tool? Explain.
- “A decrease in the national debt as a fraction of GDP requires the federal government to run budget surpluses.” True or false? Explain.
- While the national debt has been an important concern, most economists don’t believe we were truly headed for disaster in the 1980s. Explain.
- “The United States can count on large budget surpluses for the next 10 years.” True or false? Explain.

PROBLEMS AND EXERCISES

- Use the following statistics, in billions of units, to calculate the real national debt and the debt relative to GDP in 1990 and 2000 for this hypothetical country. Which figures would you use to compare the national debt in the two years?

National Debt in 1990:	1.2
National Debt in 2000:	13.84
Nominal GDP in 1990:	101.7
Nominal GDP in 2000:	552.2
Price Index in 1990:	35.2
Price Index in 2000:	113.3

2. Suppose there is a country with 30 households divided into three categories (A, B, and C), with 10 households of each type. If a household earns 20,000 zips (the country's currency) or more in a year, it must pay 15 percent in tax to the government. If the household earns less than 20,000 zips, it doesn't pay any tax. When the economy is operating at full employment, household income is 250,000 zips per year for each type A household, 50,000 zips for type B households, and 20,000 zips for type C households.
- If the economy is operating at full employment, how much revenue does the government collect in taxes for the year?
 - Suppose a recession hits and household income falls for each type of household. Type A households now earn 150,000 zips, type B households earn 30,000 zips, and type C households earn 10,000 zips for the year. How much does the government collect in tax revenue for the year? Assume the government spends all of the revenue it *would* have collected if the economy had been operating at full employment. Under this assumption, what is the effect of the recession on the government budget deficit?
 - Suppose instead that the economy expanded and household incomes rose to 400,000 zips, 75,000 zips, and 30,000 zips, respectively, for the year. How much tax would the government collect for the year?

What is the effect on the government deficit (assume again that the government spends exactly the amount of revenue it collects when household income is at the values in part (a))?

What does this problem tell you about the relationship between shocks to the economy and the budget deficit?

3. According to the minimal guideline for responsible government outlined in the text, is either of the following two countries having a national debt crisis?

Country A
(Figures in Billions of \$)

	<i>Debt</i>	<i>GDP</i>
1999	1	100
2000	2	110
2001	3	150

Country B
(Figures in Billions of \$)

	<i>Debt</i>	<i>GDP</i>
1999	1236	1400
2000	1346	1550
2001	1406	1707

C H A L L E N G E Q U E S T I O N

Suppose the United States decides to dissipate potential future surpluses by either cutting taxes or increasing government spending.

- Compared to a policy of just accruing surpluses and paying down the national debt, what will this policy do to U.S. real GDP and interest rates in the *short run*? Illustrate your answer graphically. (*Hint*: Which macro model, and which graphs, should you use to illustrate effects on output and interest rates in the short run?)
- Compared to a policy of just accruing surpluses and paying down the national debt, what will this policy do to U.S. real GDP and interest rates in the *long run*? Illustrate your answer graphically.
- Going back to the short run, suppose the Fed responds by neutralizing the impact of the fiscal change in part (a) above. What will happen to real GDP and interest rates?

E X P E R I E N T I A L E X E R C I S E

Use a search engine such as google.com, yahoo.com, or goto.com to find data for: (a) the most recent year's growth rate of output; (b) the most recent month's inflation rate; (c) the most recent month's unemployment rate; and (d) the most recent trading day's interest rate on 10-year treasury bonds.



For each of these numbers, how does the reality compare with the CBO's early 2000 forecast (given in the "Using the Theory" section of this chapter)? For each number, state whether the deviation from the

forecast tends to make the budget surplus larger or smaller than the forecast. Do all of the deviations tend to influence the budget in the same way? Or do they push the budget in different directions?

Finally, find the most recent year's *actual* budget surplus or deficit as a fraction of GDP. What is the deviation from the CBO's early 2000 projection? Is the deviation what you'd expect from your analysis of deviations from the macro projection? If not, what might explain the inconsistency?

EXCHANGE RATES AND MACROECONOMIC POLICY

CHAPTER OUTLINE

Foreign Exchange Markets and Exchange Rates

Dollars per Pound or Pounds per Dollar?

The Demand for British Pounds

The Demand for Pounds Curve
Shifts in the Demand for Pounds Curve

The Supply of British Pounds

The Supply of Pounds Curve
Shifts in the Supply of Pounds Curve

The Equilibrium Exchange Rate

What Happens When Things Change?

How Exchange Rates Change Over Time
The Very Short Run: Hot Money
The Short Run: Macroeconomic Fluctuations
The Long Run: Purchasing Power Parity

Interdependent Markets: The Role of Arbitrage

Government Intervention and Foreign Exchange Markets

Managed Float
Fixed Exchange Rates
The Euro

Exchange Rates and the Macroeconomy

Exchange Rates and Spending Shocks
Exchange Rates and Monetary Policy

Using the Theory: The Stubborn U.S. Trade Deficit

If you've ever traveled to a foreign country, you were a direct participant in the **foreign exchange market**—a market in which one country's currency is traded for that of another. For example, if you traveled to Mexico, you might have stopped near the border to exchange some dollars for Mexican pesos.

Even if you have never traveled abroad, you've been involved, at least indirectly, in all kinds of foreign exchange dealings. For example, suppose you buy some Mexican-grown tomatoes at a store in the United States, where you pay with dollars. Except for shipping and retailing services, the resources used to produce those tomatoes were Mexican. A Mexican farmer grew the tomatoes; Mexican truckers transported them to the distribution center in the nearest large city; and Mexican workers, machinery, and raw materials were used to package them. All of these people want to be paid in Mexican pesos, regardless of who buys the final product. After all, they live in Mexico, so they need pesos to buy things there. But you, as an American, want to pay for your tomatoes with dollars.

Let's think about this for a moment. You want to pay for the tomatoes in dollars, but the Mexicans who produced them want to be paid in pesos. How can this happen?

The answer: *Someone*, here or abroad, must use the foreign exchange market to exchange dollars for pesos. For example, it might work like this: You pay dollars to your supermarket, which pays them to a U.S. importer, who sends a check in dollars to the distributor in Mexico, who—finally—turns the check over to a Mexican bank in exchange for pesos. That is how the Mexican distributor is able to pay the Mexican farmer in pesos. In this case, the actual changing of dollars into pesos takes place in a Mexican bank. But why is the Mexican bank willing to accept dollars for pesos? Because the bank is a participant in the market for foreign exchange.

In this chapter, we'll look at the markets in which dollars are exchanged for foreign currency. We'll also expand our macroeconomic analysis to consider the effects of changes in exchange rates. As you'll see, what happens in the foreign exchange market affects the economy, and changes in the economy affect the foreign exchange market. This has implications for the Fed as it tries to use monetary policy to steer the economy and keep it growing smoothly. Finally, in the "Using the The-

ory” section, you’ll see how the tools of the chapter can help us understand why the United States has such a large and persistent trade deficit.

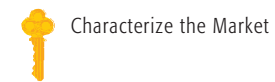
Foreign exchange market The market in which one country’s currency is traded for another country’s.

FOREIGN EXCHANGE MARKETS AND EXCHANGE RATES

Every day, all over the world, more than a hundred different national currencies are exchanged for one another in banks, hotels, stores, and kiosks in airports and train stations. Traders exchange dollars for Mexican pesos, Japanese yen, European euros, Indian rupees, Chinese yuan, and so on. In addition, traders exchange each of these foreign currencies for one another: pesos for euros, yen for yuan, euros for yen. . . . There are literally thousands of combinations. How can we hope to make sense of these markets—how they operate and how they affect us?

Our basic approach is to treat each pair of currencies as a separate market. That is, there is one market in which dollars are exchanged for euros, another in which Angolan kwanzas trade for yen, and so on. The physical locations where the trading takes place do not matter: Whether you exchange your dollars for yen in France, Germany, the United States, or even in Ecuador, you are a trader in the same dollar–yen market.

In any foreign exchange market, the rate at which one currency is traded for another is called the **exchange rate** between those two currencies. For example, if you happened to trade dollars for British pounds on March 7, 2000, each British pound would have cost you \$1.58. On that day, the exchange rate was \$1.58 per pound.



Characterize the Market

Exchange rate The amount of one country’s currency that is traded for one unit of another country’s currency.

DOLLARS PER POUND OR POUNDS PER DOLLAR?

Table 1 lists exchange rates between the dollar and various foreign currencies on a particular day in 2000. But notice that we can think of any exchange rate in two ways: as so many units of foreign currency per dollar, or so many dollars per unit of foreign currency. For example, the table shows the exchange rate between the British pound and the dollar as 0.6330 pounds per dollar, or 1.5798 dollars per pound. We can always obtain one form of the exchange rate from the other by taking its reciprocal: $1/0.6330 = 1.5798$, and $1/1.5798 = 0.6330$.

TABLE 1

FOREIGN EXCHANGE RATES,
MARCH 7, 2000

Country	Name of Currency	Symbol	Units of Foreign Currency per Dollar	Dollars per Unit of Foreign Currency
Brazil	real	R	1.7455	\$0.5729
China	yuan	Y	8.2784	0.1208
European Monetary Union Countries	euro	€	1.0422	0.9595
Great Britain	pound	£	0.6330	1.5798
India	rupee	₹	43.565	0.02295
Japan	yen	¥	105.62	0.009468
Mexico	peso	P	9.2800	0.1078
Russia	ruble	R	28.575	0.03500

In this chapter, we'll always define the exchange rate as “dollars per unit of foreign currency,” as in the last column of the table. That way, from the American point of view, the exchange rate is just another *price*. The same way you pay a certain number of dollars for a gallon of gasoline (the price of gas), so, too, you pay a certain number of dollars for a British pound (the price of pounds).

The exchange rate is the price of foreign currency in dollars.

Table 1 raises some important questions: Why, in early 2000, did a pound cost \$1.58? Why not \$1? Or \$5? Why did one Japanese yen cost a little less than a penny? And a Russian ruble about three cents?

The answers to these questions certainly affect Americans who travel abroad. Suppose you are staying in a hotel in London that costs 100 pounds per night. If the price of the pound is \$1, the hotel room will cost you \$100, but if the price is \$5, the room will cost you \$500. And exchange rates affect Americans who stay at home, too. They influence the prices of many goods we buy in the United States, they help determine which of our industries will expand and which will contract, and they affect the wages and salaries that we earn from our jobs.

How are all these exchange rates determined? In most cases, they are determined by the familiar forces of supply and demand. As in other markets, each foreign exchange market reaches an equilibrium at which the quantity of foreign exchange demanded is equal to the quantity supplied.

In the next several sections, we'll build a model of supply and demand for a representative foreign exchange market: the one in which U.S. dollars are exchanged for British pounds. Taking the American point of view, we'll call this simply “the market for pounds.” The other currency being traded—the dollar—will always be implicit.

THE DEMAND FOR BRITISH POUNDS

To analyze the demand for pounds, we start with a very basic question: *Who* is demanding them? The simple answer is, anyone who has dollars and wants to exchange them for pounds. But the most important buyers of pounds in the pound–dollar market will be American households and businesses. When Americans want to buy things from Britain, they will need to acquire pounds. To acquire them, they will need to offer U.S. dollars. To keep our analysis simple, we'll focus on just these American buyers. We'll also—for now—ignore any demand for pounds by the U.S. government.

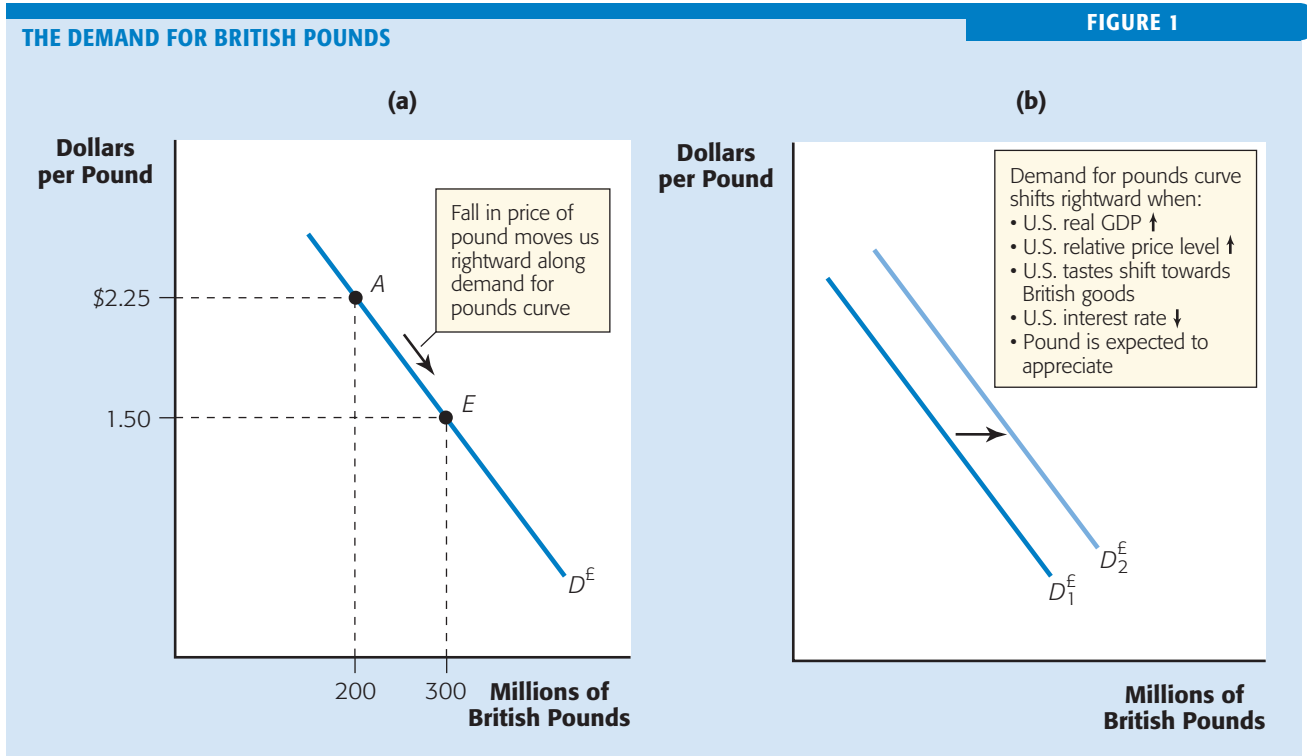
In our model of the market for pounds, we assume that American households and businesses are the only buyers.

Identify Goals and Constraints



Why do Americans want to buy pounds? There are two reasons:

- *To buy goods and services from British firms.* Americans buy sweaters knit in Edinburgh, airline tickets sold by Virgin Airways, and insurance services offered by Lloyd's. American tourists also stay in British hotels, use British taxis, and eat at British restaurants. To buy goods and services from British firms, Americans need to acquire pounds in order to pay for them.



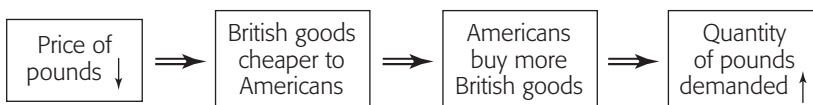
- *To buy British assets.* Americans buy British stocks, British corporate or government bonds, and British real estate. In each case, the British seller will want to be paid in pounds, so the American buyer will have to acquire them.

THE DEMAND FOR POUNDS CURVE

Panel (a) of Figure 1 shows an example of a **demand curve for foreign currency**, in this case, the demand curve for pounds. The curve tells us *the quantity of pounds Americans will want to buy in any given period, at each different exchange rate*. Notice that the curve slopes downward: The lower the exchange rate, the greater the quantity of pounds demanded. For example, at an exchange rate of \$2.25 per pound, Americans would want to purchase £200 million (point A). If the exchange rate fell to \$1.50 per pound, Americans would want to buy £300 million (point E).

Why does a lower exchange rate—a lower price for the pound—make Americans want to buy more of them? Because the lower the price of the pound, the less expensive British goods are to American buyers. Remember that Americans think of prices in dollar terms. A British compact disc that sells for £8 will cost an American \$18 at an exchange rate of \$2.25 per pound, but only \$12 if the exchange rate is \$1.50 per pound.

Thus, as we move rightward *along* the demand for pounds curve, as in the move from point A to point E:



Demand curve for foreign currency
A curve indicating the quantity of a specific foreign currency that Americans will want to buy, during a given period, at each different exchange rate.

SHIFTS IN THE DEMAND FOR POUNDS CURVE

In panel (a), you saw that a change in the exchange rate moves us *along* the demand for pounds curve. But other variables besides the exchange rate influence the demand for pounds. If any of these other variables changes, the entire curve will shift. As we consider each of these variables, keep in mind that we are assuming that only one of them changes at a time; we suppose the rest to remain constant.

U.S. Real GDP. Suppose real GDP and real income in the United States rise—say, because of continuing economic growth or a recovery from a recession. Then, Americans will buy more of everything, including goods and services from Britain. Thus, at any given exchange rate, Americans will demand more pounds. This is illustrated, in panel (b), as a rightward shift of the demand curve from D_1^{\pounds} to D_2^{\pounds} .

Relative Price Levels. Suppose that the U.S. price level rises by 8 percent, while that in Britain rises by 5 percent. Then U.S. prices will rise *relative* to British prices. Americans will shift from buying their own goods toward buying the relatively cheaper British goods, so their demand for pounds will rise. That is, the demand for pounds curve will shift rightward.

Americans' Tastes for British Goods. All else being equal, would you prefer to drive a General Motors Aurora or a Jaguar? Do you prefer British-made films, like *Mansfield Park* or *Hillary and Jackie*, or America's offerings, such as *American Beauty* or *Galaxy Quest*? These are matters of taste, and tastes can change. If Americans develop an increased taste for British cars, films, tea, or music, their demand for these goods will increase, and the demand for pounds curve will shift rightward.

Relative Interest Rates. Because financial assets must remain competitive in order to attract buyers, the rates of return on different financial assets—such as stocks and bonds—tend to rise and fall together. Thus, when one country's interest rate is high relative to that of another country, the first country's assets, *in general*, will have higher rates of return.

Now, suppose you're an American trying to decide whether to hold some of your wealth in British financial assets or in American financial assets. You will look very carefully at the rate of return you expect to earn in each country. All else being equal, a lower U.S. interest rate, relative to the British rate, will make British assets more attractive to you. Accordingly, as you and other Americans demand more British assets, you will need more pounds to buy them. The demand for pounds curve will shift rightward.

Expected Changes in the Exchange Rate. Once again, imagine you are an American deciding whether to buy an American or a British bond. Suppose British bonds pay 10 percent interest per year, while U.S. bonds pay 5 percent. All else equal, you would prefer the British bond, since it pays the higher rate of return. You would then exchange dollars for pounds at the going exchange rate and buy the bond.

But what if the price of the pound falls before the British bond becomes due? Then, when you cash in your British bond for pounds, and convert the pounds back into dollars, you'll be *selling your pounds at a lower price* than you bought them for. While you'd benefit from the higher interest rate on the British bond, you'd lose on the foreign currency transaction—buying pounds when their price is high, and selling them when their price is low. If the foreign currency loss is great

enough, you would be better off with U.S. bonds, even though they pay a lower interest rate.

As you can see, it is not just relative interest rates that matter to wealth holders; it is also *expected changes in the exchange rate*. An expectation that the price of the pound will fall will make British assets less appealing to Americans, since they will expect a foreign currency loss. In this case, the demand for pounds curve will shift leftward.

The opposite holds as well. If Americans expect the price of the pound to *rise*, they will expect a foreign currency *gain* from buying British assets. This will cause the *demand for pounds curve to shift rightward*.

THE SUPPLY OF BRITISH POUNDS

The demand for pounds is one side of the market for pounds. Now we turn our attention to the other side: the supply of pounds. And we'll begin with our basic question: *Who* is supplying them?

In the real world, pounds are supplied from many sources. Anyone who has pounds and wants to exchange them for dollars can come to the market and supply pounds. But the most important sellers of pounds are British households and businesses—who naturally have pounds and need dollars in order to make purchases from Americans. To keep our analysis simple, we'll focus on just these British sellers, and we'll ignore—for now—any pounds supplied by the British government:

In our model of the market for pounds, we assume that British households and firms are the only sellers.

The British supply pounds in the dollar–pound market for only one reason: because they want dollars. Thus, to ask why the British supply pounds is to ask why they want dollars. We can identify two separate reasons:

- *To buy goods and services from American firms.* The British buy airline tickets on United Airlines, computers made by IBM and Apple, and the rights to show films made in Hollywood. British tourists stay in American hotels and eat at American restaurants. The British demand dollars—and supply pounds—for all of these purchases.
- *To buy American assets.* The British buy American stocks, American corporate or government bonds, and American real estate. In each case, the American seller will want to be paid in dollars, and the British buyer will acquire dollars by offering pounds.

THE SUPPLY OF POUNDS CURVE

Panel (a) of Figure 2 shows an example of a **supply curve for foreign currency**—here, British pounds. The curve tells us *the quantity of pounds the British will want to sell in any given period, at each different exchange rate*. Notice that the curve slopes upward: The higher the exchange rate, the greater is the quantity of pounds supplied. For example, at an exchange rate of \$1.50 per pound, the British would want to supply £300 million (point *E*). If the exchange rate rose to \$2.25 per pound, they would supply £400 million (point *F*).

Why does a higher exchange rate—a higher price for the pound—make the British want to sell more of them? Because the higher the price for the pound, the



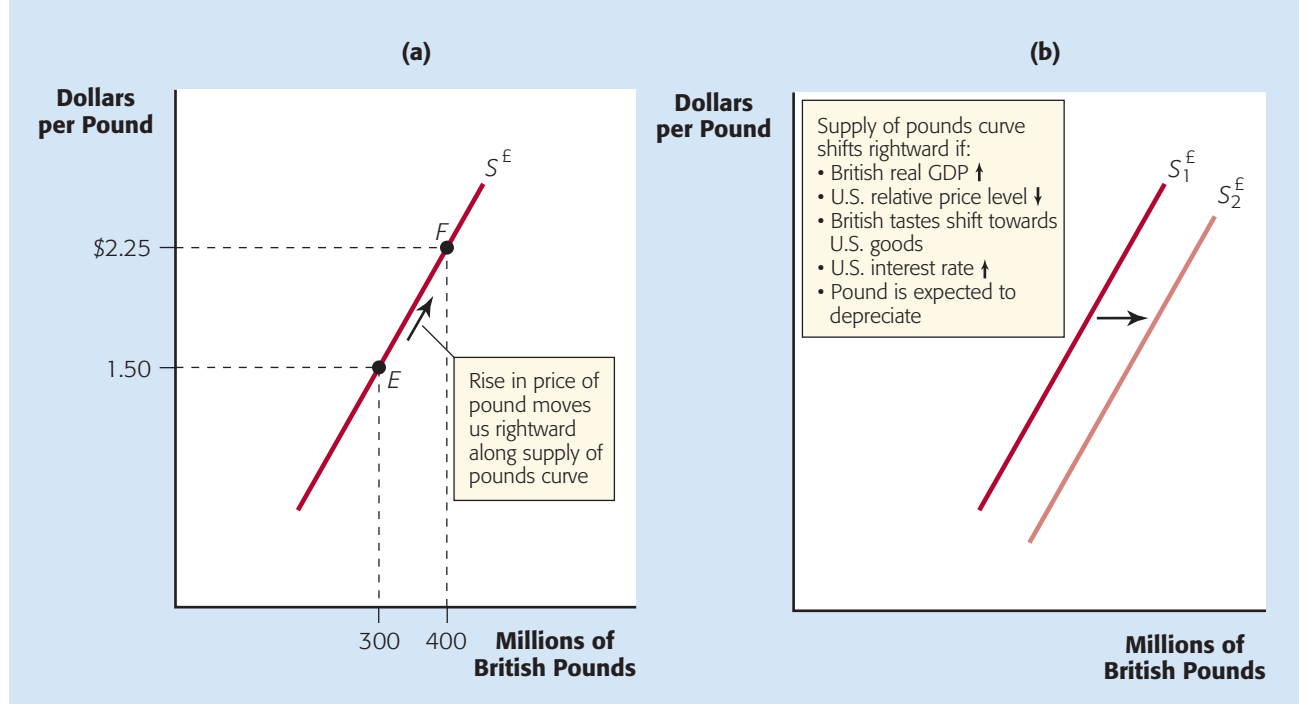
Identify Goals and Constraints

Supply curve for foreign currency

A curve indicating the quantity of a specific foreign currency that will be supplied, during a given period, at each different exchange rate.

FIGURE 2

THE SUPPLY OF BRITISH POUNDS



more dollars someone gets for each pound sold. This makes U.S. goods and services less expensive to British buyers, who will want to buy more of them—and who will therefore need more dollars.¹

To summarize, as we move rightward *along* the supply of pounds curve, such as the move from point *E* to point *F*:



SHIFTS IN THE SUPPLY OF POUNDS CURVE

When the exchange rate changes, we *move along* the supply curve for pounds, as in panel (a) of Figure 2. But other variables can affect the supply of pounds besides the exchange rate. When any of these variables change, the supply of pounds curve will shift, as shown in panel (b). What are these variables?

Real GDP in Britain. If real GDP and real income rise in Britain, British residents will buy more goods and services, including those produced in the United States.

¹ Actually, it is not a logical necessity for the supply of pounds curve to slope upward. Why not? When the price of the pound rises, it is true that the British will buy more U.S. goods and need more dollars to buy them. However, each dollar they buy costs *fewer pounds*. It might be that—even though the British obtain more dollars—they actually supply fewer pounds to get them at the higher exchange rate. In this case, the supply of pounds curve would slope downward. Economists believe, however, that a downward-sloping supply curve for foreign currency—while theoretically possible—is very rare.

Since they will need more dollars to buy U.S. goods, they will supply more pounds. In panel (b) this causes a rightward shift of the supply curve, from S_1^{\pounds} to S_2^{\pounds} .

Relative Price Levels. Earlier, you learned that a rise in the relative price level in the United States makes British goods more attractive to Americans. But it also makes *American* goods *less* attractive to the British. Since the British will want to buy fewer U.S. goods, they will want fewer dollars and will supply fewer pounds. Thus, a rise in the relative U.S. price level shifts the supply of pounds curve leftward.

British Tastes for U.S. Goods. Recall our earlier discussion about the effect of American tastes on the demand for pounds. The same reasoning applies to the effect of British tastes on the *supply* of pounds. The British could begin to crave things American—or recoil from them. A shift in British tastes toward American goods will shift the supply of pounds curve rightward. A shift in tastes *away* from American goods will shift the curve leftward.

Relative Interest Rates. You’ve already learned that a rise in the relative U.S. interest rate makes U.S. assets more attractive to Americans. It has exactly the same effect on the British. As the U.S. interest rate rises, and the British buy more U.S. assets, they will need more dollars and will supply more pounds. The supply of pounds curve will shift rightward.

Expected Change in the Exchange Rate. In deciding where to hold their assets, the British have the same concerns as Americans. They will look, in part, at rates of return; but they will *also* think about possible gains or losses on foreign currency transactions. Suppose the British *expect the price of the pound to fall*. Then, by holding U.S. assets, they can anticipate a foreign currency gain—selling pounds at a relatively high price and buying them back again when their price is relatively low. The prospect of foreign currency gain will make U.S. assets more attractive, and the British will buy more of them. *The supply of pounds curve will shift rightward.*

THE EQUILIBRIUM EXCHANGE RATE

Now we will make an important—and in most cases, realistic—assumption: that the exchange rate between the dollar and the pound *floats*. A **floating exchange rate** is one that is freely determined by the forces of supply and demand, without government intervention to change it or keep it from changing. Indeed, many of the world’s leading currencies, including the Japanese yen, the British pound, the 11-nation euro, and the Mexican peso, do float freely against the dollar most of the time.

In some cases, however, governments do not allow the exchange rate to float freely, but instead manipulate its value by intervening in the market, or even *fix* it at a particular value. We’ll discuss government intervention in foreign exchange markets later. In this section, we assume that both the British and U.S. governments leave the dollar–pound market alone.

When the exchange rate floats, the price will settle at the level where quantity supplied and quantity demanded are equal. Here, buyers and sellers are trading British pounds, and the price is the exchange rate—the *price of the pound*.

Look at panel (a) of Figure 3. The equilibrium in the market for pounds occurs at point *E*, where the supply and demand curves intersect. The equilibrium price is \$1.50 per pound. As you can verify, if the exchange rate were higher, say, \$2.25 per pound,

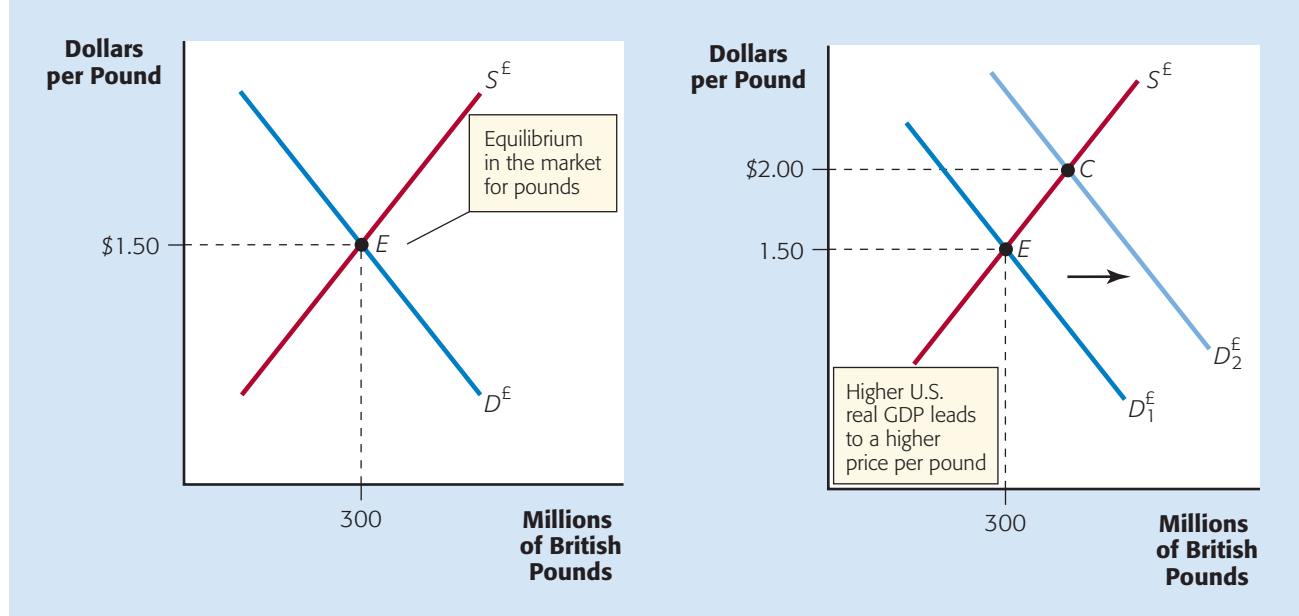


Find the Equilibrium

Floating exchange rate An exchange rate that is freely determined by the forces of supply and demand.

FIGURE 3

THE EQUILIBRIUM EXCHANGE RATE



there would be an *excess supply* of pounds, forcing the price of the pound back down to \$1.50. If the exchange rate were *lower* than the equilibrium price of \$1.50, there would be an *excess demand* for pounds, driving the price back up to \$1.50.

When the exchange rate floats—that is, when the government does not intervene in the foreign currency market—the equilibrium exchange rate is determined at the intersection of the demand curve and the supply curve.

What Happens When Things Change?



WHAT HAPPENS WHEN THINGS CHANGE?

What would cause the price of the pound to rise or fall? The simple answer to this question is, anything that shifts the demand for pounds curve, or the supply of pounds curve, or both curves together. Have another look at the right-hand panels of Figures 1 and 2. They summarize the major factors that can shift the demand and supply curves for pounds and therefore change the floating exchange rate.

Let's illustrate with a simple example. In panel (b) of Figure 3, the initial equilibrium in the market for pounds is at point E , with an exchange rate of \$1.50 per pound. Now suppose that real GDP rises in the United States. As you've learned (see Figure 1), this rise in U.S. GDP will shift the demand for pounds curve rightward, from $D_1^{\text{£}}$ to $D_2^{\text{£}}$ in the figure. At the old exchange rate of \$1.50 per pound, there would be an excess demand for pounds, which would drive the price of the pound higher. The new equilibrium—where the quantities of pounds supplied and demanded are equal—occurs at point C , and the new equilibrium exchange rate is \$2.00 per pound.

To recap, the increase in American GDP causes the price of the pound to rise from \$1.50 to \$2.00. When the price of any floating foreign currency rises because

THE EXCHANGE RATE IN THE POUND-DOLLAR MARKET

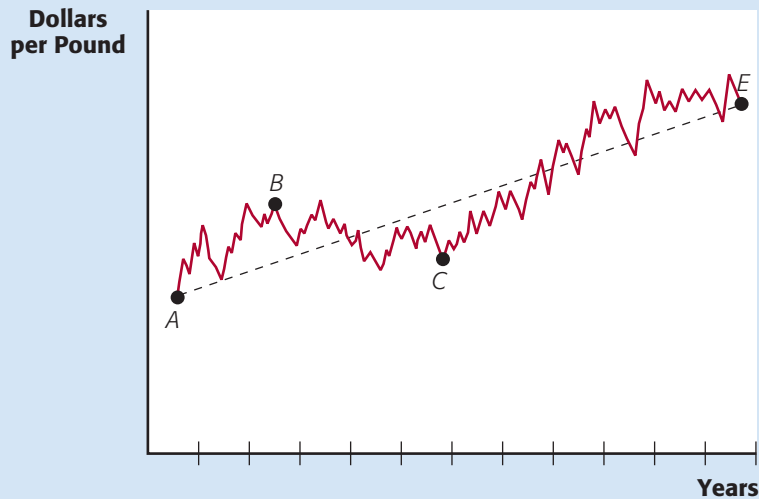


FIGURE 4

These hypothetical data show typical patterns of exchange rate fluctuations. Over the course of a few minutes, days, or weeks, the exchange rate can experience sharp up-and-down spikes. Over several months or a year or two, the exchange rate may rise or fall, as in the appreciation of the pound from points A to B and the depreciation from B to C. Over the long run, there may be a general upward or downward trend, like the appreciation of the pound illustrated by the dashed line connecting points A and E.

of a shift in the demand curve, the supply curve, or both, we call it an **appreciation** of the currency. In our example, the pound appreciates against the dollar. At the same time, there has been a **depreciation** of the dollar—a fall in its price in terms of pounds. (To see this, calculate the price of the dollar in terms of pounds before and after the shift in demand.)

When a floating exchange rates changes, one country's currency will appreciate (rise in price) and the other country's currency will depreciate (fall in price).

As you've learned, there are many other variables besides U.S. GDP that can change and affect the exchange rate. We could analyze each of these changes, using diagrams similar to panel (b) of Figure 3. However, we'll organize our discussion of exchange rate changes in a slightly different way.

HOW EXCHANGE RATES CHANGE OVER TIME

When we examine the actual behavior of exchange rates over time, we find three different kinds of movements. Look at Figure 4, which graphs the exchange rate in the pound-dollar market over time. The figure is based on hypothetical data, designed to make these three kinds of movement stand out more clearly than they usually do in practice.

Notice first the sharp up-and-down spikes. These fluctuations in exchange rates occur over the course of a few weeks, a few days, or even a few minutes—periods of time that we call the *very short run*.

Second, we see a gradual rise and fall of the exchange rate over the course of several months or a year or two. An example is the appreciation of the pound from point A to B and the depreciation of the pound from point B to C. These are *short-run* movements in the exchange rate.

Finally, notice that while the price of the pound fluctuates in the very short run and the short run, we can also discern a general *long-run* trend: The pound seems

Appreciation An increase in the price of a currency in a floating-rate system.

Depreciation A decrease in the price of a currency in a floating-rate system.

to be appreciating in the figure. This long-run trend is illustrated by the dashed line connecting points *A* and *E*.

In this section, we'll explore the causes of movements in the exchange rate over all three periods: the very short run, the short run, and the long run.

THE VERY SHORT RUN: "HOT MONEY"

Banks and other large financial institutions collectively have trillions of dollars worth of funds that they can move from one type of investment to another at very short notice. These funds are often called "hot money." If those who manage hot money perceive even a tiny advantage in moving funds to a different country's assets—say, because its interest rate is slightly higher—they will do so. Often, decisions to move billions of dollars are made in split seconds, by traders watching computer screens showing the latest data on exchange rates and interest rates around the world. Because these traders move such large volumes of funds, they have immediate effects on exchange rates.

Let's consider an example. Suppose that the relative interest rate in the United States suddenly rises. Then, as you've learned, U.S. assets will suddenly be more attractive to residents of both the United States and England, including managers of hot-money accounts in both countries. As these managers shift their funds from British to United States assets, they will be dumping billions of pounds on the foreign exchange market in order to acquire dollars to buy U.S. assets. This will cause a significant rightward shift of the supply of pounds curve.

In addition to affecting managers of hot-money accounts, the higher relative interest rate in the United States will affect ordinary investors. British investors will want to buy more American assets, helping to shift the supply of pounds curve further rightward. And American investors will want to buy fewer British assets than before, causing some decrease in the *demand* for pounds. Thus, in addition to the very large rightward shift in the supply of pounds, there will be a more moderate leftward shift in the demand for pounds.

Both of these shifts are illustrated in Figure 5: The supply of pounds curve shifts from S_1^{\pounds} to S_2^{\pounds} , and the demand for pounds curve shifts from D_1^{\pounds} to D_2^{\pounds} . The result is easy to see: The equilibrium in the market for pounds moves from point *E* to point *G*, and the price of the pound *falls* from \$1.50 to \$1.00. The pound depreciates and the dollar appreciates.

Expectations about future exchange rates can also trigger huge shifts of hot money, and Figure 5 also illustrates what would happen if American and British residents suddenly *expect* the pound to depreciate against the dollar. In this case, it would be the anticipation of foreign currency gains from holding U.S. assets, rather than a higher U.S. interest rate, that would cause the supply and demand curves to shift. As you can see in Figure 5, the expectation that the pound will depreciate actually *causes* the pound to depreciate—a self-fulfilling prophecy.

Sudden changes in relative interest rates, as well as sudden expectations of an appreciation or depreciation of a nation's currency, occur frequently in foreign exchange markets. They can cause massive shifts of hot money from the assets of one country to those of another in very short periods of time. For this reason,

relative interest rates and expectations of future exchange rates are the dominant forces moving exchange rates in the very short run.

HOT MONEY IN THE VERY SHORT RUN

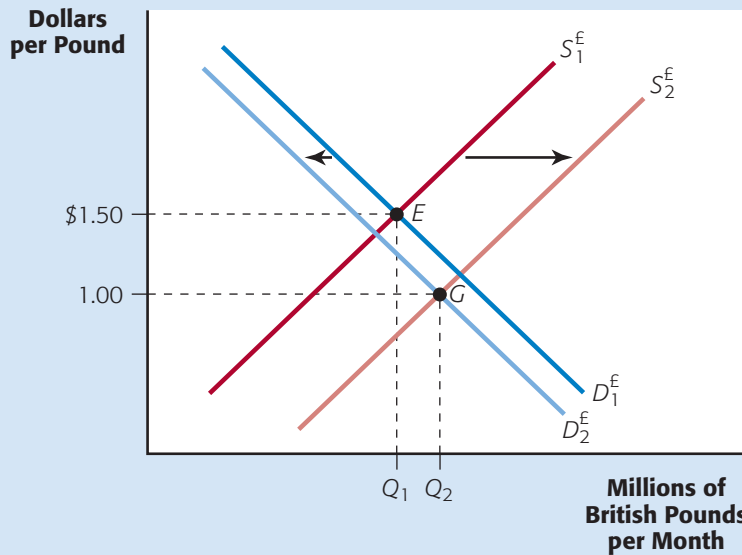


FIGURE 5

The market for pounds is initially in equilibrium at point E , with an exchange rate of \$1.50 per pound. A rise in the U.S. interest rate relative to the British rate will make U.S. assets more attractive to both Americans and Britons. Hot-money managers in both countries will shift funds from British to U.S. assets, causing a rightward shift of the supply of pounds curve. American investors will want to buy fewer British assets, causing a decrease in the demand for pounds. The net effect is a lower exchange rate—\$1.00 per pound at point G .

THE SHORT RUN: MACROECONOMIC FLUCTUATIONS

Look again at Figure 4. What explains the movements in the *short run* rate—the changes that occur over several months or a few years? In most cases, the causes are economic fluctuations taking place in one or more countries.

Suppose, for example, that both Britain and the United States are in a recession, and the U.S. economy begins to recover while the British slump continues. As real GDP rises in the United States, so does Americans' demand for foreign goods and services, including those from Britain. The demand for pounds curve will shift rightward, and—as shown in panel (a) of Figure 6—the pound will appreciate.

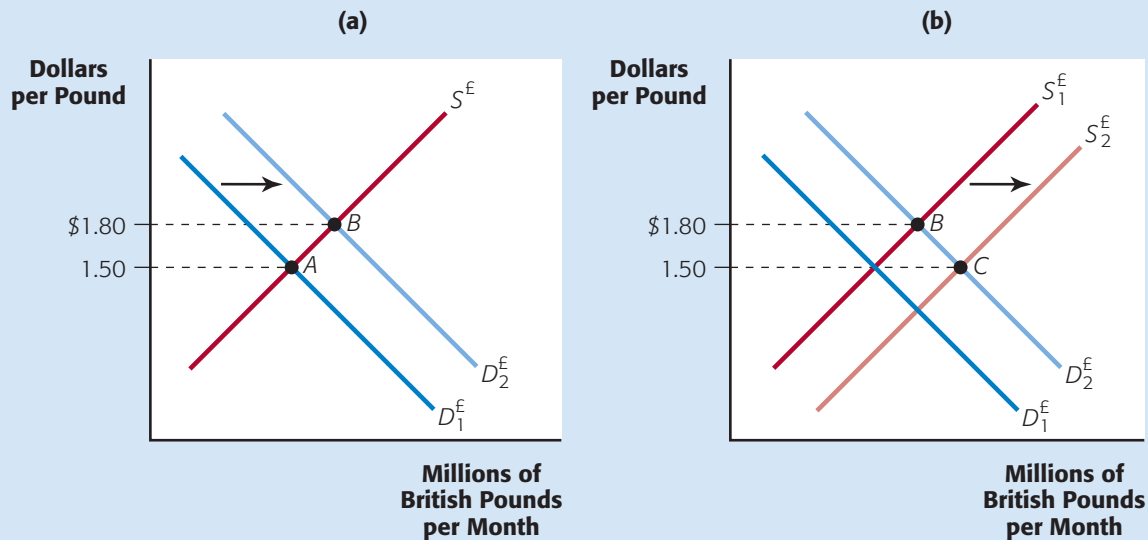
A year or so later, when Britain recovers from *its* recession, its real GDP will rise. British residents will begin to buy more U.S. goods and services, and supply more pounds so they can acquire more dollars. The supply of pounds curve will shift rightward, and—as shown in panel (b) of Figure 6—the pound will depreciate. Thus,

in the short run, movements in exchange rates are caused largely by economic fluctuations. All else equal, a country whose GDP rises relatively rapidly will experience a depreciation of its currency. A country whose GDP falls more rapidly will experience an appreciation of its currency.

This observation contradicts a commonly held myth: that a strong (appreciating) currency is a sign of economic health, and a weak (depreciating) currency denotes a sick economy. The truth may easily be the opposite. Over the course of several quarters or a few years, the dollar could appreciate because the U.S. economy is *weakening*—entering a serious recession. This would cause Americans to cut back spending on domestic *and* foreign goods, and decrease the demand for foreign currency. Similarly, a *strengthening* U.S. economy—in which Americans are earning

FIGURE 6

EXCHANGE RATES IN THE SHORT RUN



Panel (a) shows a situation in which the United States recovers from a recession first. U.S. demand for foreign goods and services increases, shifting the demand for pounds curve to the right. The result is an appreciation of the pound. Panel (b) shows Britain's subsequent recovery from its recession. As the British begin to buy more U.S. goods and services, the supply of pounds curve shifts rightward, causing the pound to depreciate.

and spending more—would increase the U.S. demand for foreign currency and—all else equal—cause the dollar to depreciate.

Keep in mind, though, that other variables can change over the business cycle besides real GDP, including interest rates and price levels in the two countries. For example, a recession can be caused by a monetary contraction that raises the relative interest rate in a country. Or a monetary stimulus in the midst of a recession could result in a relatively low interest rate. These changes, too, will influence exchange rates over the business cycle.

THE LONG RUN: PURCHASING POWER PARITY

In mid-1992, you could buy about 100 Russian rubles for one dollar. In mid-1998, that same dollar would get you more than 6,000 rubles—so many that the Russian government that year created a new ruble that was worth 1,000 of the old rubles. (The ruble exchange rate in Table 1 is for the new ruble.) What caused the ruble to depreciate so much against the dollar during those six years?

This is a question about exchange rates over many years—the long run. Movements of hot money—which explain sudden, temporary movements of exchange rates—cannot explain this kind of long-run trend. Nor can business cycles, which are, by nature, temporary. What, then, causes exchange rates to change over the long run?

In general, long-run trends in exchange rates are determined by *relative price levels* in two countries. We can be even more specific:

According to the purchasing power parity (PPP) theory, the exchange rate between two countries will adjust in the long run until the average price of goods is roughly the same in both countries.

To see why the PPP theory makes sense, imagine a basket of goods that costs \$750 in the United States and £500 in Britain. If the prices of the goods themselves do not change, then, according to the PPP theory, the exchange rate will adjust to $\$750/\pounds 500 = \1.5 dollars per pound. Why? Because at this exchange rate, \$750 can be exchanged for £500, so the price of the basket is the same to residents of either country—\$750 for Americans, and £500 for the British.

Now, suppose the exchange rate was *below* its PPP rate of \$1.50 per pound—say, \$1 per pound. Then a trader could take \$500 to the bank, exchange it for £500, buy the basket of goods in Great Britain, and sell it in the United States for \$750. She would earn a profit of \$250 on each basket of goods traded. In the process, however, traders would be increasing the demand for pounds and raising the exchange rate. When the price of the pound reached \$1.50, purchasing power parity would hold, and special trading opportunities would be gone. As you can see, trading activity will tend to drive the exchange rate toward the PPP rate. (An end-of-chapter review question asks you to explain the adjustment process when the exchange rate starts *higher* than the PPP rate.)

The PPP theory has an important implication:

In the long run, the currency of a country with a higher inflation rate will depreciate against the currency of a country whose inflation rate is lower.

Why? Because in the country with the higher inflation rate, the relative price level will be rising. As that country's basket of goods becomes relatively more expensive, only a depreciation of its currency can restore purchasing power parity. And traders—taking advantage of opportunities like those just described—would cause the currency to depreciate.

Purchasing Power Parity: Some Important Caveats. While purchasing power parity is a good general guideline for predicting long-run trends in exchange rates, it does not work perfectly. For a variety of reasons, exchange rates can deviate from their PPP values for many years.

First, some goods—by their very nature—are difficult to trade. Suppose a haircut costs £5 in London and \$30 in New York, and the exchange rate is \$1.50 per pound. Then British haircuts are cheaper for residents of both countries. Could traders take advantage of this? Not really. They cannot take \$30 to the bank in exchange for £20, buy four haircuts in London, ship them to New York, and sell them for a total of \$120 there. Haircuts and most other personal services are nontradable.

Second, high transportation costs can reduce trading possibilities even for goods that *can* be traded. Our earlier numerical example would have quite a different ending if moving the basket of goods between Great Britain and the United States involved \$500 of freight and insurance costs.

Third, artificial barriers to trade, such as the special taxes or quotas on imports can hamper traders' ability to move exchange rates toward purchasing power parity.

Still, the purchasing power parity theory is useful in many circumstances. Under floating exchange rates, a country whose relative price level is rising rapidly

Purchasing power parity (PPP)

theory The idea that the exchange rate will adjust in the long run so that the average price of goods in two countries will be roughly the same.

will almost always find that the price of its currency is falling rapidly. If not, all of its tradeable goods would soon be priced out of the world market.

Indeed, we often observe that countries with very high inflation rates have currencies depreciating against the dollar by roughly the amount needed to preserve purchasing power parity. For example, we've already mentioned the sharp depreciation of the Russian ruble from 1992 to 1998. During those six years, the number of rubles that exchanged for a dollar rose from around 100 to about 6,000. Over the same period, the annual inflation rate averaged about 200 percent in Russia, but only about 3 percent in the United States. As a result, the relative price level in Russia skyrocketed, leading to a dramatic depreciation of the ruble against the dollar. Another recent example is Turkey: From mid-1996 to mid-1997, its price level almost doubled, while the dollar price of its currency was cut in half.

INTERDEPENDENT MARKETS: THE ROLE OF ARBITRAGE

The market for pounds—like any other foreign exchange market—is not a centralized market in a single location. Rather, pounds and dollars are exchanged at tens of thousands of locations—at banks, hotels, airports, and train stations in hundreds of cities and towns around the world. How do we know that the equilibrium exchange rate, such as the one we found back in Figure 3 (a), will be the exchange rate in *all* of these locations? Couldn't it be that in New York pounds sell for \$1.50 each, while in London they sell for \$1.60, and in Paris, for \$1.35?

Actually, no. An exchange rate between two currencies will be the same in every location, except for tiny differences that will exist for only a few seconds. Why? Because of the process of **arbitrage**—the simultaneous buying and selling of a foreign currency in order to profit from any difference in exchange rates.

Figure 7 can help us visualize how **bilateral arbitrage**—in which only one pair of currencies is traded—drives an exchange rate to the same equilibrium value around the world. Suppose that in New York (panel (a)) the equilibrium price of the pound was \$1.20, while in London (panel (b)) the price was \$1.80. Then astute traders could make fortunes in minutes. American and British traders could buy pounds in New York for \$1.20 each, while simultaneously selling them in London for \$1.80 each. On each pound traded, they would make a profit of 60 cents. This may not sound like much, but in the foreign exchange market, a professional trader can easily buy and sell millions of dollars' worth of currency in a matter of seconds, with a few keystrokes on a computer. In our example, someone buying \$10 million worth of pounds in New York and selling them in London would make a nice profit of \$6 million—not bad for the few seconds it took to make the trade.

But before you decide to quit college and become a foreign exchange trader, you should know that differences in exchange rates as large as the one in Figure 7 never actually occur. Why not? Because traders—by taking advantage of even the tiniest differences in exchange rates—wipe out those differences entirely.

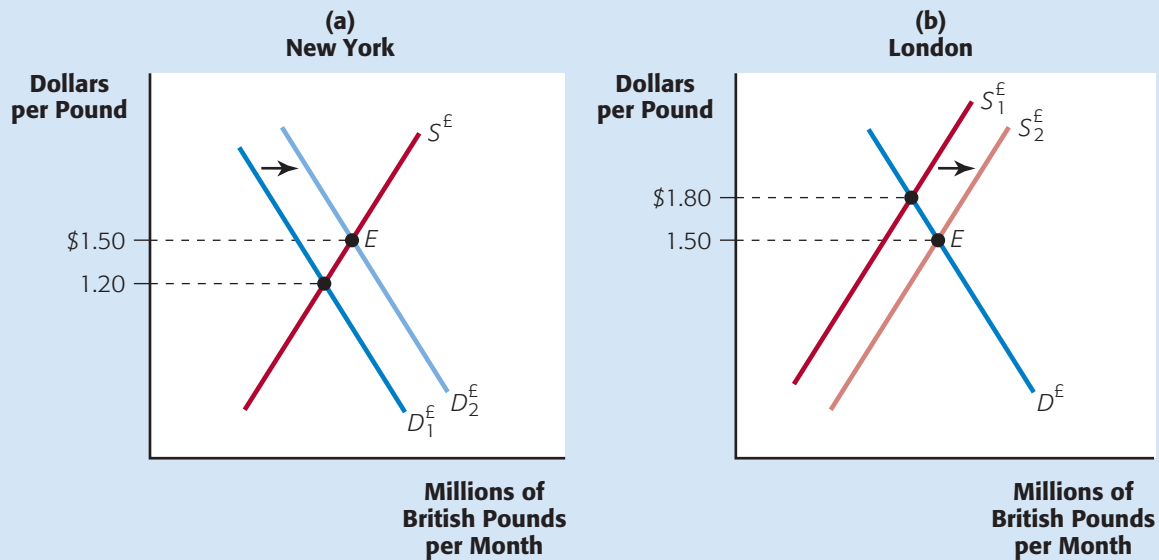
Let's go back to Figure 7 and see how bilateral arbitrage equalizes exchange rates in different locations. As traders buy pounds in New York, the demand curve there shifts rightward, from $D_1^{\$}$ to $D_2^{\$}$, thereby increasing the price of the pound in New York. As traders sell pounds in London, the supply curve there shifts rightward, from $S_1^{\$}$ to $S_2^{\$}$, thereby decreasing the price of pounds in London. The process continues until the exchange rate reaches the same value of \$1.50 in both markets and there are no more profit opportunities for traders.

Arbitrage Simultaneous buying and selling of a foreign currency in order to profit from a difference in exchange rates.

Bilateral arbitrage Arbitrage involving one pair of currencies.

BILATERAL ARBITRAGE

FIGURE 7



Initially, the price of the pound is \$1.20 in New York—panel (a)—and \$1.80 in London—panel (b). Traders take advantage of this exchange rate differential by buying pounds in New York and simultaneously selling them in London. As they do so, the demand curve shifts rightward in New York, and the supply curve shifts rightward in London. Arbitrage continues until the exchange rate attains the same value—\$1.50 per pound—in both locations.

Bilateral arbitrage ensures that the exchange rate between any two currencies is the same everywhere in the world.²

Triangular Arbitrage. Another form of arbitrage—called **triangular arbitrage**—involves trades among *three* (or more) countries' currencies. Triangular arbitrage ensures that the number of dollars that exchange for one pound is the same whether you make the trade *directly*—in the dollar–pound market—or *indirectly*, by buying and selling a third currency.

To see how triangular arbitrage works, suppose that the exchange rates among the U.S. dollar, the British pound, and the Mexican peso are as shown in the left-hand column of Table 2: The price of a pound in dollars is \$1.80, the price of a peso in dollars is \$0.10, and the price of a pound in pesos is 10 pesos.

With these exchange rates, the *direct* price of the pound to Americans is \$1.80. But the *indirect* price is \$1.00. Why? Because an American, starting with \$1.00, could purchase 10 pesos in the dollar–peso market and then use those 10 pesos to purchase 1 pound in the peso–pound market. This difference between the direct and indirect prices for the pound would allow traders to make huge profits. They could

Triangular arbitrage Arbitrage involving trades among three (or more) currencies.

² Exchange rates will sometimes *appear* to be different in different locations because a commission for the broker is often built into the rate. These commissions can differ by location, depending on the cost structure and degree of competition among brokers. For example, if you buy pounds in a small-town bank, which faces little competition and may have higher costs, you may pay more for them than if you bought them in a big-city bank. But this is only because the small-town bank is charging a higher commission.

TABLE 2

BEFORE AND AFTER
TRIANGULAR ARBITRAGE

	Exchange Rate Before Arbitrage	Exchange Rate After Arbitrage
Price of pound in dollar–pound market	\$1.80	\$1.50
Price of peso in dollar–peso market	\$0.10	\$0.125
Price of pound in pound–peso market	10 pesos	12 pesos

acquire pounds *indirectly* for \$1.00 each and then sell them *directly* for \$1.80 each, for a huge profit of 80 cents per pound sold.

However, such large potential profits from triangular arbitrage would never arise in practice. Even the tiniest potential profits would be eliminated, almost immediately, by the arbitrage process itself. In our example, when traders buy pesos with dollars, they *drive up the price of the peso in the dollar–peso market*. When they buy pounds with pesos, they *drive up the price of the pound in the pound–peso market*. Finally, when they buy dollars with pounds to make their profit, they *drive down the price of the pound in the dollar–pound market*.

Each of these movements decreases the potential profits from arbitrage, and the process ends when no opportunity for such profits remains. The third column in Table 2 shows where the exchange rates might end up after the arbitrage process is completed. With these exchange rates, the direct price of the pound is \$1.50. And this is also what it would cost to buy a pound *indirectly*: \$1.50 gets you 12 pesos, and 12 pesos gets you one pound. There are no more opportunities for arbitrage, because arbitrage has eliminated them.

Triangular arbitrage ensures that the price of a foreign currency is the same whether it is purchased directly—in a single foreign exchange market—or indirectly, by buying and selling a third currency.³

GOVERNMENT INTERVENTION IN FOREIGN EXCHANGE MARKETS

As you've seen, when exchange rates float, they can rise and fall for a variety of reasons. But a government may not be content to let the forces of supply and demand change its exchange rate. If the exchange rate rises, the country's goods will become much more expensive to foreigners, causing harm to its export-oriented industries. If the exchange rate falls, goods purchased from other countries will rise in price. Since many imported goods are used as inputs by U.S. firms (such as oil from the Middle East and Mexico, or computer screens from Japan), a drop in the exchange rate will cause a rise in the U.S. price level. Finally, if the exchange rate is too volatile, it can make trading riskier or require traders to acquire special insurance against foreign currency losses, which costs them money, time, and trouble. For all of these reasons, governments sometime *intervene* in foreign exchange markets involving their currency.

³ Because brokerage commissions are sometimes built into the price of foreign currency, small differences between the direct and indirect price may remain, even after arbitrage has eliminated all possibilities of profit. This is because two commissions are paid when a person buys indirectly, but only one commission is paid when buying directly.

MANAGED FLOAT

Many governments let their exchange rate float *most of the time*, but will intervene on occasion when the floating exchange rate moves in an undesired direction or becomes too volatile. For example, look back at Figure 5, where the price of the British pound falls to \$1 as hot money is shifted out of British assets. Suppose the British government does not want the pound to depreciate. Then its central bank—the Bank of England—could begin trading in the dollar–pound market itself. It would buy British pounds with dollars, thereby shifting the demand for pounds curve rightward. If it buys just the right amount of pounds, it can prevent the pound from depreciating at all. Alternatively, the U.S. government might not be happy with the *appreciation* of the dollar in Figure 5. In that case, the Federal Reserve can enter the market and buy British pounds with dollars, once again shifting the demand for pounds curve rightward.

The central banks of many countries—including the Federal Reserve—will sometimes intervene in this way in foreign exchange markets. When a government buys or sells its own currency or that of a trading partner to influence exchange rates, it is engaging in a “managed float” or a “dirty float.”

Under a managed float, a country’s central bank actively manages its exchange rate, buying its own currency to prevent depreciations, and selling its own currency to prevent appreciations.

Managed floats are used most often in the very short run, to prevent large, sudden changes in exchange rates. For example, on a single day—March 8, 2000—the Bank of Japan (Japan’s central bank) sold over 200 billion yen (about \$2 billion worth) in order to stop a rapid appreciation of the yen against the dollar. On the other side, during 1998, the central bank of Guatemala bought almost 2 billion quetzals (about \$300 million worth) in order to slow the depreciation of that currency.

That last example raises a question. When a country—such as Guatemala—wants to prevent or slow a depreciation against the dollar, it has to buy its own currency with dollars. Where does it get those dollars? Unfortunately for Guatemala, it is not legally permitted to print dollars—only the U.S. Federal Reserve can do that. Instead, Guatemala must use its *reserves* of dollars—the dollars its central bank keeps on hand specifically to intervene in the dollar–quetzal market.

Almost every nation holds reserves of dollars—as well as euros, yen, and other key currencies—just so it can enter the foreign exchange market and sell them for its own currency when necessary. Under a managed float, periods of selling dollars are usually short-lived, and alternate with periods of buying dollars. Thus, countries rarely use up all of their dollar reserves when they engage in managed floats.

Managed floats are controversial. Some economists believe they help to avoid wide swings in exchange rates, and thus reduce the risks for international traders and investors. But others are critical of how managed floats often work out in practice. They point out that countries often intervene when the forces behind an appreciation or depreciation are strong. In these cases, the intervention only serves to delay inevitable changes in the exchange rate—sometimes, at great cost to a country’s reserves of dollars and other key currencies.

FIXED EXCHANGE RATES

A more extreme form of intervention is a **fixed exchange rate**, in which a government declares a particular value for its exchange rate with another currency. The

Managed float A policy of frequent central bank intervention to move the exchange rate.

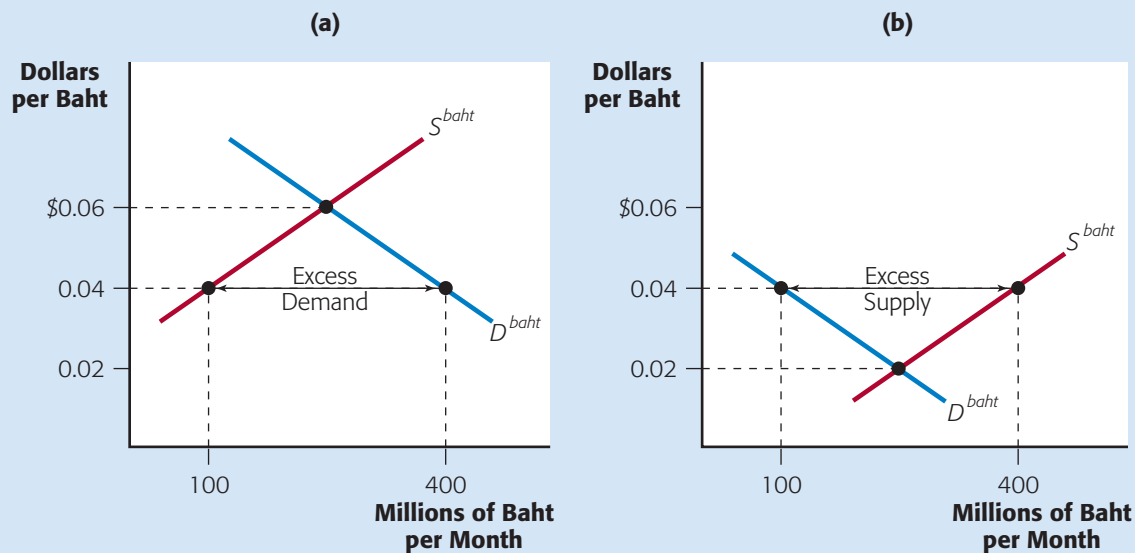


If you are interested in learning more about exchange rate systems, read “The International Financial Architecture” by Jeffrey Frankel at <http://www.brook.edu/comm/PolicyBriefs/pb051/pb51.htm>.

Fixed exchange rate A government-declared exchange rate maintained by central bank intervention in the foreign exchange market.

FIGURE 8

A FIXED EXCHANGE RATE FOR THE BAHT



In both panels, Thailand's central bank fixes the exchange rate at \$0.04 per baht. In panel (a), the equilibrium exchange rate is \$0.06 per baht—higher than the fixed rate. The central bank must sell 300 million baht per month—an amount equal to the excess demand. In panel (b), which shows a different set of supply and demand curves, the equilibrium exchange rate is \$0.02 per baht—lower than the fixed rate. The central bank must buy up the excess supply of 300 million baht.

government, through its central bank, then commits itself to intervene in the foreign exchange market any time the *equilibrium* exchange rate differs from the *fixed* rate.

For example, from 1987 to 1997, the government of Thailand fixed the value of its currency—the *baht*—at \$0.04 per baht. The two panels of Figure 8 show the different types of intervention that might be necessary in the baht–dollar market to maintain this fixed exchange rate. Each panel shows a different set of supply and demand curves—and a different equilibrium exchange rate that might exist for the baht. Look first at panel (a). Here, we assume that the equilibrium exchange rate is \$0.06 per baht, so that the fixed rate is *lower* than the equilibrium rate. At the fixed rate of \$0.04 per baht, 400 million baht would be demanded each month, but only 100 million would be supplied. There would be an *excess demand* of 300 million baht, which would ordinarily drive the exchange rate back up to its equilibrium value of \$0.06. But the Thai government prevents this by entering the market and *selling* just enough baht to cover the excess demand. In panel (a), the Central Bank of Thailand would sell 300 million baht per month to maintain the fixed rate.

When a country fixes its exchange rate below the equilibrium value, the result is an excess demand for the country's currency. To maintain the fixed rate, the country's central bank must sell enough of its own currency to eliminate the excess demand.

Panel (b) shows another possibility, where the equilibrium exchange rate is \$0.02, so that the same fixed exchange rate of \$0.04 per baht is now *above* the equilibrium rate. There is an excess *supply* of 300 million baht. In this case, to pre-

A FOREIGN CURRENCY CRISIS

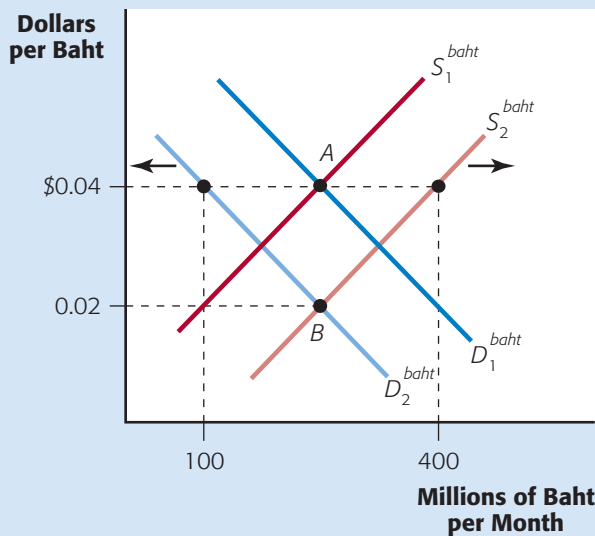


FIGURE 9

Initially, the baht is fixed at the equilibrium rate of \$0.04. When the supply and demand curves shift to D_2 and S_2 , the equilibrium exchange rate falls to \$0.02. If Thailand continues to fix the rate at \$0.04, it will have to buy up the excess supply of 300 million baht per month, using dollars. As its dollar reserves dwindle, traders will anticipate a drop in the value of the baht, shifting the curves out further, as indicated by the heavy arrows.

vent the excess supply from driving the exchange rate down, the Central Bank of Thailand must *buy* the excess baht.

When a country fixes its exchange rate above the equilibrium value, the result is an excess supply of the country's currency. To maintain the fixed rate, the country's central bank must buy enough of its own currency to eliminate the excess supply.

Fixed exchange rates present little problem for a country as long as the exchange rate is fixed at or very close to its equilibrium rate. But when the equilibrium exchange rate moves away from the fixed rate—as in the two panels of Figure 8—governments often try to maintain their fixed rate anyway, sometimes for long periods. This can create problems, especially when the exchange rate is fixed *above* the equilibrium rate.

To see why, look at Figure 9. Initially, the supply and demand curves for baht are given by S_1 and D_1 , respectively, so that the equilibrium exchange rate, \$0.04, is equal to the fixed exchange rate. At this point, the central bank is neither selling nor buying baht. Now, suppose that, for some reason (we'll be more specific in a few paragraphs), the supply and demand curves shift to S_2 and D_2 , respectively. The equilibrium rate falls, so the fixed rate of \$0.04 is above the equilibrium rate of \$0.02. The Central Bank of Thailand must now *buy* its own currency with dollars—at the rate of 300 million baht per month. Each baht costs the central bank 4 cents, so as the months go by, its dollar reserves are being depleted at the rate of 300 million \times \$0.04 = \$12 million per month. Once those reserves are gone, Thailand will have only two choices: to let its currency float (which means an immediate depreciation to the lower, equilibrium rate), or to declare a new, lower fixed rate—a **devaluation** of its currency.

Of course, at a certain point, foreign exchange speculators and traders would see that Thailand doesn't have many dollars left. (Most countries' central banks regularly

Devaluation A change in the exchange rate from a higher fixed rate to a lower fixed rate.

report their holdings of key currencies, and economists can estimate the holdings of countries that don't.) Looking ahead, these speculators and traders will begin to *anticipate* a drop in the baht. And—as you've learned in this chapter—expected changes in the exchange rate *shift* supply and demand curves for foreign currency. In this case, an expected fall in the baht causes the supply curve for baht to shift further rightward and the demand curve to shift further leftward, as indicated by the heavy arrows in the diagram. In Figure 9, these shifts will *decrease* the equilibrium value of the baht, increase the *excess supply* of baht, and make the fixed rate of \$0.04 even harder to maintain. The country is now experiencing a *foreign currency crisis*.

Foreign currency crisis A loss of faith that a country can prevent a drop in its exchange rate, leading to a rapid depletion of its foreign currency (e.g., dollar) reserves.

A foreign currency crisis arises when people no longer believe that a country can maintain a fixed exchange rate above the equilibrium rate. As a consequence, the supply of the currency increases, demand for it decreases, and the country must use up its reserves of dollars and other key currencies even faster in order to maintain the fixed rate.

Once a foreign currency crisis arises, a country typically has no choice but to devalue its currency or let it float and watch it depreciate. And ironically, because the country waited for the crisis to develop, the exchange rate may for a time drop even lower than the original equilibrium rate. For example, in Figure 9, an early devaluation to \$0.02 per dollar might prevent a crisis from occurring at all. But once the crisis begins, and the supply and demand curves shift out further than S_2 and D_2 , the currency will have to drop *below* \$0.02 to end the rapid depletion of dollar reserves.



<http://>

Professor Nouriel Roubini of New York University maintains an excellent Web page devoted to global financial crises. You can find it at <http://www.stern.nyu.edu/~nroubini/asia/AsiaHomepage.html>

The Asian Financial Crisis. In 1997, several Asian countries came very close to complete financial collapse. And fixed exchange rates and foreign currency crises—as just described—played a central role. The crisis had its roots in a practice that was common in Asia, especially in the five “frontline” countries most directly affected by the crisis—Thailand, Indonesia, South Korea, Malaysia, and the Philippines.

In these countries, banks borrowed dollars or yen in world markets, and then lent to domestic businesses in the local currency. Thus, banks were vulnerable to declines in exchange rates. For example, if the baht fell, a Thai bank would need more baht to pay back its own loans, while collecting the same number of baht from the local firms it had lent to.

Moreover, in these countries, exchange rates were managed by the government. If foreign exchange traders believed that a government was short of reserves, the exchange rate could fall dramatically. Furthermore, foreign exchange traders knew that these governments were guaranteeing the obligations of their banks. Therefore, any decline in the exchange rate would drain government reserves, making it that much harder to stabilize the exchange rate. This explains why Thailand and the other frontline countries didn't just devalue, or let their currencies float at the first sign of trouble. Even a modest devaluation would have caused their banks to fail.

But a drop in the exchange rate was inevitable. And Thailand's currency was the first to go. In July 1997, the Thai central bank—having defended its fixed rate down to almost its last dollar of reserves—had no choice but to let its currency float. The baht immediately depreciated from \$0.04 to \$0.02, and Thailand's banks were immediately in trouble.

The baht's depreciation then led to fears of depreciation or devaluation in *other* Asian countries, and served to worsen *their* crises. And there was another impact: In country after country, bank lending to businesses dried up. After all, who wants to put funds into a Korean bank when the Korean won is about to be devalued? Without sufficient funding, many businesses were forced to shut down, others lan-

guished, and millions of workers lost their jobs. Here is one way to measure the impact of the crisis: From 1990 to 1996, the average growth rate of the five Asian frontline countries was 7 percent per year. In 1998, the output of these countries *fell* by an average of 7 percent.

The Asian financial crisis lasted more than a year. Before it ended, investors—who had been awakened to the realities of devaluations and other risks—spread the crisis from country to country, and even to several Latin American countries.

In retrospect, the central cause of the crisis was the instability of banks that borrowed in dollars, yen, and other more stable currencies, and lent in their local currency. But government attempts to protect these banks by fixing or managing exchange rates proved to be a costly, and—ultimately—a losing, battle.

THE EURO

One answer to the problems that countries have encountered in managing their own currencies is to adopt another country's currency or an international currency. For example, Argentina has a currency that is locked to the U.S. dollar. In the 11 Euro-land countries, including Germany, France, Italy, and Spain, national exchange rates have already been fixed to a new European currency—the euro.

But the European nations are going even further: In 2002, their national currencies will cease to exist entirely, to be replaced by the euro. At that time, the European Central Bank will have sole authority for changing the supply of euros. It will determine a single monetary policy for all of Euroland, replacing the separate monetary policies of the different countries.

Why have these 11 European countries decided to do away with their national currencies?

There are several advantages. First, a single currency means that European firms—when they buy or sell across borders—will no longer have to pay commissions on the exchange of currency, or face the risk that exchange rates might change before accounts are settled. This should increase the volume of trade among the Euroland nations. Second, the elimination of exchange rate risk makes it easier for European firms to sell stocks and bonds to residents anywhere in Euro-land. This will help ensure that funds are channeled to the most profitable firms throughout the area. Third, adopting a single currency makes cross-country comparison shopping easier. This should help increase competition among firms, and help keep prices down to European consumers. Finally, some of these countries—such as Italy—have had a history of loose monetary policy that has generated high rates of inflation, and high expected inflation. By giving up the right to run an independent monetary policy, and leaving it to the (presumably stricter) European Central Bank, the high-inflation countries of Europe will benefit from lower inflation rates.

There are, however, downsides to the euro. In fact, some economists believe that—at least for a while—the euro will create significant problems for the Euro-land countries. Why? With a single currency, there must be a single monetary policy, making it impossible to adjust the money supply and interest rates to the problems of individual nations. For example, suppose Spain goes into a recession. In the old days before the euro, its central bank would increase the Spanish money supply and lower interest rates. But now, what if Spain's recession is accompanied by full employment or even a boom in the rest of Europe? Then, the European central bank will be tightening the money supply and raising interest rates, which will worsen conditions in Spain. Spain could always use fiscal policy. But, as you've learned, countercyclical fiscal policy is fraught with problems. Moreover, membership in



The Euro—the new common European currency—wasn't yet in circulation when this chapter was being written. But this advertisement shows prototypes of the bills that people in France, Italy, Germany, Spain, and seven other nations will carry in their wallets, beginning in 2002.

Optimum currency area A region whose economies perform better with a single currency than with separate national currencies.

Euroland requires countries to maintain strict fiscal discipline that might prevent them from using a fiscal stimulus when it is needed.

The economists who worry about these problems question whether Europe is an **optimum currency area**—a region whose economies will perform better with a single currency rather than separate national currencies. To be an optimum currency area, the different nations in a region should face common, rather than national, shocks, so that they tend to go into booms and recessions together. In that case, a single monetary policy will be appropriate, because all nations will need stimulus or restraint at the same time. In Europe, unfortunately, the shocks are often national: Different countries are dominated by different industries, face different types of labor unions, and have different institutional frameworks and laws. They are therefore susceptible to national as well as regional shocks.

Another requirement for an optimum currency area is that labor is highly mobile from one country to another. That way, if one country is experiencing a negative shock and goes into a recession that can't be addressed with monetary or fiscal policy (for the reasons discussed earlier), at least its unemployed workers can find work in other countries whose economies are performing better. Indeed, this is what happens in the United States, where labor is highly mobile among states.

But at present, labor is much less mobile across European borders than across the American states. And if unemployed workers stay within a country, its government may feel pressure to abandon the euro so that it can use expansionary fiscal and monetary policy.

In the very long run, the abolition of national currencies—and the creation of the euro—may work to increase labor mobility across Europe, especially if it changes the attitudes of European firms and workers toward cross-national employment. Europe may then move closer to being an optimum currency area in the future.

EXCHANGE RATES AND THE MACROECONOMY

Exchange rates can have important effects on the macroeconomy—largely through their effect on net exports. And although we've included net exports in our short-run macro model, we haven't yet asked how exchange rates affect them. That's what we'll do now.

What Happens When
Things Change?



EXCHANGE RATES AND SPENDING SHOCKS

Suppose that the dollar depreciates against the foreign currencies of its major trading partners. (We'll discuss *why* that might happen in a later section.) Then U.S. goods would become cheaper to foreigners, and net exports would rise at each level of output. This increase in net exports is a positive spending shock to the economy—it increases aggregate expenditure. And, as you've learned, positive spending shocks increase GDP in the short run.

A depreciation of the dollar causes net exports to rise—a positive spending shock that increases real GDP in the short run. An appreciation of the dollar causes net exports to drop—a negative spending shock that decreases real GDP in the short run.

The impact of net exports on equilibrium GDP—often caused by changes in the exchange rate—helps us understand one reason why governments are often concerned about their exchange rates. An unstable exchange rate can result in re-

peated shocks to the economy. At worst, this can cause fluctuations in GDP; at best, it makes the central bank's job more difficult as it tries to keep the economy on an even keel.

EXCHANGE RATES AND MONETARY POLICY

In several earlier chapters, we've explored how the Fed tries to keep the U.S. economy on an even keel with monetary policy. The central banks around the world are engaged in a similar struggle, and face many of the same challenges as the Fed. One challenge to central banks is that monetary policy causes changes in exchange rates, and thus has additional effects on real GDP that we have not yet considered.

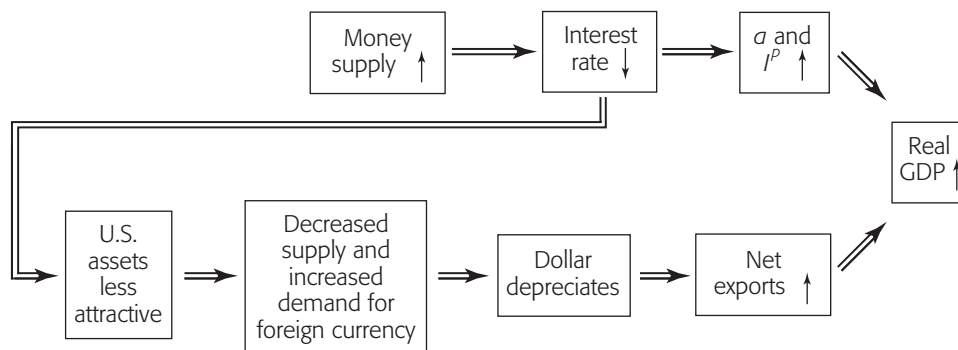
To understand this, let's run through an example. Suppose the United States is in a recession, and the Fed decides to increase equilibrium GDP. As you've learned, the Fed—by increasing the money supply—brings down the interest rate. Interest-sensitive spending rises, and so does aggregate expenditure. When we consider the foreign exchange market, however, there is an additional effect on aggregate expenditure.

By lowering the U.S. interest rate, the Fed makes foreign financial assets more attractive to Americans, which raises their demand for foreign currency. In the market for pounds, for example, this will shift the demand for pounds curve rightward. At the same time, U.S. financial assets become less attractive to foreigners, which decreases the supply of foreign exchange (in the market for pounds, a leftward shift in the supply of pounds curve). If you sketch out these shifts right now, you'll see that—as long as the exchange rate floats—the result is a *depreciation of the dollar* against the pound.

Now let's see how the depreciation of the dollar affects the economy. With dollars now cheaper for foreigners, they will buy more U.S. goods, raising U.S. exports. At the same time, with foreign goods and services more expensive to Americans, U.S. imports will decrease. Both the increase in exports and the decrease in imports contribute to a rise in net exports, NX . This, in turn, increases aggregate expenditure.

Thus, as you can see, the expansionary monetary policy causes aggregate expenditures to rise in two ways: first, by increasing interest-sensitive spending, and second, by increasing net exports. As a result, equilibrium GDP rises by more—and monetary policy is more effective—when the effects on exchange rates are included.

The channels through which monetary policy works are summarized in the following schematic:



Net Effect: GDP ↑ by more when the exchange rate's effect on net exports is included

The top line shows the familiar effect on interest-sensitive spending: An increase in the money supply causes a drop in the interest rate, which increases autonomous



What Happens When Things Change?

consumption spending (a) and investment spending (I^p). The bottom line shows the *additional* effect on net exports through changes in the exchange rate—the effects we’ve been discussing.

The analysis of contractionary monetary policy is the same, but in reverse. A decrease in the money supply will not only decrease interest-sensitive spending, it will also cause the dollar to appreciate and net exports to drop. Thus, it will cause equilibrium GDP to fall by more than in earlier chapters, where we ignored the foreign exchange market.

The channel of monetary influence through exchange rates and the volume of trade is an important part of the full story of monetary policy in the United States. And in countries where exports are relatively large fractions of GDP—such as those of Europe—the trade channel is even more important. It is the main channel through which monetary policy affects the economy.

Monetary policy has a stronger effect when we include the impact on exchange rates and net exports, rather than just the impact on interest-sensitive consumption and investment spending.

THE STUBBORN U.S. TRADE DEFICIT

Using the THEORY



Trade deficit The excess of a nation’s imports over its exports during a given period.

Trade surplus The excess of a nation’s exports over its imports during a given period.

The U.S. trade deficit is often in the news. But what, exactly, is it?

The trade deficit is the extent to which a country’s imports exceed its exports:

$$\text{Trade deficit} = \text{imports} - \text{exports.}$$

On the other hand, when exports exceed imports, a nation has a trade surplus:

$$\text{Trade surplus} = \text{exports} - \text{imports.}$$

As you can see, the trade surplus is nothing more than a nation’s net exports (NX). And when net exports are negative, we have a trade deficit.

The United States has had large trade deficits with the rest of the world since the early 1980s. In 1999, the trade deficit hit an all-time high of \$268 billion. Simply put, Americans bought \$268 billion more goods and services from other countries than their residents bought from the United States.

Why does the United States have a trade deficit with the rest of the world? A variety of explanations have been offered in the media, including the relatively low quality of U.S. goods (compared to, say, Japan), poor U.S. marketing savvy in selling to foreigners, and a greater degree of protectionism in foreign markets.

But economists believe that there is a much more important reason. In this section, we’ll use what you’ve learned about floating exchange rates to show how the U.S. trade deficit arose and why it continues. To keep our analysis simple, we’ll look at the U.S. trade deficit with just one country—Japan—but our results will hold more generally to the trade deficit with other countries as well.

Before we analyze the causes of the trade deficit, we need to do a little math. Let’s begin by breaking down the total quantity of yen demanded by Americans (D^Y) into two components: the yen demanded to purchase Japanese goods and services (U.S. imports from Japan) and the yen demanded to buy Japanese assets:

$$D^{\text{¥}} = \text{U.S. imports from Japan} + \text{U.S. purchases of Japanese assets.}$$

Similarly, we can divide the total quantity of yen supplied by the Japanese ($S^{\text{¥}}$) into two components: the yen exchanged for dollars to purchase American goods (U.S. exports to Japan), and the yen exchanged for dollars to purchase American assets like stocks, bonds, or real estate:

$$S^{\text{¥}} = \text{U.S. exports to Japan} + \text{Japanese purchases of U.S. assets.}$$

As long as the yen floats against the dollar without government intervention—which it does during most periods—we know that the exchange rate will adjust until the quantities of yen supplied and demanded are equal, or $D^{\text{¥}} = S^{\text{¥}}$. Substituting the foregoing breakdowns into this equation, we have

$$\begin{aligned} & \text{U.S. imports from Japan} + \text{U.S. purchases of Japanese assets} \\ &= \text{U.S. exports to Japan} + \text{Japanese purchases of U.S. assets.} \end{aligned}$$

Now let's rearrange this equation—subtracting U.S. exports from both sides, and subtracting American purchases of Japanese assets from both sides, to get

$$\begin{aligned} & \text{U.S. imports from Japan} - \text{U.S. exports to Japan} \\ &= \text{Japanese purchases of U.S. assets} - \text{U.S. purchases of Japanese assets.} \end{aligned}$$

The term on the left should look familiar: It is the U.S. trade deficit with Japan. And since a similar equation must hold for every country, we can generalize it this way:

$$\begin{aligned} & \text{U.S. imports from other countries} - \text{U.S. exports to other countries} \\ &= \text{foreign purchases of U.S. assets} - \text{U.S. purchases of foreign assets.} \end{aligned}$$

But what is the expression on the right? It tells us the extent to which foreigners are buying more of our assets than we are buying of theirs. It is often called the **net capital inflow** into the United States, because when the residents of other countries buy U.S. assets, funds flow into the U.S. financial market, where they are made available to U.S. firms and the U.S. government. Thus, the equation we've derived—which must hold true when exchange rates float—can also be expressed as

$$\text{U.S. trade deficit} = \text{U.S. net capital inflow.}$$

Why have we bothered to derive this equation? Because it tells us two very important things about the U.S. trade deficit. First, it tells us how the trade deficit is *financed*. Think about it: If the United States is running a trade deficit with, say, Japan, it means that the Japanese are providing more goods and services to Americans—more automobiles, VCRs, memory chips, and other goods—than Americans are providing to them. The Japanese are not doing this out of kindness. They must be getting *something* in return for the extra goods we are getting, and the equation tells us just what that is: U.S. assets. This is one reason why the trade deficit concerns U.S. policy makers: It results in a transfer of wealth from Americans to foreign residents.

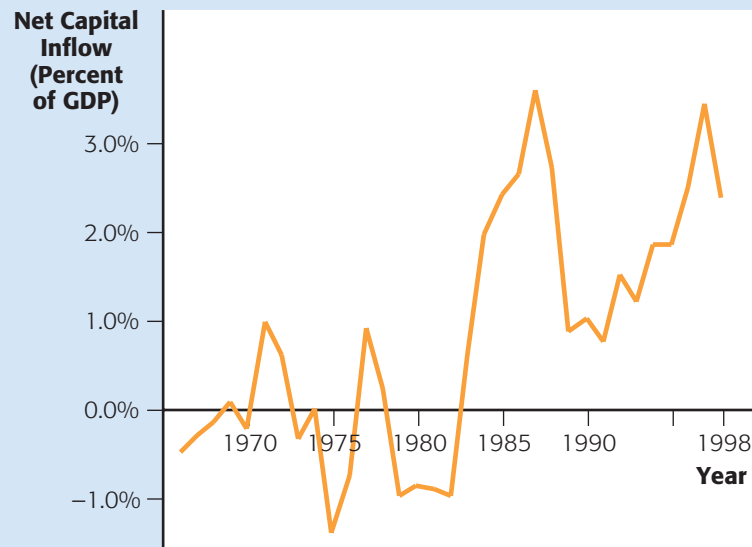
The second important insight provided by the equation is that a trade deficit can arise *because* of forces that cause a capital inflow. That is, if forces in the global economy make the right side of the equation positive, then the left side must be positive as well, and we will have a trade deficit. Indeed, economists believe this

Net capital inflow An inflow of funds equal to a nation's trade deficit.

FIGURE 10

Beginning in the early 1980s, and continuing today, a massive capital inflow has caused a U.S. trade deficit. The capital inflow was originally caused by high U.S. interest rates relative to interest rates abroad. But in the 1990s and early 2000s, the capital inflow has been sustained by the favorable investment climate in the United States, and by the explosive growth of the Internet sector.

NET CAPITAL FLOWS INTO THE UNITED STATES AS A PERCENT OF GDP



is just what has happened to the United States: that the U.S. trade deficit has been caused by the desire of foreigners to invest in the United States. The result was a massive capital inflow and trade deficit that arose in the early 1980s, as illustrated in Figure 10. That capital inflow was unprecedented in size and duration, and it reversed a long-standing pattern of ownership between the United States and other countries. For decades, American holdings of foreign assets far exceeded foreign holdings of U.S. assets. But the capital inflows of the 1980s changed that: By 1988, foreigners held about \$500 billion more in U.S. assets than Americans held in foreign assets. Ten years later, the difference in asset holdings had tripled to about \$1.5 trillion.

But how do the forces that create a capital inflow also *cause* a trade deficit? When foreigners start buying more of our assets than we are buying of theirs, the dollar will appreciate above the value it would have if there were no net capital inflow. This makes U.S. goods more expensive to foreigners, and foreign goods cheaper to Americans. Thus,

an increase in the desire of foreigners to invest in the United States contributes to an appreciation of the dollar. As a result, U.S. exports—which become more expensive for foreigners—decline. Imports—which become cheaper to Americans—increase. The result is a rise in the trade deficit.

How can we explain the huge capital inflow that began in the 1980s, and has grown larger over the past decade? In the 1980s, an important part of the story was *a rise in U.S. interest rates relative to interest rates abroad*, which made U.S. assets more attractive to foreigners, and foreign assets less attractive to Americans. In the 1990s, however, U.S. interest rates were low relative to rates in other countries, yet the inflow continued. Why?

Even when U.S. interest rates are the same or lower than abroad, it seems that residents of other countries have a strong preference for holding American assets.

In part, this is because of a favorable investment climate. The United States is a stable country with a long history of protecting individual property rights. People know that if they buy American stocks or bonds, the U.S. government is very unlikely to confiscate these assets or suddenly impose punitive taxes when foreigners want to repatriate the funds to their home countries. The United States also has a stable legal and financial system, reassuring foreign investors that they will be treated fairly and on equal footing with U.S. citizens.

And in the late 1990s, there was another reason for the growing capital inflow: American companies took the lead in exploiting the Internet. New businesses—with the prospect of high future profits—sprang up daily, issuing shares of stock to anyone in the world who wanted to buy them. Thus, an asymmetry developed: The U.S. was offering assets that foreigners found attractive, while no foreign country was offering assets that Americans found nearly as attractive. As we entered the early 2000s, the attractiveness of U.S. stock was continuing to feed the U.S. capital inflow, with no end in sight.

Remember that, under floating exchange rates, the capital inflow equals the trade deficit. Thus, the story of the U.S. capital inflow of the 1980s and 1990s is also the story of the U.S. trade deficit:

We can trace the rise in the trade deficit during the 1980s and 1990s to two important sources: first, relatively high interest rates in the 1980s; and second, a long-held preference for American assets that grew stronger in the 1990s. Each of these contributed to a large capital inflow, a higher value for the dollar, and a trade deficit.

S U M M A R Y

When residents of two countries trade with one another, one party ordinarily makes use of the foreign exchange market to trade one national currency for another. In this market, suppliers of a currency interact with demanders to determine an exchange rate—the price of one currency in terms of another.

In the market for U.S. dollars and British pounds, for example, demanders are mostly Americans who wish to obtain pounds in order to buy goods and services from British firms, or to buy British assets. A higher dollar price for the pound will lead Americans to demand fewer pounds—the demand curve slopes downward. Changes in U.S. real GDP, the U.S. price level relative to the British price level, Americans' tastes for British goods, interest rates in the United States relative to Britain, or expectations regarding the exchange rate, can each cause the demand curve to shift.

Suppliers of pounds are mostly British residents who wish to buy American goods, services, or assets. A higher dollar price for the pound will lead Britons to supply more pounds—the supply curve slopes upward. The supply curve will shift in response to changes in British real GDP, prices in Britain relative to the United States, British tastes for U.S. goods, the British interest rate relative to the U.S. rate, and expectations regarding the exchange rate.

When the exchange rate floats, the equilibrium rate is determined where the supply and demand curves cross. If the equilibrium is disturbed by, say, a rightward shift of the de-

mand curve, then the currency being demanded will appreciate—the exchange rate will rise. (The other country's currency will depreciate.) In a similar way, a rightward shift of the supply curve will cause the currency being supplied to depreciate.

In practice, each country's currency is traded in a variety of markets around the world. Currency traders, in search of profits, engage in arbitrage whenever the exchange rate differs between two markets. This activity—buying low and selling high—serves to eliminate any exchange rate differentials. A more complex form of arbitrage ensures that the direct and indirect prices of one currency in terms of another will be the same.

Governments often intervene in foreign exchange markets. Many countries manage their float—buying and selling their own currency to alter the exchange rate. Some countries fix their exchange rate to the dollar or the currency of a major trading partner. And in Europe, 11 national governments are on the way to eliminating their national currency, and replacing it with the Euro. Although these 11 nations may not yet be an optimum currency area, they are moving closer to one.

When a currency depreciates, its net exports rise—a positive spending shock. Monetary policy, in addition to its impact on interest-sensitive spending—also changes the exchange rate and net exports, adding to changes in output. This monetary policy is more effective in changing GDP when its effects on net exports are included.

KEY TERMS

foreign exchange market	floating exchange rate	bilateral arbitrage	optimum currency area
exchange rate	appreciation	triangular arbitrage	trade deficit
demand curve for foreign currency	depreciation	managed float	trade surplus
supply curve for foreign currency	purchasing power parity (PPP) theory	fixed exchange rate	net capital inflow
	arbitrage	devaluation	
		foreign currency crisis	

REVIEW QUESTIONS

- Why do Americans demand foreign currency? Why does the demand curve for foreign currency slope downward? What factors shift the demand curve for foreign currency to the right? What factors shift it to the left?
- Why do foreigners supply foreign currency? Why does the supply of foreign currency curve slope upward? What factors shift the supply curve for foreign currency to the right? What factors shift it to the left?
- Explain how an expected appreciation of a foreign currency can become a self-fulfilling prophecy.
- What forces move exchange rates in the very short run? In the short run?
- “A weak currency is a sign of a sick economy.” True or false? Explain.
- What is purchasing power parity? Why might exchange rates deviate from purchasing power parity?
- Suppose the purchasing power parity exchange rate between the dollar and the pound is \$1.50 per pound but, the actual exchange rate is \$2 per pound. Explain how a trader could profit by buying a basket of goods in one country (which country?) and selling it in the other. How would such actions by traders affect the exchange rate?
- What is a managed float and why would a government use it?
- What is the difference between bilateral arbitrage and triangular arbitrage? What would be different about foreign exchange markets if neither type of arbitrage took place?
- How does an appreciation of the dollar affect U.S. real GDP?
- According to economists, what caused the U.S. trade deficit in the 1980s? Why does the trade deficit persist?

PROBLEMS AND EXERCISES

- Do the following events cause the dollar to appreciate against the French franc or to depreciate?
 - Health experts discover that red wine, especially French red wine, lowers cholesterol.
 - France’s GDP falls.
 - The United States experiences a higher inflation rate than France does.
 - The United States runs a large budget deficit.
- Suppose the U.S. government intervenes in the foreign currency market and uses U.S. dollars to buy 2 million pounds. What happens to the exchange rate? Why might the U.S. government do this?
- Suppose the following are the exchange rates among the U.S. dollar, the Mexican peso, and the Euro:

$$\text{Dollars per peso} = 0.2.$$

$$\text{Dollars per euro} = 0.5.$$

$$\text{Euros per peso} = 0.3.$$

Is there an opportunity for triangular arbitrage? If so, how would it work?

- Let the demand for British pounds and the supply of British pounds be described by the following equations:

$$\text{Demand for pounds} = 10 - 2e$$

$$\text{Supply of pounds} = 4 + 3e,$$

where the quantities are in millions of pounds and e is dollars per pound.

- Find the equilibrium exchange rate.

- Suppose the United States and Mexico are each other’s sole trading partners. The Fed, afraid that the economy is about to overheat, decreases the U.S. money supply.

- a. Will the dollar appreciate or depreciate against the Mexican peso? Illustrate with a diagram of the dollar–peso foreign exchange market.
- b. What will happen to equilibrium GDP in the United States?
- c. How would your analyses in (a) and (b) change if, at the same time that the Fed was increasing the U.S. interest rate, the Mexican central bank increased the Mexican interest rate by an equivalent amount?

C H A L L E N G E Q U E S T I O N S

1. It is often stated that the U.S. trade deficit with Japan results from Japanese trade barriers against U.S. goods.
 - a. Suppose that Japan and the U.S. trade goods but not assets. Show—with a diagram of the dollar–yen market—that a trade deficit is impossible. (*Hint:* With no trading in assets, the quantity of yen demanded at each exchange rate is equal in value to U.S. imports, and the quantity of yen supplied at each exchange rate is equal in value to U.S. exports.)
 - b. In the diagram, illustrate the impact of a reduction in Japanese trade barriers. Would the dollar appreciate or depreciate against the yen? What would be the impact on U.S. net exports?
 - c. Now suppose that the United States and Japan also trade assets, but that the Japanese buy more U.S. assets than we buy of theirs. Could the elimination of Japanese trade barriers wipe out the U.S. trade deficit with Japan? Why, or why not? (*Hint:* What is the relationship between the U.S. trade deficit and U.S. net capital inflow?)
2. Suppose that the U.S. government raises spending without increasing taxes. Will there be any effects on the foreign exchange market? (*Hint:* What does this policy do to U.S. interest rates?) When we take the foreign exchange market and net exports into account, will this policy be more effective or less effective in changing equilibrium GDP in the short run?

E X P E R I E N T I A L E X E R C I S E S

1. Trade among European nations will be bolstered by the introduction of the euro—a common European currency. Not surprisingly, there are special Web pages devoted to the euro. The official European Union Web site—<http://europa.eu.int/euro/html/entry.html>—is one of them. Go to this Web page to review the latest developments.



2. The latest data on exchange rates appear in the Currency Trading column in the *Wall Street Journal*. You can find it in the Money & Investing section. Try tracking a particular currency over the course of several weeks. Has the dollar been appreciating or depreciating relative to that currency? Try to explain the behavior of the exchange rate based on what you've learned in this chapter.

Using All the THEORY

THE STOCK MARKET AND THE MACROECONOMY

CHAPTER OUTLINE

Basic Background

Why Do People Hold Stock?
Tracking the Stock Market

Explaining Stock Prices

The Stock Market and the Macroeconomy

How the Stock Market Affects
the Economy
How the Economy Affects the
Stock Market

What Happens When Things Change?

A Shock to the Economy
A Shock to the Economy *and*
the Stock Market: The 1990s
The Fed's Dilemma in the Late
1990s and Early 2000

In December 1996, Alan Greenspan—the chair of the Federal Reserve Board—uttered two sentences that caught the world's attention. Speaking to a Washington research organization, he asked, “How do we know when irrational exuberance has unduly escalated asset values which then become the subject of unexpected and prolonged contractions . . . ? And how do we factor that assessment into monetary policy?”

Greenspan was referring to the rapid rise in stock prices that had occurred over the previous several years. By one broad measure, the average stock's price had doubled over this period—a very rapid rise by historical standards. But when the markets opened for trading at 9:30 A.M. on the morning after Greenspan's speech, stock prices dropped by about 2 percent almost immediately.

That evening, on *Larry King Live* and *ABC News Nightline* and *CNN Moneyline*, pundits debated the meaning and wisdom of Greenspan's remarks. Everyone agreed that the purpose of Greenspan's remarks was to bring down stock prices, and that he had succeeded somewhat. But there were two opposing reactions to what the Fed chair had done. One group of commentators believed that Greenspan was making a mistake, that government officials have no business deciding when stock prices are too high or too low, and should leave the market alone. The other group believed that stock prices had, indeed, risen too high and too fast, and that Greenspan was entirely justified in trying to bring them down.

The debate that took place at the end of December 1996—and continued for the next several years—raises a number of questions. *Why* does the stock market matter? What is its role in the economy? Why should public officials worry when stock prices are too high? And why should two little words—“irrational exuberance”—uttered by one man rock the stock market and drive down share prices? In this chapter—after providing some basic background about the stock market—we'll answer all these questions.

BASIC BACKGROUND

Let's start with the most basic question of all: What is a share of stock?

First, a share of stock is a private financial asset, like a corporate bond. In fact, stocks and corporate bonds are alike in two ways. Both are issued by corpo-



rations to raise funds for investment projects, and both offer future payments to their owners.

But there is also an important difference between these two types of assets. When a corporation issues a bond, it is *borrowing* funds; the bond is just a promise to pay back the loan. A share of stock, by contrast, is a share of *ownership* in a corporation. When a firm issues new shares of stock, those who pay for those shares provide the firm with new funds, and in return, the firm owes them—at some future date or dates—a share of the firm's profits.

When a firm issues new shares of stock—in what is called a *public offering*—the sale of stock generates funds for the firm. Once the newly issued shares are sold, however, the buyer is free to sell them to someone else. Indeed, virtually all of the shares traded in the stock market are previously issued shares, and this trading does not involve the firm that issued the stock.

But a firm is still *concerned* about the price of its previously issued shares for two reasons. First, the firm's owners—its stockholders—want high share prices because that is the price they can sell at. A management team that ignores the desires of stockholders for too long might find itself replaced by other managers who will pay more attention.

Moreover, the price of previously issued shares has an important impact on firms that are planning new public offerings. That's because previously issued shares are perfect substitutes for the firm's new shares, so the firm cannot expect to receive a higher price for its new shares than the going price on its old shares. The higher the price for previously issued shares, the higher the price the firm will receive for *new* shares, and the more funds it will obtain from any given public offering.

WHY DO PEOPLE HOLD STOCK?

Stock ownership in the United States is growing rapidly. In 1983, only 19 percent of Americans owned shares of stock either directly or through mutual funds—companies that invest in a variety of stocks for their clients. By early 2000, the percentage of Americans who owned stock in these two ways had increased dramatically, to 48 percent. If we included stocks in employer-managed retirement accounts, the percentage of Americans with a stake in the stock market would be much higher. And the stakes are significant. By early 2000, the average U.S. household held more wealth in the stock market than in real estate, including the value of their own home.

Why do so many individuals choose to hold their wealth in stocks? You already know part of the answer: When you own a share of stock, you own part of the corporation. The fraction of the corporation that you own is equal to the fraction of the company's total stock that you own. For example, in April 2000, there were 497,476,000 shares outstanding in Southwest Airlines corporation (no relation to the publisher of this book). If you owned 10,000 shares of Southwest stock, then you owned $10,000/497,476,000 = 0.00002$, or about two-thousandths of a percent of the company. That means that you are, in a sense, entitled to two-thousandths of a percent of the firm's after-tax profits.

In practice, however, most firms do not pay out *all* of their profit to shareholders. Instead, some is kept as *retained earnings*, for later use by the firm. The part of profit that is distributed to shareholders is called *dividends*. A firm's dividend payments benefit stockholders in much the same way that interest payments benefit bondholders, providing a source of steady income. Of course, as part owner of a firm, you are part owner of any retained earnings as well, even if you will not benefit from them until later.

Aside from dividends, a second—and usually more important—reason that people hold stocks is that they hope to enjoy *capital gains*. A capital gain is the return someone gets when they sell an asset at a higher price than they paid for it. For example, if you buy shares of Southwest at \$15 per share, and later sell them at \$19 per share, your capital gain is \$4 per share. This is in addition to any dividends you earned while you owned the stock.

Some stocks pay no dividends at all, because the management believes that stockholders are best served by reinvesting all profits within the firm so that *future* profits will be even higher. The idea is to increase the value of the stock, and create capital gains for the shareholders when the stock is finally sold. New or fast-growing companies—such as Yahoo, America Online, and Microsoft—typically pay no dividends at all.

Over the past century, corporate stocks have generally been a good investment. Holding stocks was especially rewarding during the 1990s, as you'll see in the next section.

TRACKING THE STOCK MARKET

In the United States, financial markets are so important that stock and bond prices are monitored on a continuous basis. If you wish to know the value of a stock, you can find out instantly by checking with a broker or logging onto a Web site (such as Yahoo.com, Morningstar.com, or thomsoninvest.com). In addition, stock prices and other information are reported daily in local newspapers and in specialized financial publications such as the *Wall Street Journal* and the *Financial Times*.

In addition to monitoring individual stocks, the media keep a close watch on many stock market indices or averages. These averages track movements in stock prices as a whole, or movements in particular types of stocks. The oldest and most popular average is the *Dow Jones Industrial Average (DJIA)*, which tracks the prices of 30 of the largest companies in the United States, including AT&T, IBM, and Wal-Mart. Another popular average is the much broader *Standard & Poor's 500 (S&P 500)*, which tracks stock prices of 500 corporations chosen to represent all stocks in the market. Finally, the *NASDAQ* index tracks share prices of about 5,000 mostly newer companies whose shares are traded on the Nasdaq stock exchange—an association of stockbrokers who execute trades electronically. The companies in the NASDAQ include most of the new high-tech companies that are closely connected to the Internet sector.

Often, the three stock market averages will rise and fall at the same time, sometimes by the same percentage. That's because many of the shocks that hit the stock market affect most share prices *together*. But the indices can and do behave differently—sometimes very differently. For example, in early 2000, Internet stocks fluctuated wildly from day to day, as new information changed public opinion about the future of industry. There were many days on which the NASDAQ rose substantially while the Dow and the Standard & Poor's 500 fell, and vice versa.

Table 1 shows how the three averages performed over different lengths of time ending in December 31, 1999. The entries in the table tell us the average annual increase in each index over the period. For example, the entry 15.3 percent in the table (be sure you can find it) tells us that—over the period January 1, 1990 to December 31, 1999—the S&P 500 rose an average of 15.3 percent per year.

As you can see, while the S&P 500 and the DJIA have moved in tandem, NASDAQ has moved upward much more rapidly over the decade. You can also see that the 1990s were a good decade for stocks. Someone who invested \$10,000 in a typical group of S&P 500 stocks on January 1, 1990 would have been able to

TABLE 1

THE PERFORMANCE OF
THREE STOCK MARKET
INDEXES

Index	Increase, 1 Year Ending December 31, 1999	Average Annual Increase, 3 Years Ending December 31, 1999	Average Annual Increase, 5 Years Ending December 31, 1999	Average Annual Increase, 10 Years Ending December 31, 1999
Dow Jones Industrial Average	25.2%	21.3%	24.6%	15.4%
Standard & Poor's 500	19.5%	25.6%	26.2%	15.3%
NASDAQ	85.6%	46.6%	40.2%	24.5%

Source: Dow Jones Web site (http://averages.dowjones.com/dja_fact.html), accessed April 17, 2000, and author's calculations.

sell them for \$41,523 on December 31, 1999. And someone who had invested \$10,000 in a typical collection of NASDAQ stocks would have \$89,473 at the end of the period.

EXPLAINING STOCK PRICES

Why do stock prices change? And why do they change so often?

We can answer these questions—as we answer most questions about the economy—by using our four-step process.

KEY STEP #1: CHARACTERIZE THE MARKET

The price of a share of stock—like any other price—is determined in a market. But which market? Initially, we'll be focusing on price changes for shares of a particular stock, so the most useful way to organize our thinking is to look at the market for a single corporation's shares. That is, we'll view the "stock market" as a collection of *individual* markets, one for shares of stock in America Online, another for shares in Kmart, another for shares in Starbucks, and so on.

Further, we'll characterize the market for a company's shares as perfectly competitive. Indeed, markets for shares *do* satisfy the three requirements of perfect competition rather closely. There are many buyers and sellers (so many that no one of them can do much to change the market price of the stock).¹ There is a standardized product (it makes no difference to the buyer whether her Kmart shares are being sold to her by Smith or by Jones). And there is easy entry (virtually anyone with funds to invest can open up a brokerage account and buy or sell any publicly traded stock).

In sum,

we'll view the stock market as a collection of individual, perfectly competitive markets for particular corporations' shares.



Characterize the Market

¹ In some cases, a single buyer or seller holds such a large fraction of a company's shares that his or her decisions have a significant impact on market price. But these exceptions are rare for publicly traded shares.

Identify Goals and Constraints



KEY STEP #2: IDENTIFY THE GOALS AND CONSTRAINTS

In each market for shares, the buyers and sellers are both individuals and institutions (money market funds, insurance companies, and retirement funds). In either case, a high rate of return is a primary goal. Since most of the return on stocks comes from capital gains, we can state this goal very simply: Buyers will want to buy and hold shares in a company when they believe the stock price will rise, and sell shares in a company when they believe the price will fall.

In addition, stockholders are concerned about risk. All else equal, most of us would prefer to hold financial assets with relatively stable prices, rather than those whose prices fluctuate widely from day to day or month to month. Thus, when people choose *between* stocks and other financial assets, and when they choose *which* companies' shares to hold, they will look at risk as well as the expected rise in price.

In many cases, there is a trade-off between risk and return: The stocks that offer the highest expected return are also subject to the greatest risk. (If that weren't the case, few people would want to hold the riskier stocks.) Therefore, a part of every stockholder's goal is to strike the right balance between risk and return, based on their own attitudes toward risk.

What about constraints? Since stock holding is one way of holding wealth, individuals and institutions are constrained by the amount of wealth at their disposal. For a household, the constraint is the household's net worth. For a mutual fund, the constraint is the total amount of funds that households have contributed.

Stockholders are concerned about both the rate of return and the risk associated with stocks. In practice, they try to allocate their total wealth among a collection of assets—including stocks—that strikes the right balance between risk and return.

Find the Equilibrium



KEY STEP #3: FIND THE EQUILIBRIUM

Like all prices in competitive markets, stock prices are determined by supply and demand. However, in stock markets, our supply and demand curves require careful interpretations.

Figure 1 presents a supply and demand diagram for the shares of Southwest Airlines. Unlike most supply curves you've studied in this book—which show the quantity of something that suppliers want to *sell* over a given period of time—the supply curve in Figure 1 is somewhat different. It tells us the quantity of shares of Southwest stock *in existence* at any moment in time. This is the number of shares that people are *actually* holding.

On any given day, the number of Southwest shares in existence is just the number that the firm has issued previously. Therefore, no matter what happens to the price of the stock, the number of shares remains unchanged, and so does the quantity supplied. This is why the supply curve in the figure is a vertical line at 497 million: Over the time period we're analyzing, there are 497 million shares in existence regardless of the price.

Now, just because 497 million shares of Southwest stock exist, that does not mean that this is the number of shares that people *want* to hold. The desire to hold Southwest shares is given by the downward-sloping demand curve. As you can see, all else equal, the lower the price of the stock, the more shares of Southwest that people will want to hold. Why is this?

First, people have different expectations about the firm's future profits. Some may believe that Southwest will continue to grow as it has in the past. Others will think it is poised for a spurt of higher growth, while still others—more pessimistic—

THE MARKET FOR SHARES OF SOUTHWEST AIRLINES CORPORATION

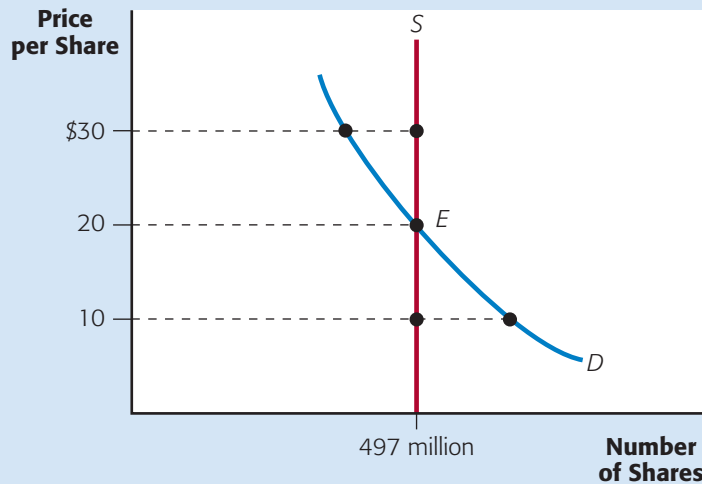


FIGURE 1

The supply curve shows the number of shares of Southwest Airlines stock people *are* holding. The curve is vertical at the number of shares outstanding—which was 497 million in early 2000. The demand curve tells us how many shares people *want* to hold. It slopes downward—the lower the price, the more shares people want to hold. At any price other than the equilibrium price of \$20, there would be either an excess supply or an excess demand for shares.

may believe that Southwest's best days are behind it. There will also be different opinions about the *risk* of those future profits. Thus, at any given moment, with an array of opinions about the company's future, each person will have a different price in mind that would make the stock an attractive buy. As the price per share falls, more and more people will find the stock to be a bargain, and want to hold it. This is what the downward-sloping demand curve tells us.

In the figure, you can see that at any price other than \$20 per share, the number of shares people *are* holding (on the supply curve) will differ from the number they *want* to hold (on the demand curve). For example, at a price of \$10 per share, people would want to hold more shares than they are currently holding. Many would try to buy the stock, and the price would be bid up. At \$30 per share, the opposite occurs: People find themselves holding more shares than they want to hold, and they will try to get rid of the excess by selling them. The sudden sales would cause the price to drop. Only at the *equilibrium price* of \$20—where the supply and demand curves intersect—are people satisfied holding the number of shares they are *actually* holding.

Stocks achieve their equilibrium prices almost instantly. There are so many stock traders—both individuals and professional fund managers—poised at their computers, ready to buy or sell a particular firm's shares at a moment's notice, that any excess supply or excess demand will cause the price to move within seconds. Thus, we can have confidence that the price of a share at any time is the equilibrium price.

But why do stock prices *change* so often? To answer that question, we need Key Step #4.

KEY STEP #4: WHAT HAPPENS WHEN THINGS CHANGE?

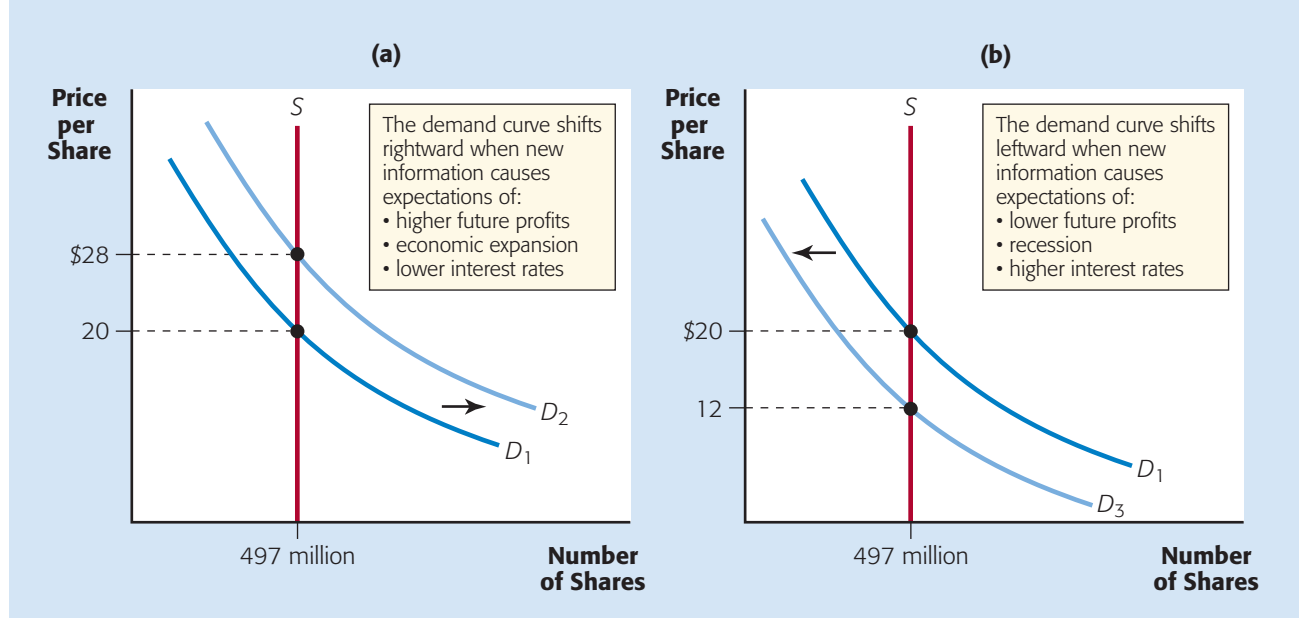
The *supply* curve for a corporation's shares, like the one in Figure 1, shifts rightward whenever there is a public offering. Can this explain changes in share prices? Not really. Public offerings occur only occasionally and with great fanfare. Moreover, most public offerings by existing companies are for a relatively small number of shares. They shift the supply curve only a little, and therefore have little impact on the



What Happens When Things Change?

FIGURE 2

SHIFTS IN THE DEMAND FOR SHARES CURVE



market price of the stock. Thus, the changes in equilibrium prices we observe for most stocks are *not* caused by shifts of the supply curve.

That leaves only one explanation: shifts in *demand*.

The changes we observe in a stock's price—over a few minutes, a few days, or a few years—are virtually always caused by shifts in the demand curve.

Panel (a) of Figure 2 shows how a rightward shift of the demand curve for shares of Southwest Airlines could cause the equilibrium price to rise to \$28 per share. Indeed, on rare occasions, the demand curve for a firm's shares has shifted so far rightward in a single day that the share price doubled or even tripled.

But what causes these sudden changes in demand for a share of stock?

In almost all cases, it is one or more of the following three factors:

1. *Changes in expected future profits of the firm.* At any given time, people have an idea about the expected profits of every firm. But these ideas can change as new information becomes available. The new information can pertain to a scientific discovery, a corporate takeover or merger with another company, or a new government policy. Even information that suggests that one of these events *might* occur can change the attractiveness of stocks. After all, a stock in a company that has a 50 percent chance of making huge profits from a new invention is more attractive than a stock that has only a 20 percent chance of such profits.

New information can be positive—shifting the demand curve rightward and increasing the price of the stock. But it can also be negative, shifting the demand curve leftward. A dramatic example of the latter occurred on March 14, 2000. On that day, President Clinton and British Prime Minister Blair issued a joint statement that they would work to make data from the human genome publicly available. Some observers interpreted the statement to indicate a possible tilt in public policy. Perhaps the government was suggesting that it would work to eliminate or shorten the duration of gene-based patents, vastly reducing the future profits of biomedical re-

search companies that held those patents. Within minutes of the statement, demand curves for shares of biotech companies shifted leftward, and share prices plummeted—some by as much as 30 percent.

Any new information that increases expectations of firms' future profits—including announcements of new scientific discoveries, business developments, or changes in government policy—will shift the demand curves of the affected stocks rightward. New information that decreases expectations of future profits will shift the demand curves leftward.

2. *Macroeconomic fluctuations.* When the economy is expanding, and real GDP is rising, firms *in general* tend to earn higher profits, and these profits are less risky. By contrast, in a recession, sales and profits decrease. For this reason,

any news that suggests the economy will enter an expansion, or that an expansion will continue, will shift the demand curves for most stocks rightward. Any news that suggests an economic slowdown or a coming recession shifts the demand curves for most stocks leftward.

3. *Changes in the interest rate.* Stocks are not the only way that people can hold their wealth. They can also hold money and—more importantly—they can hold interest-earning assets like certificates of deposit or bonds. If the interest rate rises, these other assets become more attractive, and many people will want to shift their wealth *out* of stocks so they can buy them. Thus,

a rise in the interest rate in the economy will shift the demand curves for most stocks to the left. Similarly, a drop in the interest rate will shift the demand curves for most stocks to the right.²

Even *expectations* of a future interest rate change can shift demand curves for stocks. This can create some rather convoluted—but logical—explanations for movements in stock prices. For example, suppose that a report comes out suggesting that real GDP is growing very rapidly. All else equal, this makes stocks more attractive. But . . . all else may *not* remain equal. In fact, you may surmise that the U.S. Federal Reserve and its influential chair—currently Alan Greenspan—want to prevent inflation at almost any cost. You might then *anticipate* that the Fed—concerned about the economy overheating—will raise interest rates in the near future to slow down the growth in real GDP. You also know that—if the interest rate *does* rise—stock prices will fall, for the reasons we've just discussed. What should you do? *Dump your stocks now*, to avoid a capital loss later. Since you and many others will no doubt have access to the same information, and feel the same way, the announcement of rapid economic growth could lead—almost immediately—to a *decrease* in stock prices.

Similarly, bad news about economic growth—if it leads to an expected decrease in interest rates—can cause stock prices to rise.

News that causes people to anticipate a rise in the interest rate will shift the demand curves for stocks leftward. Similarly, news that suggests a future drop in the interest rate will shift the demand curves for stocks rightward.

² If you've studied *microeconomics*, you've learned another way to view the impact of interest rate changes on stock prices: Higher interest rates reduce the *present value* of any given stream of future profits.

Panels (a) and (b) of Figure 2 summarize the different forces that cause the demand curve for a stock to shift rightward or leftward.

THE STOCK MARKET AND THE MACROECONOMY

As you can see in Figure 3, there is a *two-way* relationship between the stock market and the economy. That is, the performance of the stock market affects the performance of the economy, and vice versa. In the next two sections of this chapter, we'll look at this two-way relationship. Let's start with the impact of the stock market on the economy, as illustrated by the upper arrow in the figure.

HOW THE STOCK MARKET AFFECTS THE ECONOMY

On October 19, 1987, there was a dramatic drop in the stock market. That day, the Dow Jones Industrial Average fell by 508 points—a drop of 23 percent—and about \$500 billion in household wealth disappeared. That same evening, as President Reagan boarded his helicopter, a breathless Sam Donaldson of ABC News thrust a microphone in front of him and asked, “Mr. President, are you concerned about the 500-point drop in the Dow?” As Reagan entered his helicopter, he smiled calmly and replied, “Why, no, Sam. I don't own any stocks.”

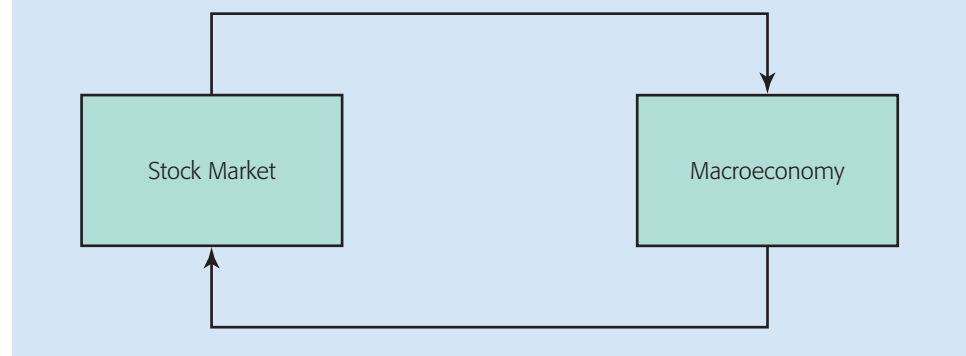
It was a curious exchange. Reagan was probably kidding—perhaps trying to calm a worried nation with his trademark humor. Or perhaps he was annoyed at a frantic reporter invading his personal space. Or he might have been caught off guard and said the first thing that popped into his head.

Whatever Reagan's intent, the statement was startling because, in fact, the stock market *does* matter to all Americans, whether they own stocks or not. As you are about to see, the ups and downs of stock prices—if they are big enough and sustained enough—can cause ups and downs in the overall economy.

The Wealth Effect. To understand how the market affects the economy, let's run through the following mental experiment: We'll suppose that, for *some* reason (we'll discuss specific reasons later), stock prices rise. As a result, those who own stock will feel wealthier. In fact, they *are* wealthier. After all, just as you measure the value of your house by the price at which you could sell it, the same is true of your financial assets, like stocks. When stock prices rise, so does household wealth.

FIGURE 3

THE TWO-WAY RELATIONSHIP BETWEEN THE STOCK MARKET AND THE ECONOMY



What do households do when their wealth increases? Typically, they increase their spending. In our short-run macro model, we would classify this as an increase in *autonomous consumption*—an increase in consumption spending at *any* level of disposable income.

The link between stock prices and consumer spending is an important one, so economists have given it a name: the *wealth effect*. And the wealth effect works in both directions: Just as an increase in stock prices increases autonomous consumption, so will a drop in stock prices—which decreases household wealth—cause autonomous consumption spending to fall.

More generally,

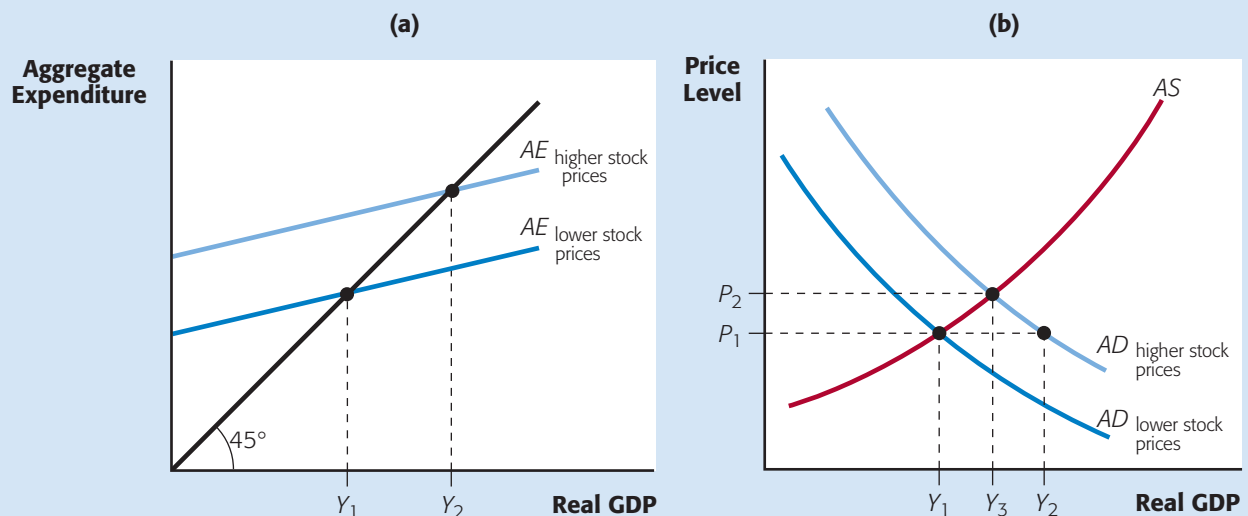
the wealth effect tells us that autonomous consumption spending tends to move in the same direction as stock prices. When stock prices rise, autonomous consumption spending rises; when stock prices fall, autonomous consumption spending falls with it.

The Wealth Effect and Equilibrium GDP. As you learned when you studied the short-run macroeconomic model, autonomous consumption is a component of total spending. And an increase in total spending tends to increase equilibrium real GDP, as shown in panel (a) of Figure 4. There, when stock prices rise, the increase in real wealth causes the aggregate expenditure line to shift upward, and increases the economy's equilibrium GDP from Y_1 to Y_2 .

Panel (b) of Figure 4 shows a more complete way to view the impact of rising stock prices. In this panel, the increase in equilibrium GDP at any given price level is shown as a rightward shift in the economy's *AD* curve. And—in the absence of

THE EFFECT OF HIGHER STOCK PRICES ON THE ECONOMY

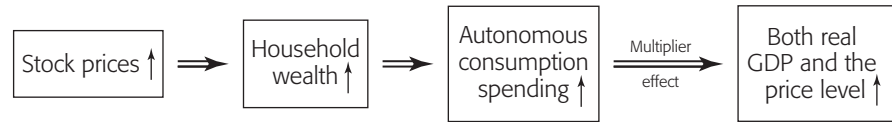
FIGURE 4



Higher stock prices have a wealth effect on spending, increasing consumption spending at any level of real GDP. In panel (a), the wealth effect of higher stock prices shifts the aggregate expenditure line upward, raising equilibrium GDP from Y_1 to Y_2 . Panel (b) shows a more complete way of illustrating the wealth effect: Higher stock prices shift the aggregate demand curve rightward, increasing both equilibrium real GDP and the price level.

any change in government policy—this shift of the AD curve will increase both equilibrium GDP (to Y_3) and the price level (to P_2). (Why does equilibrium GDP increase by less in panel (b) than in panel (a)?)

We can summarize the logic of the wealth effect as follows:



In words:

Changes in stock prices—through the wealth effect—cause both equilibrium GDP and the price level to move in the same direction. That is, an increase in stock prices will raise equilibrium GDP and the price level, while a decrease in stock prices will decrease both equilibrium GDP and the price level.

How important is the wealth effect? Economic research shows that the *marginal propensity to consume out of wealth*—the change in consumption spending for each one-dollar rise in wealth—is between 0.03 and 0.05. In other words, when household wealth rises by a dollar, all else remaining the same, consumption spending tends to rise by between 3 and 5 cents. Moreover, recent research suggests that virtually *all* of the increase in consumption comes rather quickly—within one quarter (3 months) after the quarter in which stock prices rise.³ Let's translate this into some practical numbers.

First, as a rule of thumb, a 100-point rise in the DJIA—which generally means a rise in stock prices in general—causes household wealth to rise by about \$100 billion. This rise in household wealth, we've now learned, will increase autonomous consumption spending by between \$3 billion and \$5 billion—we'll say \$4 billion. As you learned several chapters ago, the multiplier in the real world—after we take account of all the automatic stabilizers that reduce its value—is equal to about 1.5, with most of its impact in the first nine months to a year after a shock. Thus, a 100-point rise in the DJIA, which causes consumption spending to rise by about \$4 billion, will cause real GDP to increase by about \$4 billion \times 1.5 = \$6 billion. Extrapolating from these results, a 6,000- or so point rise in the Dow—such as we saw in the second half of the 1990s—would generate about \$240 billion in additional consumption spending, and drive up real GDP by about \$360 billion—an increase of about 4 percent. This is in addition to the normal rise in real GDP that would be occurring anyway, as income grows and spending grows with it. Thus,

rapid increases in stock prices—such as those that have occurred over the past five years—can cause significant positive demand shocks to the economy, shocks that policy makers cannot ignore. Similarly, rapid decreases in stock prices can cause significant negative demand shocks to the economy, which would be a major concern for policy makers.

³ Sydney Ludvigson and Charles Steindel, "How Important Is the Stock Market Effect on Consumption?" New York Federal Reserve Bank *Policy Review*, July 1999. (Also available at http://www.ny.frb.org/rmaghome/econ_poll/799lud.htm.)

HOW THE ECONOMY AFFECTS THE STOCK MARKET

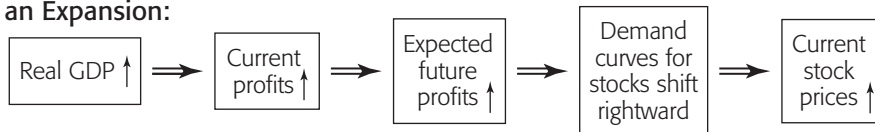
Now that we've explored how the stock market affects the economy, let's look at the other side of the two-way relationship: how the economy affects stock prices.

Actually, many different types of changes in the overall economy can affect the stock market. Some—like the revolution in telecommunications that took place in the 1990s—are rare, happening once or twice a century. Others—like the impact of macroeconomic fluctuations—happen much more frequently. In this section, we'll focus on the more frequent scenario: how the stock market responds as the economy goes through expansions and recessions in the short run.

Let's start by looking at the typical expansion, in which real GDP rises rapidly over several years. In the typical expansion, profits will rise along with GDP. Higher profits are themselves enough to make stocks look more attractive. But the process is further helped by another factor: an improvement in investor psychology. In an expansion, not only are corporate profits rising, but also the unemployment rate falls, and household incomes rise. Memories of the last recession are dim, and it looks as if the economy will continue to grow and grow, perhaps forever. This optimistic outlook raises estimates of future profits—sometimes dramatically. The demand curves for stocks will shift rightward, and stock prices will rise.

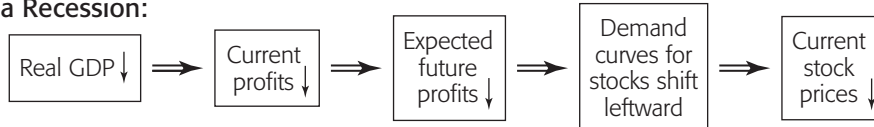
We can summarize the impact of an expansion on the market as follows:

In an Expansion:



Of course, the process also works in reverse. When a recession strikes, corporate profits drop, unemployment rises, and the economy begins to look bleak. Stockholders turn pessimistic, and expect lower profits in the future. The demand curves for stocks shift leftward, driving stock prices down:

In a Recession:



In sum,

in the typical expansion, higher profits and stockholder optimism cause stock prices to rise. In the typical recession, lower profits and stockholder pessimism cause stock prices to fall.

WHAT HAPPENS WHEN THINGS CHANGE?

Now that you understand how stock prices affect the overall economy, and how the economy can affect stock prices—it's time to apply Key Step #4 one more time. But this time, we'll apply it very broadly: We'll observe how *both* the stock market and the macroeconomy are affected when *something* changes.

But . . . *what* changes?

FIGURE 5

THREE TYPES OF SHOCKS

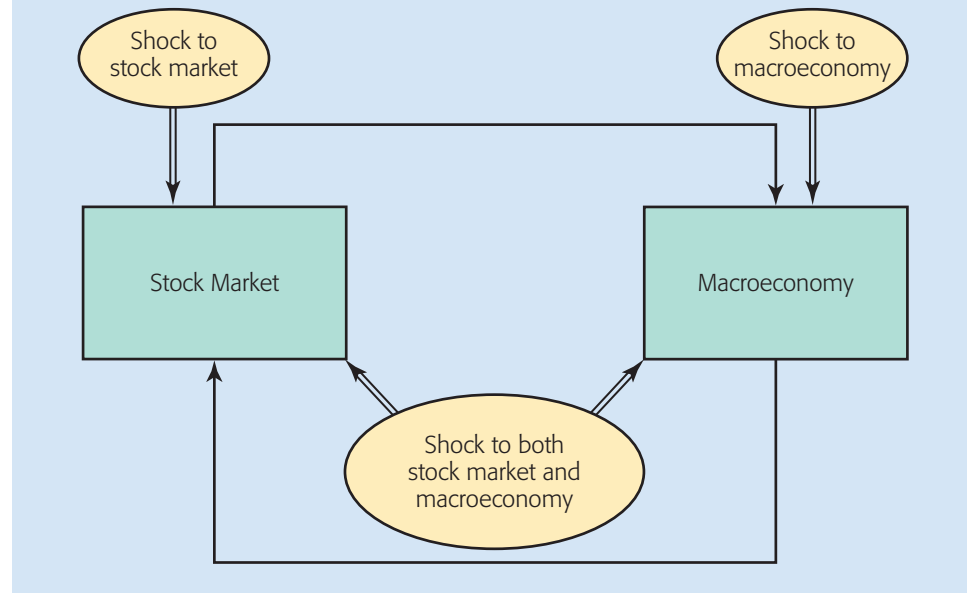


Figure 5 illustrates three different types of changes we might explore. A change might have most of its initial impact on the overall economy, rather than the stock market. For example, a change in government spending or taxes—with an unchanged interest rate target by the Fed—would initially affect real GDP, rather than the stock market. Ultimately, stock prices would be affected, but primarily *through* the change in real GDP.

Alternatively, there might be a shock that initially affects the stock market. An example would be a change in the duration of patent protection for intellectual property, which would change the expected profits of firms and shift the demand curves for stocks.

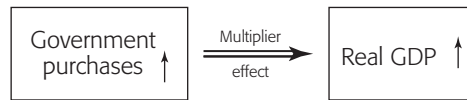
Finally, a shock could have powerful, initial impacts on *both* the stock market *and* the overall economy. An example is the technological revolution of the late 1990s and early 2000s, which has rocked both the economy and the stock market.

In the next section, we'll explore the consequences of an initial shock to the *economy*. Then, we'll turn our attention to a shock that simultaneously hits the market and the economy, as occurred during the 1990s. Finally, in the end-of-chapter questions, you'll be asked to address the remaining case: a shock that initially hits just the stock market.

A SHOCK TO THE ECONOMY

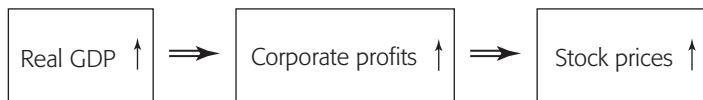
Imagine that new legislation greatly increases government purchases—say, to equip public schools with more sophisticated telecommunications equipment, or to increase the strength of our armed forces. This spending shock—and increase in government purchases—will have its primary initial impact on the overall economy, rather than the stock market. Let's suppose, too, that the Fed maintains its interest rate target, so there is no direct impact on the stock market from changes in the interest rate. What will happen?

As you've learned in your study of macroeconomics, the rise in government purchases will first increase real GDP through the expenditure multiplier:

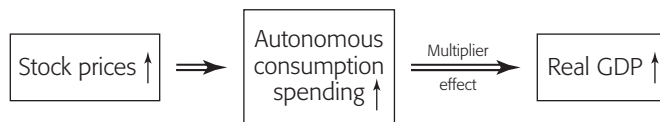


In previous chapters, the multiplier process was a simple one: Increases in output and income cause increases in consumption, which cause further increases in output and income and then further increases in consumption, and so on. But now, there is a *new* contributor to the multiplier process: the stock market.

First, remember that increases in real GDP cause increases in corporate profits. This, in turn, leads to investor optimism about *future* profits, shifting the demand curves for stocks rightward, and increasing the average price of stocks.



But the story doesn't end there. The increase in stock prices will—through the wealth effect—cause an increase in autonomous consumption spending. Note that this is *in addition* to the increase in consumption spending caused by the normal multiplier process. Indeed, the increase in autonomous consumption spending caused by the wealth effect sets off its *own* multiplier process, further increasing real GDP:



Now, look back at the three cause-and-effect chains just presented. You can see that an increase in government purchases will cause a *larger* rise in real GDP when we include the effects of the stock market. Another way of saying this is that,

when we include the effects of the stock market, the expenditure multiplier is larger. An increase in spending that increases real GDP will also cause stock prices to rise, causing still greater increases in real GDP. Similarly, a decrease in spending that causes real GDP to fall will also cause stock prices to fall, causing still greater decreases in real GDP.

When you first learned about the multiplier, you learned about automatic stabilizers—features of the economy, such as the income tax or unemployment insurance payments, that make the expenditure multiplier smaller, and thus help to stabilize real GDP. Now you can see that the normal behavior of the stock market—which makes the expenditure multiplier larger—works as an *automatic destabilizer*. This is one reason why stock prices are so carefully watched by policy makers, and matter for everyone—whether they own stocks themselves or not.

A SHOCK TO THE ECONOMY AND THE STOCK MARKET: THE 1990s

The 1990s—especially the second half of the 1990s—saw a dramatic rise in stock prices. Both the Dow Jones Industrial Average and the Standard & Poor's 500 more than quadrupled over the period, and the NASDAQ increased almost ninefold.

The 1990s were also a period of rapid expansion, especially the period from 1995 to 1999, in which economic growth averaged 4.2 percent per year—much faster than in previous decades.

In part, the economic expansion and the rise in stock prices were reinforcing: each contributed to the other, as we've seen in this chapter. But the expansion and the climb in stocks were *initiated* by a common shock: a technological revolution, led by the Internet.

The Internet had a direct impact on the stock market through its effect on expected future profits of U.S. firms. In particular, stockholders (and potential stockholders) believed that the new technology would enable firms to produce goods and services at much lower costs than before, and that this reduction in costs would translate into an increase in profit. The increase in expected future profits translated into a rightward shift of the demand curve for stocks at virtually any firm that had the potential to exploit the new technology, or help other firms exploit it.

For example, during this period, AT&T positioned itself to be a major supplier of information and voice communication using the Internet and other new technologies. As a result, the demand curve for AT&T stock shifted rightward—enough to drive the price of AT&T stock from \$28 per share in early 1997 to \$58 by the end of 1999.

At the same time, the technological revolution was having a huge impact on the overall economy. Investment spending rose, as business firms—in order to take advantage of the new technology—invested in new plant and equipment. Autonomous consumption spending also rose: consumers wanted new gadgets that would enable them to enjoy new types of services—new cellular phones, new computers, palm pilots, high-speed Internet connections, and more.

Faced with these demand shocks, the Federal Reserve would ordinarily have raised its interest rate target to prevent real GDP from exceeding potential output. But the technological revolution of the 1990s was having *another* effect on the economy: It *increased* potential output more rapidly than before. New computers, new software, and other forms of new capital equipment—along with the increased skills and training of the workforce—raised the typical worker's hourly output by about 22 percent over the decade.

The technological changes of the 1990s were an example of a shock to both the stock market *and* the economy. But remember that each of these also influences the other. As the expansion gained steam and real GDP was growing steadily and rapidly, profits and expected profits soared, pushing stock prices up further. And as stock prices rose, the wealth effect worked to propel consumer spending still higher. The result was a market and an economy that were feeding on each other, sending both to new performance heights. Was this a good thing?

Yes, and no. Higher stock prices certainly make stockholders happy. And a rapid expansion is good for workers, since it makes it easy to find jobs and forces firms to compete for workers by offering higher wages and better fringe benefits. Indeed, from a high of almost 8 percent in 1991, the unemployment rate dropped steadily during the 1990s, reaching 4 percent at the end of the decade.

But in spite of all this good news, there were dark clouds on the horizon . . . at least from the Fed's point of view.

THE FED'S DILEMMA IN THE LATE 1990s AND EARLY 2000

As stocks soared during the late 1990s, many people began to wonder: Did the realities of the late-1990s economy justify the heights to which stock prices had risen? Clearly, many people in the market thought so, or they wouldn't have been willing

to hold stocks at those high prices. But around 1995 and 1996, others—including some officials at the U.S. Federal Reserve—began to worry that share prices were rising out of proportion to the future profits they would be able to deliver to their owners. The Fed was worried that the market was experiencing a speculative *bubble*—a frenzy of buying that encouraged people to buy stocks and drive up their price because . . . well, just because their prices were rising.

In this view, the market in the late 1990s resembled the stock market in the 1920s, which is also often considered a bubble. While there were indeed reasons for optimism in the 1920s, there also seemed to be a speculative frenzy: Many investors borrowed money to buy stocks in companies they knew nothing about, just because of an anonymous tip or because they were watching the price of the stock go up. Indeed, the Dow Jones Industrial Average almost quadrupled from early 1920 to September 1929—just as it did during the 1990s. But when the bubble burst, it burst hard. From September 3, 1929 to July 8, 1932, the DJIA fell from 381 to 41—about a 90 percent decline.⁴ Many stocks of the most reliable and successful corporations (the so-called “blue chip” corporations) fell to only tiny fractions of their highs. General Electric stock, for example, fell from a high of 396¼ in 1929 to 8½ in 1932; Bethlehem Steel from 140¾ to 7¼; and RCA from 101 to 2½. Millions of people were financially wiped out—in itself, a human tragedy.

In 1996, when Alan Greenspan first made his “irrational exuberance” speech, he seemed to side with those who believed that the stock market was in the midst of a speculative bubble. His fear was that when the bubble burst—when people realized that there weren’t sufficient buyers to keep propping up stock prices out of proportion to their future profits—then stock prices would come plummeting down to earth. And a burst bubble would be painful—millions of people would lose substantial amounts of wealth. Moreover, the Fed would be forced to intervene to prevent the wealth effect—this time in a negative direction—from creating a recession.

Could the Fed do so? Probably. We understand how the economy works much better today than we did in 1929, when—in retrospect—the Fed made several mistakes after the stock market crashed. But the Fed’s knowledge isn’t perfect and—as you’ve learned—Fed intervention is still fraught with uncertainty. There is always a chance the Fed will react too strongly, or not enough. From the Fed’s point of view in the mid-1990s, the best economy would be one that hummed along without needing any policy intervention. That is, an economy with stock prices rising steadily and *slowly*, rather than a bubble that might burst and require a big policy shift.

In the mid-1990s, Greenspan seemed to be trying to “talk the market down” by letting stockholders know that he thought share prices were too high. The implied threat: If stocks rose any higher, the Fed would raise interest rates and bring them down. Indeed, according to many observers, merely hinting that the Fed *might* raise interest rates was designed to keep stock prices from rising too rapidly, and perhaps bring them down gently.

Only it didn’t work. While Greenspan’s irrational exuberance speech did bring the market down for a day or so, the relentless rise in stock prices continued. In October 1996, just before Greenpan’s speech, the DJIA stood at about 6,500. By March 1999—less than three years later—it had reached 10,000.

Not only were Greenspan’s efforts to “talk the market down” unsuccessful, they were also widely criticized. In the view of his critics, the value of stocks should be based on the decisions of those who buy and sell them. If people believe that a

⁴ The Dow Jones Industrial Average measures *nominal* stock prices. Since the price level decreased over this period, the decline in *real* stock prices was less than 90 percent, but still a substantial loss.

company is onto something good and that its future profits justify a doubling or tripling of its stock price within a short time, what business is it of the Fed to say they are wrong? After all, stock buying—and the funds it has made available to American corporations—is partly responsible for the remarkable rise in U.S. living standards over the past century. Moreover, the stock market has been especially effective in funneling funds to good ideas and away from bad ones because it relies on *decentralized* decision making. Those who put their money at risk decide for themselves what is and is not a good idea.

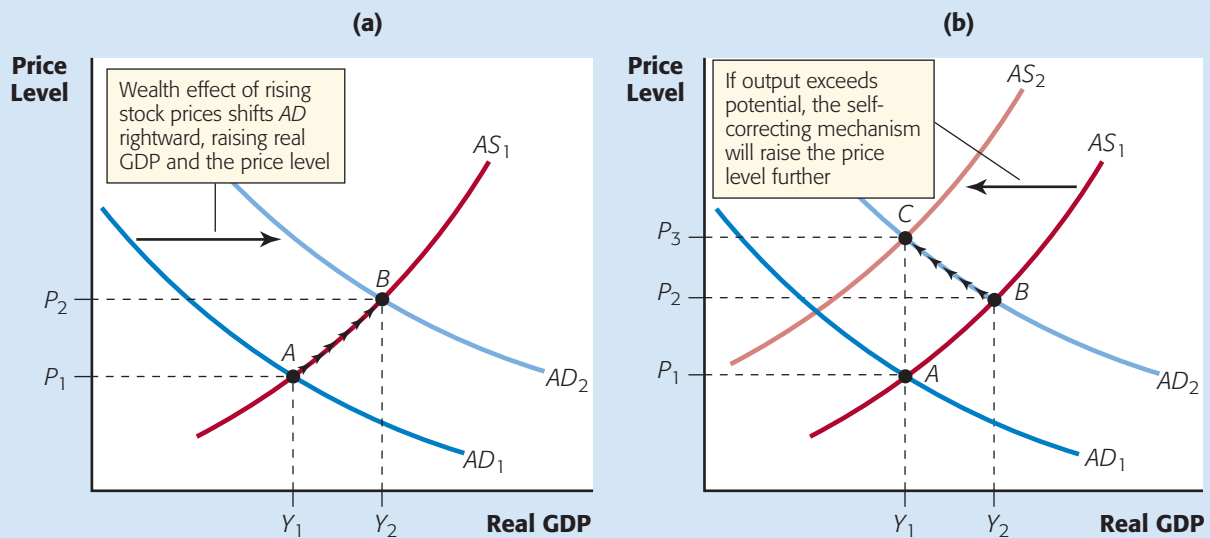
Greenspan himself seemed to change his tune as the 1990s continued. By 1998, he had stopped referring to exuberance—rational or irrational. Instead, he began to stress the remarkable changes in the economy, the rapid rise in productivity and potential output, and the fact that the American people—who buy and sell stocks—have a certain wisdom that should not be second-guessed by government officials. It was almost a complete reversal.

But as the 1990s came to a close, and the stock market continued to soar, the Fed faced a new problem: *the wealth effect*. Justified or not, share prices had continued to rise—and they rose a lot. In the two and a half years after Greenspan's famous irrational exuberance remarks in 1996, about \$3 trillion in new wealth was created. Consumer spending was rising dramatically, and the Fed began to worry that the economy might be exceeding—or would soon exceed—its potential output.

Figure 6 shows one way we can view the Fed's problem: with aggregate demand and supply curves. In panel (a), the wealth effect of rising stock prices shifts the aggregate demand curve from AD_1 to AD_2 , causing an increase in real GDP from Y_1 to Y_2 along with a rise in the price level. The question is: What happens next? That depends on where our potential output is relative to Y_2 . If Y_2 is greater than potential output, the self-correcting mechanism will begin to work: The price level will rise further, bringing the economy back to potential output (assumed to be Y_1 in the figure). This is something the Fed has worked hard to avoid. As you've learned, inflation—once it begins—tends to be self-perpetuating. People begin to expect it. And once the

FIGURE 6

THE FED'S PROBLEM: AN AS-AD VIEW



inflation is embedded in the economy, eradicating it is painful: The Fed would have to raise interest rates and slow the economy by more than would have been necessary to prevent the inflation in the first place. Moreover, in the past, efforts to bring the inflation rate down have triggered deep recessions. From the Fed's point of view, preventing inflation in the first place is always the preferred alternative.

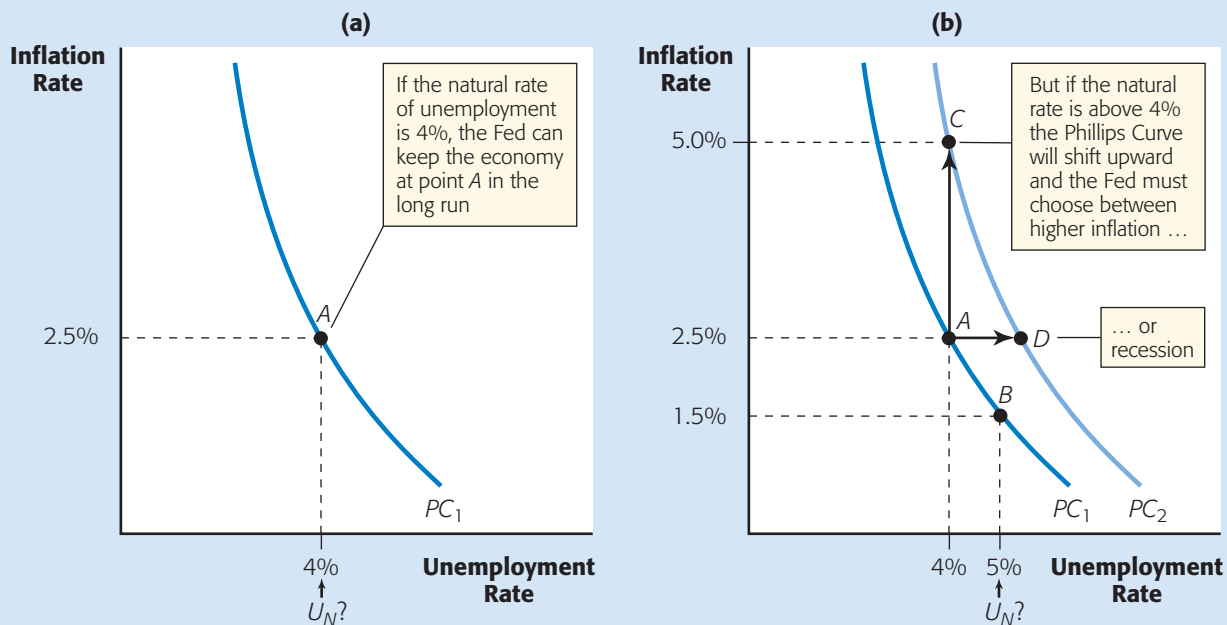
Figure 6 is useful, but it has a serious limitation: It doesn't take account of the rise in potential output. Each year, potential output increases because the population is growing, and because productivity—output per worker—is growing. In the 1990s and through early 2000, potential output was growing even more rapidly than in previous decades. We could illustrate this on an AS-AD diagram by shifting the AS curve rightward and downward over time. That is, due to changes in population and productivity, we could produce more output at any given price level, or have a lower price level at any given level of output. With a shifting AS curve, the Fed's goal is to shift the AD curve rightward each year by just enough to prevent inflation.

But the Phillips curve can illustrate the Fed's goal more easily. Look first at panel (a) of Figure 7, where the position of the economy in late 1999 and early 2000 is represented by point A on the Phillips curve PC_1 : 4 percent unemployment and a 2.5 percent annual inflation rate. As you learned a few chapters ago (Inflation and Monetary Policy), the Fed can *keep* the economy at point A only if the actual unemployment rate, 4 percent, is also the *natural* rate of unemployment. In that case, the Fed—by keeping the economy at point A—would be allowing actual output to rise each year by just enough to keep it equal to potential output.

But what if the natural rate of unemployment is *greater* than 4 percent—say, 5 percent? Then—as you can see in panel (b)—the economy would need to operate at point B to be at the natural rate. Point A now represents an overheated economy, with output greater than potential output. If we remain at point A, then over time

THE FED'S PROBLEM: A PHILLIPS CURVE VIEW

FIGURE 7



the entire Phillips curve would shift upward, to the curve labeled PC_2 , and point A would no longer be an option. If the Fed tried to maintain a 4 percent unemployment rate, the economy would then be at point C , with a rise in the inflation rate to 5 percent. To prevent any rise in inflation, the Fed would have to engineer a recession, bringing the economy to point D with unemployment above the natural rate.

To keep inflation low and stable without needing corrective recessions, the Fed strives to maintain unemployment at its natural rate. But no one—including the Fed—knows what the natural rate of unemployment *is* during any given year. We know that it is lower today than in the early 1990s—when it was believed to be about 5.5 or 6 percent. But no one knows how far it has fallen since then.

You might think that the Fed can estimate the natural rate by a process of trial and error—bringing the unemployment rate to a certain level (such as 4 percent) and seeing what happens to inflation. Then, if the inflation rate rises, the Fed would know that it had underestimated the natural rate. It could then slow the economy and bring unemployment back up to the natural rate.

Unfortunately, things are not so simple. In the real world, when the economy overheats, the inflation rate begins to rise only after a lag of several quarters or longer. By the time the Fed notices a rise in the inflation rate, the economy may have been overheated for months, and the Fed may have to take even more drastic action to bring us back to potential output. For this reason, the Fed looks ahead, and determines whether *current* economic conditions are likely to raise the inflation rate in the *future*.

And that is just what the Fed did beginning in mid-1999. With the unemployment rate near 4 percent, the economy growing at a rapid 3.8 percent clip for a year, and the stock market continuing to rise to record levels, Fed officials believed that the wealth effect would overheat the economy if nothing were done. So the Fed took action, even though inflation was still low and stable. From June 1999 through May 2000, the Fed raised its target for the federal funds rate six times: from 4.75 percent to 6.5 percent.

As this is being written (May 2000), the Fed's caution seems to have been warranted. In spite of the rise in interest rates, the unemployment rate remained at 4 percent. And far from slowing down, the growth rate of real GDP increased in the second half of 1999 and the first quarter of 2000. This suggests that if the Fed had *not* raised interest rates, the economy would, indeed, have overheated.

But there is a problem with the Fed's approach. Raising interest rates to rein in the economy can *also* bring down stock prices. Thus the Fed, in trying to steer the economy, can be accused of trying to regulate stock prices. Indeed, the Fed itself seems to regard stock prices as one of its tools for steering the economy: Higher interest rates decrease (or slow the rise in) stock prices, and thus slow spending through the wealth effect. Thus, we are back to the debate of the mid-1990s: Who should be setting the general level of share prices—the millions of stockholders who buy and sell shares, or the Federal Reserve?

PROBLEMS AND EXERCISES

1. The chapter contains the following statement: "You can also see that the 1990s were a good decade for stocks. Someone who invested \$10,000 in a typical group of S&P 500 stocks on January 1, 1990 would have been able to sell them for \$41,523 on December 31, 1999." Using information in Table 1, demonstrate that this statement is correct.
2. Suppose the corporate profits tax rate is reduced, and other taxes in the economy are increased by just enough to leave total tax revenue unchanged. Thus, the economy's equilibrium real GDP is unaffected, at least initially.
 - a. Would this event have any impact on the stock market? Illustrate, using supply and demand curves for

- the typical stock. What will happen to the price of the typical stock?
- Using cause-and-effect diagrams like the ones in this chapter, show how this change in tax policy would first affect the stock market, then affect the economy, and then create feedback effects in the stock market.
 - When we include feedback effects from the macro-economy, is the ultimate effect on stock prices greater or smaller than the initial impact in (a) above? Explain.
- Sometimes corporations will use their profits to buy back their own shares.
 - Explain why this action—using funds that could have been given to shareholders as dividends—might actually benefit shareholders. (*Hint:* Draw a supply and demand diagram for the corporation's shares. Which curve is affected by a stock buy-back?)
 - In the United States, the tax rate on long-term capital gains (capital gains on assets held longer than one year) is lower than the tax rate on ordinary income, including income from dividends. Does this help explain why corporations sometimes buy back their own shares? Explain.
 - Suppose that, over time, people become more sophisticated about changes in stock prices. Specifically, they realize that while stock prices go down in a recession, they tend to rise when the recession ends. Would this change the way the economy affects stock prices? Would it change our view of the stock market as an automatic destabilizer over the business cycle? Explain.
 - Classify each of the following events as a shock that initially affects (a) primarily the economy; (b) primarily the stock market; (c) both the stock market and the economy. Justify your answer in each case.
 - Government spending increases, while the Fed leaves its interest rate target unchanged.
 - The Fed—beginning to worry about inflation—increases its interest rate target.
 - In the section “*A Shock to the Economy*,” we explored the impact of an increase in government purchases with an unchanged interest rate target by the Fed.
 - In order to maintain an unchanged interest rate target, will the Fed have to increase or decrease the money supply? Illustrate with a diagram of the money market.
 - Suppose the Fed instead decides to pursue a completely passive monetary policy—leaving the *money supply* unchanged. What will happen to the interest rate? (Illustrate with another diagram of the money market.)
 - Under a passive monetary policy, does the change in government spending have more or less of an initial impact on the stock market (compared to the policy of maintaining an unchanged interest rate target)?
 - When population and productivity are increasing, potential output increases each year. One way to illustrate this is to shift the economy's *AS* curve rightward (and downward) each year. Assume that there is no expected inflation embedded in the economy, so that the ongoing inflation rate is zero. Using *AS-AD* diagrams, illustrate each of the following scenarios.
 - Potential output is increasing, and the Fed allows actual output to rise just enough to keep up with potential.
 - Potential output is increasing, and the Fed allows actual output to rise above potential.
 - Potential output is increasing, and the Fed allows so little growth that output falls below potential.
 - Using a diagram similar to panel (b) of Figure 7, show what happens over time if the Fed maintains an unemployment rate of 4 percent when the *natural* rate of unemployment is actually 3.5 percent.

C H A L L E N G E Q U E S T I O N

- In addition to its short-run effects on the economy via the wealth effect, the stock market affects the economy in another way: It is part of the loanable funds market in which households make their saving available to firms. Thus, the *existence* of a stock market should affect the economy in the long run.
 - Do stocks have any advantages for households over other forms of saving? If so, what are they? (Think of yourself or your family. Why might you want to hold some of your wealth in the form of stocks, rather than hold all of it in other forms such as bonds or cash?)
 - Using a loanable funds diagram, and your answer in part (a), show what happens—in an economy that is initially without a stock market—when a viable stock market is introduced. In particular, which curve will shift?
 - Using your graph from part (b), how does introducing a stock market into the economy affect the level of investment spending and the standard of living over the long run?
 - Do stocks have any advantages for business firms over other ways of obtaining funds for investment projects? If so, what are they?
 - On your loanable funds diagram, and using your answer from part (d), illustrate the impact of the stock market on the investment demand curve. Does this contribute to, or work against, the impact of the stock market on the economy that you found in part (c)?

GLOSSARY

A

- Absolute advantage** The ability to produce a good or service, using fewer resources than other producers use.
- Accounting profit** Total revenue minus accounting costs.
- Active monetary policy** When the Fed changes the money supply to achieve some objective.
- Agent** A person hired to do a job.
- Aggregate demand (AD) curve** A curve indicating equilibrium GDP at each price level.
- Aggregate expenditure (AE)** The sum of spending by households, business firms, the government, and foreigners on final goods and services produced in the United States.
- Aggregate production function** The relationship showing how much total output can be produced with different quantities of labor, and with land, capital, and technology held constant.
- Aggregate supply (AS) curve** A curve indicating the price level consistent with firms' unit costs and markups for any level of output over the short run.
- Aggregation** The process of combining different things into a single category.
- Allocative efficiency** When there is no change in quantity consumed of any good by any consumer that would be a Pareto improvement.
- Alternate goods** Other goods that a firm could produce, using some of the same types of inputs as the good in question.
- Appreciation** An increase in the price of a currency in a floating-rate system.
- Arbitrage** Simultaneous buying and selling of a foreign currency in order to profit from a difference in exchange rates.
- Automatic stabilizers** Forces that reduce the size of the expenditure multiplier and diminish the impact of spending shocks.
- Autonomous consumption spending** The part of consumption spending that is independent of income; also, the vertical intercept of the consumption function.
- Average cost pricing** The regulatory strategy of setting price equal to a natural monopolist's long-run average total cost.
- Average fixed cost** Total fixed cost divided by the quantity of output produced.
- Average standard of living** Total output (real GDP) per person.
- Average tax rate** The fraction of a given income paid in taxes.

- Average total cost** Total cost divided by the quantity of output produced.
- Average variable cost** Total variable cost divided by the quantity of output produced.
- Averch-Johnson effect** The tendency of regulated natural monopolies to overinvest in capital.

B

- Balance sheet** A financial statement showing assets, liabilities, and net worth at a point in time.
- Banking panic** A situation in which depositors attempt to withdraw funds from many banks simultaneously.
- Bilateral arbitrage** Arbitrage involving one pair of currencies.
- Black market** A market in which goods are sold illegally at a price above the legal ceiling.
- Bond** A promise to pay a specific sum of money at some future date, or dates, issued by a corporation or government agency when it borrows funds.
- Boom** A period of time during which real GDP is above potential GDP.
- Budget constraint** The different combinations of goods a consumer can afford with a limited budget, at given prices.
- Budget deficit** The excess of government purchases over net taxes.
- Budget line** The graphical representation of a budget constraint.
- Budget surplus** The excess of net taxes over government purchases.
- Business cycles** Fluctuations in real GDP around its long-term growth trend.
- Business firm** A firm, owned and operated by private individuals, that specializes in production.

C

- Capital gain** The return someone gets by selling a financial asset at a price higher than they paid for it.
- Capital gains tax** A tax on profits earned when a financial asset is sold at more than its acquisition price.
- Capital per worker** The total capital stock divided by total employment.
- Capital stock** The total value of all goods that will provide useful services in future years.
- Capital** Long-lasting tools used in producing goods and services.

- Capitalism** A type of economic system in which most resources are owned privately.
- Cartel** A group of firms that selects a common price that maximizes total industry profits.
- Cash in the hands of the public** Currency and coins held outside of banks.
- Central bank** A nation's principal monetary authority.
- Change in demand** A shift of a demand curve in response to a change in some variable other than price.
- Change in quantity demanded** A movement along a demand curve in response to a change in price.
- Change in quantity supplied** A movement along a supply curve in response to a change in price.
- Change in supply** A shift of a supply curve in response to some variable other than price.
- Circular flow** A diagram that shows how goods, resources, and dollar payments flow between households and firms.
- Classical model** A macroeconomic model that explains the long-run behavior of the economy, assuming that all markets clear.
- Command or centrally planned economy** An economic system in which resources are allocated according to explicit instructions from a central authority.
- Communism** A type of economic system in which most resources are owned in common.
- Comparative advantage** The ability to produce a good or service at a lower opportunity cost than other producers or countries.
- Compensating wage differential** A difference in wages that makes two jobs equally attractive to a worker.
- Complement** A good that is used *together with* some other good.
- Complementary input** An input whose utilization increases the marginal product of another input.
- Complete crowding out** A dollar-for-dollar decline in one sector's spending caused by an increase in some other sector's spending.
- Constant cost industry** An industry in which the long-run supply curve is horizontal because each firm's *ATC* curve is unaffected by changes in industry output.
- Constant returns to scale** Long-run average total cost is unchanged as output increases.
- Consumer Price Index** An index of the cost, through time, of a fixed market basket of goods purchased by a typical household in some base period.
- Consumption (C)** The part of GDP purchased by households as final users.
- Consumption function** A positively sloped relationship between real consumption spending and real disposable income.
- Consumption tax** A tax on the part of their income that households spend.
- Consumption-income line** A line showing aggregate consumption spending at each level of income or GDP.
- Copyright** A grant of exclusive rights to sell a literary, musical, or artistic work.
- Corporate profits tax** A tax on the profits earned by corporations.
- Corporation** A firm owned by those who buy shares of stock and whose liability is limited to the amount of their investment in the firm.
- Countercyclical fiscal policy** Changes in taxes or government spending designed to counteract economic fluctuations.
- Coupon payments** A series of periodic payments that a bond promises before maturity.
- Critical assumption** Any assumption that affects the conclusions of a model in an important way.
- Cross-price elasticity of demand** The percentage change in the quantity demanded of one good caused by a 1-percent change in the price of another good.
- Crowding out** A decline in one sector's spending caused by an increase in some other sector's spending.
- Cyclical deficit** The part of the federal budget deficit that varies over the business cycle.
- Cyclical unemployment** Joblessness arising from changes in production over the business cycle.
- D**
- Decreasing cost industry** An industry in which the long-run supply curve slopes downward because each firm's *ATC* curve shifts downward as industry output increases.
- Deflation** A *decrease* in the price level from one period to the next.
- Demand curve facing the firm** A curve that indicates, for different prices, the quantity of output that customers will purchase from a particular firm.
- Demand curve for foreign currency** A curve indicating the quantity of a specific foreign currency that Americans will want to buy, during a given period, at each different exchange rate.
- Demand deposit multiplier** The number by which a change in reserves is multiplied to determine the resulting change in demand deposits.
- Demand deposits** Checking accounts that do not pay interest.
- Demand schedule** A list showing the quantities of a good that consumers would choose to purchase at different prices, with all other variables held constant.
- Demand shock** Any event that causes the *AD* curve to shift.
- Depreciation** A decrease in the price of a currency in a floating-rate system.
- Depression** An unusually severe recession.
- Derived demand** The demand for an input that arises from, and varies with, the demand for the product it helps to produce.
- Devaluation** A change in the exchange rate from a higher fixed rate to a lower fixed rate.
- Diminishing marginal returns to labor** The marginal product of labor decreases as more labor is hired.
- Discount rate** two meanings: (1) The interest rate used to compute present values; (2) The interest rate the Fed charges on loans to banks.

Discounting The act of converting a future value into its present-day equivalent.

Discouraged workers Individuals who would like a job, but have given up searching for one.

Discrimination When a group of people have different opportunities because of personal characteristics that have nothing to do with their abilities.

Diseconomies of scale Long-run average total cost increases as output increases.

Disequilibrium A situation in which a market does not clear—quantity supplied is not equal to quantity demanded.

Disposable income The part of household income that remains after paying taxes.

Diversification The process of reducing risk by spreading sources of income among different alternatives.

Dividends The part of a firm's current profit that is distributed to shareholders.

Dominant strategy A strategy that is best for a firm no matter what strategy its competitor chooses.

Dow Jones Industrial Average An index of the prices of stocks of 30 large U.S. firms.

Duopoly An oligopoly market with only two sellers.

E

Economic efficiency A situation in which every Pareto improvement has occurred.

Economic luxury A good with an income elasticity of demand greater than 1.

Economic necessity A good with an income elasticity of demand between 0 and 1.

Economic profit Total revenue minus all costs of production, explicit and implicit.

Economic system A system of resource allocation and resource ownership.

Economics The study of choice under conditions of scarcity.

Economies of scale Long-run average total cost decreases as output increases.

Efficient market A market that instantaneously incorporates all available information relevant to a stock's price.

Elastic demand A price elasticity of demand less than -1 .

Equilibrium GDP In the short run, the level of output at which output and aggregate expenditure are equal.

Equilibrium A state of rest; a situation that, once achieved, will not change unless some external factor, previously held constant, changes.

Excess demand for bonds The amount of bonds demanded exceeds the amount supplied at a particular interest rate.

Excess demand At a given price, the excess of quantity demanded over quantity supplied.

Excess reserves Reserves in excess of required reserves.

Excess supply of money The amount of money supplied exceeds the amount demanded at a particular interest rate.

Excess supply At a given price, the excess of quantity supplied over quantity demanded.

Exchange rate The amount of one country's currency that is traded for one unit of another country's currency.

Exchange The act of trading with others to obtain what we desire.

Excise tax A tax on a specific good or service.

Excludability The ability to exclude those who do not pay for a good from consuming it.

Exit A permanent cessation of production when a firm leaves an industry.

Expansion A period of increasing real GDP.

Expenditure approach Measuring GDP by adding the value of goods and services purchased by each type of final user.

Explicit collusion Cooperation involving direct communication between competing firms about setting prices.

Explicit costs Money actually paid out for the use of inputs.

Exports Goods and services produced domestically, but sold abroad.

Externality A by-product of a good or activity that affects someone not immediately involved in the transaction.

F

Factor markets Markets in which resources—capital, land, labor, and natural resources—are sold to firms.

Factor payments approach Measuring GDP by summing the factor payments made by all firms in the economy.

Factor payments Payments to the owners of resources that are used in production.

Federal funds rate The interest rate charged for loans of reserves among banks.

Federal Open Market Committee A committee of Federal Reserve officials that establishes U.S. monetary policy.

Federal Reserve System The central bank and national monetary authority of the United States.

Fiat money Anything that serves as a means of payment by government declaration.

Final good A good sold to its final user.

Financial asset A promise to pay future income in some form, such as future dividends or future interest payments.

Financial intermediary A business firm that specializes in brokering between savers and borrowers.

Firm's quantity supplied The total amount of a good or service that an individual firm would choose to produce and sell at a given price.

Firm's supply curve A curve that shows the quantity of output a competitive firm will produce at different prices.

Fiscal policy A change in government purchases or net taxes designed to change total spending and total output.

Fixed costs Costs of fixed inputs.

Fixed exchange rate A government-declared exchange rate maintained by central bank intervention in the foreign exchange market.

Fixed input An input whose quantity remains constant, regardless of how much output is produced.

Floating exchange rate An exchange rate that is freely determined by the forces of supply and demand.

Flow variable A measure of a process that takes place over a period of time.

Foreign currency crisis A loss of faith that a country can prevent a drop in its exchange rate, leading to a rapid depletion of its foreign currency (e.g., dollar) reserves.

Foreign exchange market The market in which one country's currency is traded for another country's.

Frictional unemployment Joblessness experienced by people who are between jobs or who are just entering or re-entering the labor market.

Friendly takeover When a firm's management arranges a takeover by another firm deemed unlikely to fire them.

Full employment A situation in which there is no cyclical unemployment.

Fundamental analysis A method of predicting a stock's price based on the fundamental forces driving the firm's future earnings.

G

Game theory An approach to modeling the strategic interaction of oligopolists in terms of moves and countermoves.

GDP price index An index of the price level for all final goods and services included in GDP.

General human capital Knowledge, education, or training that is valuable at many different firms.

Gini coefficient A measure of income inequality; the ratio of the area above a Lorenz curve and under the complete equality line to the area under the diagonal.

Government demand for funds curve Indicates the amount of government borrowing at various interest rates.

Government franchise A government-granted right to be the sole seller of a product or service.

Government purchases (G) Spending by federal, state, and local governments on goods and services.

Gross Domestic Product (GDP) The total value of all final goods and services produced for the marketplace during a given year, within the nation's borders.

H

Herfindahl-Hirschman Index The sum of squared market shares of all firms in an industry.

Hostile takeover When outsiders buy up a firm's shares with the goal of replacing the management team and increasing profits.

(Household) saving The portion of after-tax income that households do not spend on consumption goods.

Human capital The skills and training of the labor force.

I

Imperfectly competitive market A market in which a single buyer or seller has the power to influence the price of the product.

Implicit costs The cost of inputs for which there is no direct money payment.

Imports Goods and services produced abroad, but consumed domestically.

Income The amount that a person or firm earns over a particular period.

Income effect As the price of a good decreases, the consumer's purchasing power increases, causing a change in quantity demanded for the good.

Income elasticity of demand The percentage change in quantity demanded caused by a 1-percent change in income.

Increasing cost industry An industry in which the long-run supply curve slopes upward because each firm's *ATC* curve shifts upward as industry output increases.

Increasing marginal returns to labor The marginal product of labor increases as more labor is hired.

Index A series of numbers used to track a variable's rise or fall over time.

Indexation Adjusting the value of some nominal payment in proportion to a price index, in order to keep the real payment unchanged.

Individual demand curve A curve showing the quantity of a good or service demanded by a particular individual at each different price.

Individual's quantity demanded The total amount of a good an individual would choose to purchase at a given price.

Inelastic demand A price elasticity of demand between 0 and -1 .

Inferior good A good that people demand less of as their income rises.

Inflation rate The percent change in the price level from one period to the next.

Injections Spending from sources other than households.

Interest rate target The interest rate the Federal Reserve aims to achieve by adjusting the money supply.

Intermediate goods Goods used up in producing final goods.

Investment demand curve Indicates the level of investment spending firms plan at various interest rates.

Investment tax credit A reduction in taxes for firms that invest in certain favored types of capital.

Investment Firms' purchases of new capital over some period of time.

Involuntary part-time workers Individuals who would like a full-time job, but who are working only part time.

L

Labor The time human beings spend producing goods and services.

Labor demand curve Indicates how many workers firms will want to hire at various wage rates.

Labor force Those people who have a job or who are looking for one.

Labor productivity Total output (real GDP) per worker.

Labor shortage The quantity of labor demanded exceeds the quantity supplied at the prevailing wage rate.

Labor supply curve A curve indicating the number of people who want jobs in a labor market at each wage rate.

Labor surplus The quantity of labor supplied exceeds the quantity demanded at the prevailing wage rate.

- Land** The physical space on which production occurs, and the natural resources that come with it.
- Law of demand** As the price of a good increases, the quantity demanded decreases.
- Law of diminishing marginal returns** As more and more of any input is added to a fixed amount of other inputs, its marginal product will eventually decline.
- Law of diminishing marginal utility** As consumption of a good or service increases, marginal utility decreases.
- Law of increasing opportunity cost** The more of something that is produced, the greater the opportunity cost of producing one more unit.
- Law of supply** As the price of a good increases, the quantity supplied increases.
- Leakages** Income earned, but not spent, by households during a given year.
- Liquidity** The property of being easily converted into cash.
- Loan** An IOU issued by a household or noncorporate business when it borrows funds.
- Loanable funds market** Arrangements through which households make their saving available to borrowers.
- Long run** A time horizon long enough for a firm to vary all of its inputs.
- Long-run aggregate supply curve** A vertical line indicating all possible output and price-level combinations at which the economy could end up in the long run.
- Long-run average total cost** The cost per unit of output in the long run, when all inputs are variable.
- Long-run elasticity** An elasticity measured a year or more after a price change.
- Long-run labor supply curve** A curve indicating how many (qualified) people will want to work in a labor market after full adjustment to a change in the wage rate.
- Long-run Phillips curve** A vertical line indicating that in the long run, unemployment must equal its natural rate, regardless of the rate of inflation.
- Long-run supply curve** A curve indicating the quantity of output that all sellers in a market will produce at different prices, after all long-run adjustments have taken place.
- Long-run total cost** The cost of producing each quantity of output when the least-cost input mix is chosen in the long run.
- Lorenz curve** When households are arrayed according to their incomes, a line showing the cumulative percent of income received by each cumulative percent of households.
- Loss** A negative profit—when total cost exceeds total revenue.
- M**
- M1** A standard measure of the money supply, including cash in the hands of the public, checking account deposits, and travelers checks.
- M2** M1 plus savings account balances, noninstitutional money market mutual fund balances, and small time deposits.
- Macroeconomics** The study of the economy as a whole.
- Managed float** A policy of frequent central bank intervention to move the exchange rate.
- Marginal approach to profit** A firm maximizes its profit by taking any action that adds more to its revenue than to its cost.
- Marginal cost** The increase in total cost from producing one more unit of output.
- Marginal decision making** To understand and predict the behavior of individual decision makers, we focus on the incremental or marginal effects of their actions.
- Marginal product of labor** The additional output produced when one more worker is hired.
- Marginal propensity to consume** The amount by which consumption spending rises when disposable income rises by one dollar.
- Marginal revenue product (MRP)** The change in revenue from hiring one more worker.
- Marginal revenue product of capital** The increase in output due to a one-unit increase in the capital input.
- Marginal revenue** The change in total revenue from producing one more unit of output.
- Marginal tax rate** The fraction of an additional dollar of income paid in taxes.
- Marginal utility** The change in total utility an individual obtains from consuming an additional unit of a good or service.
- Market** A group of buyers and sellers with the potential to trade with each other.
- Market clearing** Adjustment of prices until quantities supplied and demanded are equal.
- Market demand curve** The graphical depiction of a demand schedule; a curve showing the quantity of a good or service demanded at various prices, with all other variables held constant.
- Market economy** An economic system in which resources are allocated through individual decision making.
- Market failure** A market equilibrium that fails to take advantage of every Pareto improvement.
- Market labor demand curve** A curve indicating the total number of workers all firms in a labor market want to employ at each wage rate.
- Market quantity demanded** The total amount of a good that all buyers in the market would choose to purchase at a given price.
- Market quantity supplied** The total amount of a good or service that all producers in a market would choose to produce and sell at a given price.
- Market signals** Price changes that cause firms to change their production to more closely match consumer demand.
- Market structure** The characteristics of a market that influence how trading takes place.
- Market supply curve** A curve indicating the quantity of output that all sellers in a market will produce at different prices.
- Maturity date** The date at which a bond's principal amount will be paid to the bond's owner.
- Means of payment** Anything acceptable as payment for goods and services.
- Microeconomics** The study of the behavior of individual households, firms, and governments; the choices they make; and their interaction in specific markets.

Minimum efficient scale (MES) The level of output at which economies of scale are exhausted and minimum *LRATC* is achieved.

Model An abstract representation of reality.

Money supply curve A line showing the total quantity of money in the economy at each interest rate.

Monopolistic competition A market structure in which there are many firms selling products that are differentiated, yet are still close substitutes, and in which there is free entry and exit.

Monopoly firm The only seller of a good or service that has no close substitutes.

Monopoly market The market in which a monopoly firm operates.

Mutual fund A corporation that specializes in owning shares of stock in other corporations.

N

National debt The total amount of government debt outstanding.

Natural monopoly A market in which, due to economies of scale, one firm can operate at lower average cost than can two or more firms.

Natural rate of unemployment The unemployment rate when there is no cyclical unemployment.

Net capital inflow An inflow of funds equal to a nation's trade deficit.

Net exports (NX) Total exports minus total imports.

Net investment Total investment minus depreciation.

Net taxes Government tax revenues minus transfer payments.

Net worth The difference between assets and liabilities.

Nominal interest rate The annual percent increase in a lender's dollars from making a loan.

Nominal variable A variable measured in current dollars, without adjustment for the *dollar's* changing value.

Nonmarket production Goods and services that are produced, but not sold in a market.

Nonmonetary job characteristic Any aspect of a job—other than the wage—that matters to a potential or current employee.

Nonprice competition Any action a firm takes to increase the demand for its product, other than cutting its price.

Normal good A good that people demand more of as their income rises.

Normal profit Another name for zero economic profit.

Normative economics The study of what *should* be; it is used to make value judgments, identify problems, and prescribe solutions.

O

Oligopoly A market structure in which a small number of firms are strategically interdependent.

Open market operations Purchases or sales of bonds by the Federal Reserve System.

Opportunity cost The value of the best alternative, or alternatives, sacrificed when taking an action.

Optimum currency area A region whose economies perform better with a single currency than with separate national currencies.

P

Pareto improvement An action that makes at least one person better off, and harms no one.

Partnership A firm owned and usually operated by several individuals who share in the profits and bear personal responsibility for any losses.

Passive monetary policy When the Fed keeps the money supply constant regardless of shocks to the economy.

Patent protection A government grant of exclusive rights to use or sell a new technology.

Patent A temporary grant of monopoly rights over a new product or scientific discovery.

Payoff matrix A table showing the payoffs to each of two players for each pair of strategies they choose.

Peak The point at which real GDP reaches its highest level during an expansion.

Perfect competition A market structure in which there are many buyers and sellers, the product is standardized, and sellers can easily enter or exit the market.

Perfect price discrimination Charging each customer the most he or she would be willing to pay for each unit purchased.

Perfectly (infinitely) elastic demand A price elasticity of demand approaching minus infinity.

Perfectly competitive labor market A market with many indistinguishable sellers of labor and many buyers, and that involves no barriers to entry or exit.

Perfectly competitive market A market in which no buyer or seller has the power to influence the price.

Perfectly inelastic demand A price elasticity of demand equal to 0.

Phillips curve A curve indicating possible combinations of inflation and unemployment in the short run.

Planned investment spending Business purchases of plant and equipment.

Plant The collection of fixed inputs at a firm's disposal.

Positive economics The study of what *is*, of how the economy works.

Potential output The level of output the economy could produce if operating at full employment.

Poverty line The income level below which a family is considered to be in poverty.

Poverty rate The percent of families whose incomes fall below a certain minimum—the poverty line.

Present value The value, in today's dollars, of a sum of money to be received or paid at a specific date in the future.

Price The amount of money that must be paid to a seller to obtain a good or service.

Price ceiling A government-imposed maximum price in a market.

Price discrimination Charging different prices to different customers for reasons other than differences in cost.

Price elasticity of demand The sensitivity of quantity demanded to price; the percentage change in quantity demanded caused by a 1-percent change in price.

Price floor A government-imposed minimum price in a market.

Price leadership A form of tacit collusion in which one firm sets a price that other firms copy.

Price level The average level of dollar prices in the economy.

Price taker Any firm that treats the price of its product as given and beyond its control.

Primary market The market in which newly issued financial assets are sold for the first time.

Principal A person or group that hires someone to do a job.

Principal (face value) The amount of money a bond promises to pay when it matures.

Principal-agent problem The situation that arises when an agent has interests that conflict with the principal's, and has the ability to pursue those interests.

Principle of asset valuation The idea that the value of an asset is equal to the total present value of all the future benefits it generates.

Private good A good that is rival and excludable, and is supplied by private firms in the marketplace.

Private investment (*I*) The sum of business plant and equipment purchases, new home construction, and inventory changes.

Product markets Markets in which firms sell goods and services to households or other firms.

Production function A function that indicates the maximum amount of output a firm can produce over some period of time from each combination of inputs.

Production possibilities frontier (PPF) A curve showing all combinations of two goods that can be produced with the resources and technology currently available.

Productive efficiency When it is impossible to produce more of one good without producing less of some other good.

Productive inefficiency A situation in which more of at least one good can be produced without sacrificing the production of any other good.

Profit Total revenue minus total cost.

Progressive tax A tax whose rate increases as income increases.

Property income Income derived from supplying capital, land, or natural resources.

Protectionism The belief that a nation's industries should be protected from foreign competition.

Public good A good that is non-rivalrous and non-excludable; the market cannot, and should not, provide such goods.

Purchasing power parity (PPP) theory The idea that the exchange rate will adjust in the long run so that the average price of goods in two countries will be roughly the same.

Pure discount bond A bond that promises no payments except for the principal it pays at maturity.

Q

Quota A limit on the physical volume of imports.

R

Rational preferences Preferences that satisfy two conditions: (1) Any two alternatives can be compared, and one is preferred or else the two are valued equally, and (2) the comparisons are logically consistent.

Real interest rate The annual percent increase in a lender's *purchasing power* from making a loan.

Real variable A variable measured in the dollars of a base year, thereby adjusting for changes in the dollar's value.

Recession A period of declining or abnormally low real GDP.

Relative price The price of one good relative to the price of another.

Rent controls Government-imposed maximum rents on apartments and homes.

Rent-seeking activity Any costly action a firm undertakes to establish or maintain its monopoly status.

Repeated play A situation in which strategically interdependent sellers compete over many time periods.

Required reserve ratio The minimum fraction of checking account balances that banks must hold as reserves.

Required reserves The minimum amount of reserves a bank must hold, depending on the amount of its deposit liabilities.

Reservation wage The lowest wage rate at which an individual would supply labor to a particular labor market.

Reserves Vault cash plus balances held at the Fed.

Resource allocation A method of determining which goods and services will be produced, how they will be produced, and who will get them.

Resources The land, labor, and capital that are used to produce goods and services.

Rivalry A situation in which one person's consumption of a good or service means that no one else can consume it.

Run on the bank An attempt by many of a bank's depositors to withdraw their funds simultaneously.

S

Say's law The idea that total spending will be sufficient to purchase the total output produced.

Scarcity A situation in which the amount of something available is insufficient to satisfy the desire for it.

Seasonal unemployment Joblessness related to changes in weather, tourist patterns, or other seasonal factors.

Secondary market The market in which previously issued financial assets are sold.

Self-correcting mechanism The adjustment process through which price and wage changes return the economy to full-employment output in the long run.

Share of stock A share of ownership in a corporation.

Short run A time horizon during which at least one of the firm's inputs cannot be varied.

Short side of the market The smaller of quantity supplied and quantity demanded at a particular price.

Short-run elasticity An elasticity measured just a short time after a price change.

- Short-run macro model** A macroeconomic model that explains how changes in spending can affect real GDP in the short run.
- Short-run macroeconomic equilibrium** A combination of price level and GDP consistent with both the *AD* and *AS* curves.
- Shutdown price** The price at which a firm is indifferent between producing and shutting down.
- Shutdown rule** In the short run, the firm should continue to produce if total revenue exceeds total variable costs; otherwise, it should shut down.
- Simplifying assumption** Any assumption that makes a model simpler without affecting any of its important conclusions.
- Single-price monopoly** A monopoly firm that is limited to charging the same price for each unit of output sold.
- Socialism** A type of economic system in which most resources are owned by the state.
- Sole proprietorship** A firm owned by a single individual.
- Specialization** A method of production in which each person concentrates on a limited number of activities.
- Specific human capital** Knowledge, education, or training that is valuable only at a specific firm.
- Spending shock** A change in spending that ultimately affects the entire economy.
- Stagflation** The combination of falling output and rising prices.
- Standard & Poor's 500** An index of the prices of stocks of 500 large U.S. firms.
- Statistical discrimination** When individuals are excluded from an activity based on the statistical probability of behavior in their group, rather than their personal characteristics.
- Stock options** Rights to purchase shares of stock at a prespecified price.
- Stock variable** A measure of an amount that exists at a moment in time.
- Stockholder revolt** When owners, dissatisfied with the profits they are earning, replace the firm's management team.
- Structural deficit** The part of the federal budget deficit that is independent of the business cycle.
- Structural unemployment** Joblessness arising from mismatches between workers' skills and employers' requirements or between workers' locations and employers' locations.
- Substitute** A good that can be used in place of some other good and that fulfills more or less the same purpose.
- Substitute input** An input whose utilization decreases the marginal product of another input.
- Substitution effect** As the price of a good falls, the consumer substitutes that good in place of other goods whose prices have not changed.
- Sunk cost** A cost that was incurred in the past and does not change in response to a present decision.
- Supply curve** A graphical depiction of a supply schedule; a curve showing the quantity of a good or service supplied at various prices, with all other variables held constant.
- Supply curve for foreign currency** A curve indicating the quantity of a specific foreign currency that will be supplied, during a given period, at each different exchange rate.
- Supply of funds curve** Indicates the level of household saving and any budget surplus at various interest rates.
- Supply schedule** A list showing the quantities of a good or service that firms would choose to produce and sell at different prices, with all other variables held constant.
- Supply shock** Any event that causes the *AS* curve to shift.
- T**
- Tacit collusion** Any form of oligopolistic cooperation that does not involve an explicit agreement.
- Tariff** A tax on imports.
- Technical analysis** A method of predicting a stock's price based on that stock's past behavior.
- Technological change** The invention or discovery of new inputs, new outputs, or new production methods.
- Technology** The set of methods a firm can use to turn inputs into outputs.
- Terms of trade** The ratio at which a country can trade domestically produced products for foreign-produced products.
- Tit for tat** A game-theoretic strategy of doing to another player this period what he has done to you in the previous period.
- Tort** A wrongful act that harms someone.
- Total cost** The costs of all inputs—fixed and variable.
- Total demand for funds curve** Indicates the total amount of borrowing at various interest rates.
- Total fixed cost** The cost of all inputs that are fixed in the short run.
- Total product** The maximum quantity of output that can be produced from a given combination of inputs.
- Total revenue** The total inflow of receipts from selling a given amount of output.
- Total variable cost** The cost of all variable inputs used in producing a particular level of output.
- Traditional economy** An economy in which resources are allocated according to long-lived practices from the past.
- Tragedy of the commons** The problem of overuse when a good is rival but nonexcludable.
- Transaction costs** The time costs and other costs required to carry out market exchanges.
- Transfer payment** Any payment that is not compensation for supplying goods or services.
- Triangular arbitrage** Arbitrage involving trades among three (or more) currencies.
- Trough** The point at which real GDP reaches its lowest level during a recession.
- U**
- Unemployment rate** The fraction of the labor force that is without a job.
- Unit of value** A common unit for measuring how much something is worth.
- Unitary elastic demand** A price elasticity of demand equal to -1 .

Utility Pleasure or satisfaction obtained from consuming goods and services.

V

Value added The revenue a firm receives minus the cost of the intermediate goods it buys.

Value-added approach Measuring GDP by summing the value added by all firms in the economy.

Variable costs Costs of variable inputs.

Variable input An input whose usage changes as the level of output changes.

W

Wage taker Any firm that takes the market wage rate as a given when making employment decisions.

Wealth constraint At any point in time, wealth is fixed.

Wealth The total value of everything a person or firm owns, at a point in time, minus the total value of everything owed.

White knight A firm that undertakes a friendly takeover.

Y

Yield The rate of return a bond earns for its owner.

