# ECONOMICS (ECONOMICS SCIENCE) 

MICHAEL PARKIN
University of Western Ontario

GLOBAL EDITION
TWELFTH EDITION
Publisher: PEARSON EDUCATION LIMITED
ISBN 9781-292-09450-2

## CONTENTS

## PART 1: INTRODUCTION

1. What Is Economics?
2. The Economic Problem

## PART 2: HOW MARKETS WORK

3. Demand and Supply
4. Elasticity
5. Efficiency and Equity
6. Government Actions in Markets
7. Global Markets in Action

## PART 3: HOUSEHOLDS' CHOICES

8. Utility and Demand
9. Possibilities, Preferences, and Choices

## PART 4: FIRMS AND MARKETS

10. Organizing Production
11. Output and Costs
12. Perfect Competition
13. Monopoly
14. Monopolistic Competition
15. Oligopoly

## PART 5: MARKET FAILURE AND GOVERNMENT

16. Public Choices, Public Goods, and Healthcare
17. Externalities

## PART 6: FACTOR MARKETS, INEQUALITY, AND UNCERTAINTY

18. Markets for Factors of Production
19. Economic Inequality
20. Uncertainty and Information

## PART 7: MONITORING MACROECONOMIC PERFORMANCE

21. Measuring GDP and Economic Growth
22. Monitoring Jobs and Inflation

## PART 8: MACROECONOMIC TRENDS

23. Economic Growth
24. Finance, Saving, and Investment
25. Money, the Price Level, and Inflation
26. The Exchange Rate and the Balance of Payments

## PART 9: MACROECONOMIC FLUCTUATIONS

27. Aggregate Supply and Aggregate Demand
28. Expenditure Multipliers
29. The Business Cycle, Inflation, and Deflation

## PART 10: MACROECONOMIC POLICY

30. Fiscal Policy
31. Monetary Policy

## PART 1: INTRODUCTION

## CHAPTER 1: What is Economics?

After studying this chapter, you will be able to:

- Define economics and distinguish between microeconomics and macroeconomics
- Explain the two big questions of economics
- Explain the key ideas that define the economic way of thinking
- Explain how economists go about their work as social scientists and policy advisers

Is economics about money: How people make it and spend it? Is it about business, government, and jobs? Is it about why some people and some nations are rich and others poor? Economics is about all thesethings. But its core is the study of choices and their consequences. This chapter gets you started by describing the questions that economists try to answer and looking at how economists think as they search for the answers.

## Definition of Economics

A fundamental fact dominates our lives: We want more than we can get. Our inability to get everythingwe want is called scarcity. What you can afford to buy is limited by your income and by the prices you must pay. And your time is limited by the fact that your day has 24 hours. You want some other things that only governments provide. What governments can afford is limited by the taxes they collect. Because we can't get everything we want, we must make choices. You can't afford both a laptop and an iPhone, so you must choose which one to buy. You can't spend tonight both studying for your next test and going to the movies, so again, you must choose which one to do. Governments can't spend a tax dollar on both national defense and environmental protection, so they must choose how to spend that dollar.
Your choices must somehow be madeconsistent with the choices of others. If you choose to buy a laptop, someone else must choose to sell it. Incentives reconcile choices. An incentive is a reward that encourages an action or a penalty that discourages one. Prices act as incentives. If the price of a laptop is too high, more will be offered for sale than people want to buy. And if the price is too low, fewer will be offered for sale than people want to buy. But there is a price at which choices to buy and sell are consistent.

Economics is the social science that studies the choices that individuals, businesses, governments, and entire societies make asthey cope with scarcity and the incentives that influence and reconcile those choices.
The subject has two parts:

1. Microeconomics is the study of the choices that individuals and businesses make, the way these choices interact in markets, and the influence of governments.
2. Macroeconomics is the study of the performance of the national economy and the global economy.

Two Big Economic Questions summarize the scope of economics:

1. How do choices end up determining what, how, and for whom goods and services are produced?
2. Do choices made in the pursuit of self-interest also promote the social interest?

## What, How, and For Whom?

Goods and services are the objects that people value and produce to satisfy wants. Goods are physical objects such as cellphones and automobiles. Services are tasks performed for people such as cellphone service and auto-repair service.

What? What we produce varies across countries and changes over time. In the United States today, agriculture accounts for $1 \%$ of total production, manufactured goods for $20 \%$, and services (retail and wholesale trade, healthcare, and education are the biggest ones) for $79 \%$. In contrast, in China today, agriculture accounts for $10 \%$ of total production, manufactured goods for $45 \%$, and services for $45 \%$ Figure 1.1 on page 41 shows these numbers and also the percentages for Brazil, which fall between those for the United States and China. What determines these patterns of production? How do choices end up determining the quantities of cellphones, automobiles, cellphone service, auto-repair service, and the millions of other items that are produced in the United States and around the world?

How? How we produce is described by the technologies and resources that we use. The resources used to produce goods and services are called factors of production, which are grouped into four categories:

Land The "gifts of nature" that we use to producegoods and services are called land. In economics, land is what in everyday language we call naturalresources. It includes land in the everyday sense together with minerals, oil, gas, coal, water, air, forests, and fish.

Labor The work time and work effort that people devote to producing goods and services is called labor. Labor includes the physical and mental efforts of all the people who work on farms and construction sites and in factories, shops, and offices.
The quantity of labor depends on human capital, which is the knowledge and skill that people obtainfrom education, on-the-job training, and work experience. Figure 1.2 on page 42 shows these measures of human capital in the United States and its growth over the past 110 years.

Capital The tools, instruments, machines, buildings, and other constructions that businesses use to produce goods and services are called capital. In everyday language, we talk about money, stocks, and bonds as being "capital." These items arefinancial capital. Financial capital plays an important role in enabling businesses to borrow the funds that they use to buy physical capital. But financial capitalis not used to produce goods and services and it is not a factor of production.

Entrepreneurship The human resource that organizes labor, land, and capital is called entrepreneurship. Entrepreneurs are the drivers of economic progress. They develop new ideas about what and how to produce, make business decisions, and bear the risksthat arise from these decisions. What determines how the factors of production are used to produce each good and service?

ForWhom? Whoconsumes the goods and services that are produced depends on the incomes that people earn. People with large incomes can buy a wide range of goods and services. People with small incomes have fewer options and can afford a smaller range of goods and services. People earn their incomes by selling the services of the factors of production they own:

- Land earns rent.
- Labor earns wages.
- Capital earns interest.
- Entrepreneurship earns profit.

Which factor of production earns the most income? The answer is labor. In 2011, wages were 68 percent of total income and the incomes from land, capital, and entrepreneurship totaled 32 percent. These shares remain remarkably constant overtime.
We can get a good sense of who consumes the goods and services produced by looking at the percentages of total income earned by different groupsof people. The 20 percent of people with the lowestincomes earn about 5 percent of total income, whilethe richest 20 percent earn close to 50 percent oftotal income. So on average, people in the richest 20 percent earn more than 10 times the incomes of those in the poorest 20 percent. There is even huge inequality within the richest 20 percent and the top 1 percent earns almost 15 percent of total income. Why is the distribution of income so unequal?

## Do Choices Made in the Pursuit of Self-Interest also Promote the Social Interest?

Every day, you and 320 million other Americans, along with 7.2 billion people in the rest of the world, make economic choices that result in what, how, and for whom goods and services are produced. These choices are made by people who are pursuing their self-interest.

Self-Interest All the choices that people make about how to use their time and other resources are made in the pursuit of self-interest. The big question is: Is it possible that all the choices that each one of us makes in the pursuit of self-interest could end up achieving an outcome thatis best for everyone?

Social Interest An outcome is in the social interest if it is best for society as a whole. It is easy to see how youdecide what is in your self-interest. But how do we decide if something is in the social interest? To help you answer this question, imagine a scene like that in Economics in the News on page 50. The economist's answer is "Yes." It is in the social interest because it makes everyone better off. There are nolosers.

Efficiency and the Social Interest Economists usethe everyday word "efficient" to describe a situation that can't be improved upon. Resource use is efficient if it is not possible to make someone better off without making someone else worse off. If it is possible to make someone better off without making anyone worse off, society can be made better off and the situation is not efficient. In the Ted story everyone is better off, so it improves efficiency and the outcome is in the social interest.

FairShares and the Social Interest The idea that the social interest requires "fair shares" is a deeply held one. But what is fair? There isn't a crisp definition of fairness to match that of efficiency. Reasonable people have a variety of views about it. Almost everyone agrees that too much inequality is unfair. But how much is too much? And inequality of what: income, wealth, or the opportunity to work, earn an income, and accumulate wealth? You will examine efficiency again in Chapter 2 and efficiency and fairness in Chapter 5.

Questions about the social interest are hard ones to answer and they generate discussion, debate, anddisagreement. Four issues in today's world put someflesh on these questions. The issues are:

- Globalization
- Information-age monopolies
- Climate change
- Financial instability

Globalization The term globalization means the expansion of international trade, borrowing and lending, and investment. Globalization is in the self-interest of those consumers who buy low-cost goods and services produced in other countries; and it is in the self-interest of the multinational firms that produce in low-cost regions and sell in high-price regions. But is globalization in the self-interest of the low-wage worker in Malaysia who sews your new running shoes and the displaced shoemaker in Atlanta? Is it in the social interest?

Information-Age Monopolies The technological change of the past forty years has been called the Information Revolution. An absence of competition gave Microsoft the power to sell Windows at prices far above the cost of production. With lower prices, many more people would have been able to afford and buy a computer. The information revolution has clearly served your self-interest: It has provided your cellphone, lap-top, loads of handy applications, and the Internet. It has also served the self-interest of Bill Gates who has seen his wealth soar. But did the information revolution best serve the social interest? Did Microsoft produce the best possible Windows operating system and sell it at a price that was in the social interest? Or was the quality too low and the price too high?

Climate Change Burning fossil fuels to generate electricity and to power airplanes, automobiles, and trucks pours a staggering 28 billion tons -4 tons per person - of carbon dioxide into the atmosphere eachyear. These carbon emissions, two thirds of which comes from the United States, China, the European Union, Russia, and India, bring global warming and climate change. Must governments change the incentives we face so that our
self-interested choices are also in the social interest? How can governments change incentives? How can we encourage the use of wind and solar power to replace the burning of fossil fuels that brings climate change?

Financial Instability In 2008, banks were in trouble. They had made loans that borrowers couldn't repay and they were holding securities the values of which had crashed.
Banks' choices to take deposits and make loans are made in self-interest, but does this lending and borrowing serve the social interest? Do banks lend too much in the pursuit of profit? When banks got into trouble in 2008, the Federal Reserve (the Fed) bailed them out with big loans backed by taxpayer dollars. Did the Fed's bailout of troubled banks serve the social interest? Or might the Fed's rescue action encourage banks to repeat their dangerous lending in the future?

We've looked at four topics and asked many questions that illustrate the potential conflict between the pursuit of self-interest and the social interest. We've asked questions but not answered them because we've not yet explained the economic principles needed to do so. We will answer these questions in future chapters.

## The Economic Way of Thinking

The questions that economics tries to answer tell us about the scope ofeconomics, but they don't tell us how economists think and go about seeking answers to these questions. You're now going to see how economists go about their work. We're going to look at six key ideas that define the economic way of thinking. These ideas are

- A choice is a tradeoff.
- People make rational choices by comparing benefits and costs.
- Benefit is what you gain from something.
- Cost is what you must give up to get something.
- Most choices are "how-much" choices made at the margin.
- Choices respond to incentives.


## A Choice is a Tradeoff

Because we face scarcity, we must make choices. You can think about your choices as tradeoffs. A tradeoff is an exchange - giving up one thing to get something else.

## Making a Rational Choice

Economists view the choices that people make as rational. A rational choice is one that compares costs and benefits and achieves the greatest benefit over cost for the person making the choice. Only the wants of the person making a choice are relevant to determine its rationality. The idea of rational choice provides an answer to the first question: What goods and services will be produced and in what quantities? The answer is those that people rationally choose to buy! But how do people choose rationally? Why do more people choose an iPad rather than a Microsoft Surface? Why has the U.S.
government chosen to build an interstate highway system and not an interstate highspeed railroad system? The answers turn on comparing benefits and costs.

## Benefit: What You Gain

The benefit of something is the gain or pleasure that it brings and is determined by preferences - by what a person likes and dislikes and the intensity of those feelings. Some benefits are large and easy to identify, such as the benefit that you get from being in school. Economists measure benefit as the most that a person is willing to give up to get something. You are willing to give up a lot to be in school.

## Cost: What You Must Give Up

The opportunity cost of something is the highest-valued alternative that must be given up to get it. To make the idea of opportunity cost concrete, think about your opportunity cost of being in school. It has two components: the things you can't afford to buy and the things you can't do with your time.
Start with the things you can't afford to buy. You've spent all your income on tuition, residence fees, books, and a laptop. If you weren't in school, you would have spent this money on tickets to ball games and movies and all the other things that you enjoy. But that's only the start of your opportunity cost. You've also given up the opportunity to get a job. Suppose that the best job you could get if you weren't in school is working at Citibank as a teller earning $\$ 25,000$ a year. Another part of your opportunity cost of being in school is all the things that you could buy with the extra $\$ 25,000$ you would have.
As you well know, being a student eats up many hours in class time, doing homework assignments, preparing for tests, and so on. To do all these schoolactivities, you must give up many hours of what would otherwise be leisure time spent with your friends.
So the opportunity cost of being in school is all the good things that you can't afford and don't have the spare time to enjoy. You might want to put a dollar value on that cost or you might just list all the items that make up the opportunity cost.
The examples of opportunity cost that we've just considered are all-or-nothing costs you're either in school or not in school. Most situations are not like this one. They involve choosing how much of an activity to do.

## How Much? Choosing at the Margin

You can allocate the next hour between studying and chatting online with your friends, but the choice is not all or nothing. You must decide how many minutes to allocate to each activity. To make this decision, you compare the benefit of a little bit more study time with its cost - you make your choice at the margin. The benefit that arises from an increase in an activity is called marginal benefit. For example, your marginal benefit from one more night of study before a test is the boost it gives to your grade. Your marginal benefit doesn't include the grade you're already achieving without that extra night of work.
The opportunity cost of an increase in an activity is called marginal cost. For you, the marginal cost of studying one more night is the cost of not spending that night on your favorite leisure activity. To make your decisions, you compare marginal benefit and marginal cost. If the marginal benefit from an extra night of study exceeds its marginal cost,
you study the extra night. If the marginal cost exceedsthe marginal benefit, you don't study the extra night.

## Choices Respond to Incentives

Economists take human nature as given and view people as acting in their self-interest. All people - you, other consumers, producers, politicians, and public servants pursue their self-interest. Self-interested actions are not necessarily selfish actions. The central idea of economics is that we can predict the self-interested choices that people make bylooking at the incentives they face. People undertakethose activities for which marginal benefit exceeds marginal cost; and they reject options for which marginal cost exceeds marginal benefit. Economists emphasize the crucial role that institutions play in influencing the incentives that peopleface as they pursue their self-interest. Laws that protect private property and markets that enable voluntary exchange are the fundamental institutions. You will learn as you progress with your study of economics that where these institutions exist, self-interest can indeed promote the social interest.

## Economics as Social Science and Policy Tool

Economics is both a social science and a toolkit for advising on policy decisions.

## Economist as Social Scientist

As social scientists, economists seek to discover how the economic world works. In pursuit of this goal, like all scientists, economists distinguish between positive and normative statements.

Positive Statements A positive statement is about what is. It says what is currently believed about the way the world operates. A positive statement might be right or wrong, but we can test it by checking itagainst the facts. A central task of economists is to test positive statements about how the economic world works and to weed out those that are wrong. Economics first got off the ground in the late 1700s, so it is a young science compared with, for example, physics, and much remains to be discovered.

Normative Statements A normative statement is about what ought to be. It depends on values and cannot be tested. Policy goals are normative statements.For example, "We ought to cut our use of coal by 50 percent" is a normative policy statement. You may agree or disagree with it, but you can't test it. Itdoesn't assert a fact that can be checked.

Unscrambling Cause and Effect Economists are particularly interested in positive statements about cause and effect. Economists create and test economic models. An economic model is a description of some aspect of the economic world that includes only those features that are needed for the purpose at hand. A model is tested by comparing its predictions with the facts. But testing an economic model is difficult because we observe the outcomes of the simultaneous change of many factors. To cope with this problem, economists look for natural experiments (situations in the ordinary course of economic life in which theone factor of interest is different and other things are equal or similar); conduct statistical
investigations to find correlations; and perform economic experimentsby putting people in decision-making situations and varying the influence of one factor at a time to dis-cover how they respond.

## Economist as Policy Advisor

Economics is useful. It is a toolkit for advising governments and businesses and for making personal decisions. All the policy questions on which economists provide advice involve a blend of the positive and the normative. Economics can't help with the normative part the policy goal. But it can help to clarify the goal. And for a given goal, economics provides the tools for evaluating alternative solutions - comparing marginal benefits and marginal costs and finding the solution that makes the best use of the available resources.

## APPENDIX

Graphs in Economics

## After studying this appendix (Pages 53 - 65), you will be able to:

- Make and interpret a scatter diagram
- Identify linear and nonlinear relationships and relationships that have a maximum and a minimum
- Define and calculate the slope of a line
- Graph relationships among more than two variables


## Graphing Data

A graph represents a quantity as a distance on a line.In Fig. A1.1, a distance on the horizontal line represents temperature, measured in degrees Fahrenheit. A movement from left to right shows an increase intemperature. The point 0 represents zero degrees Fahrenheit. To the right of 0 , the temperature is positive. To the left of 0 , the temperature is negative (as indicated by the minus sign). A distance on the vertical line represents height, measured in thousands of feet. The point O represents sea level. Points above O represent feet above sea level. Points below O represent feet below sea level (indicated by a minus sign).
In Fig. A1.1, two scale lines are perpendicular to each other and are called axes. The vertical line is the $y$-axis, and the horizontal line is the $x$-axis. Each axis has a zero point, which is shared by the two axes and called the origin.

To make a two-variable graph, we need two pieces of information: the value of the variable $x$ and the value of the variable $y$. For example, off the coast of Alaska, the temperature is 32 degrees - the value of $x$. A fishing boat is located at O feet above sea level - the value of $y$. These two bits of information appear as point $A$ in Fig. A1.1. A climber at the top of Mount McKinley on a cold day is 20,320 feet above sea level in a zero-degree gale. These two pieces of information appear as point $B$. On a warmer day, a climber might be at the peak of Mt. McKinley when the temperature is 32 degrees, at point $C$.

## FIGURE A1.1 Making a Graph



Graphs have axes that measure quantities as distances. Here, the horizontal axis (x-axis) measures temperature, and the vertical axis ( $y$-axis) measures height. Point $A$ represents a fishing boat at sea level ( 0 on the $y$-axis) on a day when the temperature is $32^{\circ} \mathrm{F}$. Point $B$ represents a climber at the top of Mt. McKinley, 20,320 feet above sea level at a temperature of $0^{\circ} \mathrm{F}$. Point $C$ represents a climber at the top of Mt. McKinley, 20,320 feet above sea level at a temperature of $32^{\circ} \mathrm{F}$.

We can draw two lines, called coordinates, from point C. One, called the x-coordinate, runs from C to the vertical axis. This line is called "the $x$-coordinate" because its length is the same as the value marked off on the $x$-axis. The other, called the $y$-coordinate, runs from $C$ to the horizontal axis. This line is called "the $y$-coordinate" because its length is the same as the value marked off on the $y$-axis.
We describe a point on a graph by the values of its $x$-coordinate and its $y$-coordinate. For example, at point $C, x$ is 32 degrees and $y$ is 20,320 feet.
A graph like that in Fig. A1.1 can be made using any quantitative data on two variables. The graph can show just a few points, like Fig. A1.1 or many points.

## Scatter Diagrams

A scatter diagram is a graph that plots the value of one variable against the value of another variable for a number of different values of each variable. Such a graph reveals whether a relationship exists between two variables and describes their relationship.
The table in Fig. A1.3 shows some data on two variables: the number of tickets sold at the box office and the number of DVDs sold for nine of the mostpopular movies in 2013.

What is the relationship between these two variables? Does a big box office success generate a large volume of DVD sales? Or does a box office success mean that fewer DVDs are sold?
We can answer these questions by making a scatter diagram. We do so by graphing the data in the table. In the graph in Fig. A1.3, each point shows the number of box office tickets sold (the $x$ variable) and the number of DVDs sold (the $y$ variable) of one of the movies. There are nine movies, so there are ninepoints "scattered" within the graph.
The point labeled $A$ tells us that Monsters University sold 33 million tickets at the box office and 2.3 million DVDs. The points in the graph don't form a distinct pattern. They suggest that large box office sales do not directly bring large DVD sales. If you want to predict a movie's DVD sales in a given year with any confidence, you need to know more than the number of tickets sold at the box office in that year.
Figure A1 . 4 shows two scatter diagrams of economic variables. Part (a) shows the relationship between income and expenditure, on average, from 2001 to 2013. Each point represents income and expenditure in a given year. For example, point Ashows that in 2006, income was $\$ 38,000$ and expenditure was $\$ 31,000$. This graph shows that as income increases, so does expenditure, and the relationship is a close one.

FIGURE AI. 3 A Scatter Diagram


## Graphs Used in Economic Models

The graphs used in economics are not always designedto show real-world data. Often they are used to show general relationships among the variables in an economic model.
An economic model is a stripped-down, simplified description of an economy or of a component of an economy such as a business or a household. It consists of statements about economic behavior that can be expressed as equations or as curves in a graph.

Economists use models to explore the effects of different policies or other influences on the economy in ways that are similar to the use of model airplanes in wind tunnels and models of the climate.
You will encounter many different kinds of graphs in economic models, but there are some repeating patterns. Once you've learned to recognize these patterns, you will instantly understand the meaning of a graph. Here, we'll look at the different types of curves that are used in economic models, and we'll see some everyday examples of each type of curve. The patterns to look for in graphs are the fourcases in which

- Variables move in the same direction.
- Variables move in opposite directions.
- Variables have a maximum or a minimum.
- Variables are unrelated.

Let's look at these four cases.

## Variables That Move in the Same Direction

Figure Al. 5 shows graphs of the relationships between two variables that move up and down together. A relationship between two variables thatmove in the same direction is called a positive relationship or a direct relationship. A line that slopes upward shows such a relationship.
Figure A1.5 shows three types of relationships: one that has a straight line and two that have curved lines. All the lines in these three graphs are called curves. Any line on a graph - no matter whether it is straight or curved - is called a curve.
A relationship shown by a straight line is called a linear relationship. Figure A1.5 (a) shows a linear relationship between the number of miles traveled in5 hours and speed. For example, point $A$ shows thatyou will travel 200 miles in 5 hours if your speed is 40 miles an hour. If you double your speed to 80 miles an hour, you will travel 400 miles in 5 hours.
Figure A1.5 (b) shows the relationship between distance sprinted and recovery time (the time it takes the heart rate to return to its normal resting rate).
This relationship is an upward-sloping one that starts out quite flat but then becomes steeper as we move along the curve away from the origin. The reason this curve becomes steeper is that the additional recovery time needed from sprinting an additional 100 yards increases. It takes less than 5 minutes to recover from sprinting 100 yards but more than 10 minutes to recover from 200 yards.
Figure A1.5(c) shows the relationship between the number of problems worked by a student andthe amount of study time. This relationship is anupward-sloping one that starts out quite steep andbecomes flatter as we move along the curve away from the origin. Study time becomes less productive as the student spends more hours studying and becomes more tired.

FIGURE A1.5 Positive (Direct) Relationships

(a) Positive linear relationship

(b) Positive, becoming steeper

(c) Positive, becoming less steep

## The Slope of a Relationship

We can measure the influence of one variable on another by the slope of the relationship. The slope of a relationship is the change in the value of the variable measured on the $y$ axis divided by the change in the value of the variable measured on the $x$-axis. We use the Greek letter $\Delta$ (delta) to represent "change in." Thus $\Delta y$ means the change in the value of the variable measured on the $y$-axis, and $\Delta x$ means the change in the value of the variable measured on the $x$-axis. Therefore the slope of the relationship is

$$
\text { Slope }=\frac{\Delta y}{\Delta x}
$$

If a large change in the variable measured on the $y$-axis $(\Delta y)$ is associated with a small change in the variable measured on the $x$-axis ( $\Delta x$ ), the slope is large and the curve is steep. If a small change inthe variable measured on the $y$-axis ( $\Delta y$ ) is associated with a large change in the variable measured on the $x$-axis $(\Delta x)$, the slope is small and the curve is flat.
We can make the idea of slope clearer by doing some calculations.

## The Slope of a Straight Line

The slope of a straight line is the same regardless of where on the line you calculate it. The slope of a straight line is constant. Let's calculate the slope ofthe positive relationship in Fig. A1.9. In part (a), when $x$ increases from 2 to $6, y$ increases from 3 to 6 . The change in $x$ is +4 that is, $\Delta$ is 4 . The change in $y$ is +3 -that is, $\Delta y$ is 3 . The slope of that line is

$$
\frac{\Delta y}{\Delta x}=\frac{3}{4}
$$

In part (b), when $x$ increases from 2 to $6, y$ decreases from 6 to 3 . The change in $y$ is $-3-$ that is $\Delta$ is -3 . The change in $x$ is plus $4-$ that is, $\Delta x$ is 4 . The slope of the curve is,

$$
\frac{\Delta y}{\Delta x}=\frac{-3}{4}
$$

FIGURE A1.9 The Slope of a Straight Line

(a) Positive slope

To calculate the slope of a straight line, we divide the change in the value of the variable measured on the $y$-axis $(\Delta y)$ by the change in the value of the variable measured on the $x$ axis $(\Delta x)$ as we move along the line.

Part (a) shows the calculation of a positive slope. When $x$ increases from 2 to $6, \Delta x$ equals 4 . That change in $x$

(b) Negative slope
brings about an increase in $y$ from 3 to 6 , so $\Delta y$ equals 3 . The slope $(\Delta y / \Delta x)$ equals $3 / 4$.

Part (b) shows the calculation of a negative slope. When $x$ increases from 2 to $6, \Delta x$ equals 4 . That increase in $x$ brings about a decrease in $y$ from 6 to 3 , so $\Delta y$ equals -3 . The slope $(\Delta y / \Delta x)$ equals $-3 / 4$.

Notice that the two slopes have the same magnitude (3/4), but the slope of the line in part (a) is positive $(+31+4=3 / 4)$ while that in part (b) is negative $(-3 /+4=-3 / 4)$. The slope of a positive relationship is positive; the slope of a negative relationship is negative.

## The Slope of a Curved Line

The slope of a curved line is trickier. The slope of a curved line is not constant, so the slope depends on where on the curved line we calculate it. There aretwo ways to calculate the slope of a curved line: Youcan calculate the slope at a point, or you can calculate the slope across an arc of the curve. Let's look at the two alternatives.

Slope at a Point To calculate the slope at a point on a curve, you need to construct a straight line that has the same slope as the curve at the point in question. Figure AI. 10 shows how this is done. Suppose you want to calculate the slope of the curve at point A. Place a ruler on the graph so that the ruler touches point $A$ and no other point on the curve, then draw a straight line along the edge of the ruler. The straight red line is this line, and it is the tangent to the curve at point $A$. If he ruler touches the curve only at point $A$, then the slope of the curve at point $A$ must be the same as the slope of the edge of the ruler. If the curve and the ruler do not have the same slope, the line along the edge of the ruler will cut the curve instead of just touching it.
figure al. 10 Slope at a Point


To calculate the slope of the curve at point $A$, draw the red line that just touches the curve at $A$-the tangent. The slope of this straight line is calculated by dividing the change in $y$ by the change in $x$ along the red line. When $x$ increases from 0 to $4, \Delta x$ equals 4 . That change in $x$ is associated with an increase in $y$ from 2 to 5 , so $\Delta y$ equals 3 . The slope of the red line is $3 / 4$, so the slope of the curve at point $A$ is $3 / 4$.
Now that you have found a straight line with the same slope as the curve at point $A$, you can calculate the slope of the curve at point $A$ by calculating the slope of the straight line Along the straight line, as $x$ increases from 0 to 4 ( $\Delta x$ is 4 ) $y$ increases from 2 to 5 ( $\Delta y$ is 3 ). Therefore the slope of the straight line is

$$
\frac{\Delta y}{\Delta x}=\frac{3}{4}
$$

So the slope of the curve at point $A$ is $3 / 4$.
Slope Across an Arc An arc of a curve is a piece of a curve. Figure AI . 11 shows the same curve as in Fig. Al.10, but instead of calculating the slope at point $A$, we are now going to calculate the slope across the arcfrom point $B$ to point $C$. You can see that the slope ofthe curve at point $B$ is greater than at point C . When we calculate the slope across an arc, we are calculating the average slope between two points. As we move along the arc from $B$ to $C, x$ increases from 3 to 5 and $y$ increases from 4.0 to 5.5 . The change in $x$ is $2(\Delta x$ is 2$)$, and the change in $y$ is $1.5(\Delta y$ is 1.5). Therefore the slope is

$$
\frac{\Delta y}{\Delta x}=\frac{1.5}{2}=\frac{3}{4}
$$

## figure Al. 11 Slope Across an Arc



To calculate the average slope of the curve along the arc $B C$, draw a straight line from point $B$ to point $C$. The slope of the line $B C$ is calculated by dividing the change in $y$ by the change in $x$. In moving from $B$ to $C$, the increase in $x$ is 2 ( $\Delta x$ equals 2 ) and the change in $y$ is 1.5 ( $\Delta y$ equals 1.5 ). The slope of the line $B C$ is 1.5 divided by 2 , or $3 / 4$. So the slope of the curve across the arc $B C$ is $3 / 4$.

So the slope of the curve across the $\operatorname{arc} B C$ is $3 / 4$.

This calculation gives us the slope of the curve between points $B$ and $C$. The actual slope calculated is the slope of the straight line from $B$ to $C$. This slope approximates the average slope of the curve along the arc $B C$. In this particular example, theslope across the arc $B C$ is identical to the slope of the curve at point $A$, but the calculation of the slope of acurve does not always work out so neatly. You might have fun constructing some more examples and a few counterexamples.

You now know how to make and interpret a graph. So far, we've limited our attention to graphs of two variables. We're now going to learn how to graphmore than two variables.

## Graphing Relationships Among More Than Two Variables

We have seen that we can graph the relationship between two variables as a point formed by the $x$-and $y$-coordinates in a two-dimensional graph. You might be thinking that although a two-dimensional graph is informative, most of the things in which you are likely to be interested involve relationships among many variables, not just two. For example, the
amount of ice cream consumed depends on the price of ice cream and the temperature. If ice cream is expensive and the temperature is low, people eat much less ice cream than when ice cream is inexpensive and the temperature is high. For any given price of ice cream, the quantity consumed varies with the temperature; and for any given temperature, the quantity of ice cream consumed varies with its price. Figure A1.12 shows a relationship among three variables. The table shows the number of gallons of ice cream consumed each day at two different temperatures and at a number of different prices of ice cream. How can we graph these numbers?
To graph a relationship that involves more than two variables, we use the ceteris paribus assumption.

Ceteris Paribus (often shortened to cetpar) means "if all other relevant things remain the same." To isolate the relationship of interest in a laboratory experiment a scientist holds everything constant except for the variable whose effect is being studied. Economists use the same method to graph a relationship that has more than two variables. (Fig. A1.12 on page 63 shows an example) There, you can see what happens to the quantity of ice cream consumed when the price of ice cream varies but the temperature is held constant.

## When Other Things Change

The temperature is held constant along each of the curves in Fig. A1.12, but in reality the temperature changes. When that event occurs, you can think of what happens in the graph as a shift of the curve. When the temperature rises from $70^{\circ} \mathrm{F}$ to $90^{\circ} \mathrm{F}$, the curve that shows the relationship between ice cream consumption and the price of ice cream shifts rightward from the blue curve to the red curve. You will encounter these ideas of movements along and shifts of curves at many points in your study of economics. Think carefully about what you've just learned and make up some examples (with assumed numbers) about other relationships.

With what you have learned about graphs, you can move forward with your study of economics. There areno graphs in this book that are more complicated thanthose that have been explained in this appendix.

## MATHEMATICAL NOTE

## Equations of Straight Lines

If a straight line in a graph describes the relationshipbetween two variables, we call it a linear relationship. Figure 1 shows the linear relationship between a person's expenditure and income. This person spends $\$ 100$ a week (by borrowing or spending previous savings) when income is zero. Out of each dollar earned, this person spends 50 cents (and saves 50 cents). All linear relationship are described by the same general equation. We call the quantity that is measured on the horizontal axis (or x-axis) $x$, and we call the quantity that is measured on the vertical axis (or $y$-axis) $y$. In the case of Fig.1, $x$ is income and $y$ is expenditure.

## A Linear Equation

The equation that describes a straight-line relation- ship between $x$ and $y$ is

$$
y=a+b x
$$

In this equation, $a$ and $b$ are fixed numbers and they are called constants. The values of $x$ and $y$ vary, so these numbers are called variables. Because the equation describes a straight line, the equation is called a linear equation.
The equation tells us that when the value of $x$ is zero, the value of $y$ is $a$. We call the constant $a$ the axis intercept. The reason is that on the graph the straight line hits they-axis at a value equal to $a$.

Figure 1 illustrates they-axis intercept.
For positive values of $x$, the value of $y$ exceeds $a$.
The constant $b$ tells us by how much $y$ increases above $a$ as $x$ increases. The constant $b$ is the slope of the line.

## Slope of Line

As we explain in the chapter, the slope of a relation- ship is the change in the value of $y$ divided by thechange in the value of $x$. We use the Greek letter !),(delta) So $\Delta y$ means the change in the value of the variable measured on the $y$-axis, and/).. $x$ means the change in the value of the variable measured on the $x$-axis. Therefore the slope of the relationship is

$$
\text { Slope }=\frac{\Delta y}{\Delta x}
$$

To see why the slope is $b$, suppose that initially the value of $x$ is $x i$, or $\$ 200$ in Fig. 2. The corresponding value of $y$ is $y_{1}$, also $\$ 200$ in Fig. 2. The equation of the line tells us that

$$
\begin{equation*}
y_{1}=a+b x_{1} . \tag{1}
\end{equation*}
$$

Now the value of $x$ increases by $\Delta x$ to $x_{1}+\Delta x$ (or $\$ 400$ in Fig. 2). And the value of $y$ increases by $\Delta y$ to $Y I+\Delta$ (or $\$ 300$ in Fig. 2).
The equation of the line now tells us that

$$
\begin{equation*}
y_{1}+\Delta y=a+b\left(x_{1}+\Delta x\right) \tag{2}
\end{equation*}
$$

To calculate the slope of the line, subtract equation (1) from equation (2) to obtain

$$
\begin{equation*}
\Delta y=b \Delta x \tag{3}
\end{equation*}
$$

and now divide equation (3) by $\Delta x$ to obtain

$$
\Delta y / \Delta x=b .
$$

So the slope of the line is $b$.


Figure 1 Linear Relationship


Figure 2 Calculating Slope

## Position of a Line

The $y$-axis intercept determines the position of the line on the graph. Figure 3 illustrates the relationship between the $y$-axis intercept and the position of the line. Inthis graph, the $y$-axis measures saving and the $x$-axis measures income.
When the $y$-axis intercept, $a$, is positive, the line hits the $y$-axis at a positive value of $y$-as the first line (from the top) does. Its $y$-axis intercept is 100 . When the $y$-axis intercept, $a$, is zero, the line hits they-axis at the origin- as the second line does. Itsy-axis intercept is 0 . When the $y$-axis intercept, $a$, is negative, the line hits the $y$-axis at a negative value of $y$ as the third line does. Its $y$-axis intercept is -100 .
As the equations of the three lines show, the value of the $y$-axis intercept does not influence the slope ofthe line. All three lines have a slope equal to 0.5 .

## Positive Relationships

Figure 1 shows a positive relationship-the two variables $x$ and $y$ move in the same direction. All positive relationships have a slope that is positive. In the equation of the line, the constant $b$ is positive. In thisexample, the $y$-axis intercept, $a$, is 100 . The slope $b$ equals $\Delta y / \Delta x$, which in Fig. 2 is $100 / 200$ or 0.5 . The equation of the line is

$$
y=100+0.5 x
$$

## Negative Relationships

Figure 4 shows a negative relationship - the two variables $x$ and $y$ move in the opposite direction. All negative relationships have a slope that is negative. In the equation of the line, the constant $b$ is negative. In the example in Fig. 4, the $y$-axis intercept, $a$, is 30 . The slope, $b$, equals $\Delta y / \Delta x$, which is $-20 / 2$ or -10 . The equation of the line is

$$
y=30+(-10) x
$$

or

$$
y=30-10 x .
$$

## Example

A straight line has ay-axis intercept of 50 and a slope of 2 . What is the equation of this line?
The equation of a straight line is

$$
y=a+b x
$$

where $a$ is the $y$-axis intercept and $b$ is the slope. So the equation is

$$
y=50+2 x
$$



Figure 3 The $\boldsymbol{y}$-Axis Intercept


Figure 4 Negative Relationship

## CHAPTER 2: The Economic Problem

After studying this chapter, you will be able to:

- Define the production possibilities frontier and use it to calculate opportunity cost
- Distinguish between production possibilities andpreferences and describe an efficient allocation of resources
- Explain how current production choices expand future production possibilities
- Explain how specialization and trade expand production possibilities
- Describe the economic institutions that coordinate decisions


## Hydraulic fracturing, or "fracking," is expanding

U.S. oil and gas production and cutting our imports. Should we produce more oil and gas?

How do weknow when we are using our energy and other re- sources efficiently?
In this chapter, you study an economic model that answers questions about the efficiency of production and trade.
At the end of the chapter, in Economics in the News, we'll apply what you learn to understand how frocking is expanding our production possibilities butwhy we are still better off importing rather than producing some of our oil and gas.

## Production Possibilities and Opportunity Cost

Every working day, in mines, factories, shops, and offices and on farms and construction sites across the United States, 142 million people produce a vast variety of goods and services valued at $\$ 60$ billion.
But the quantities of goods and services that we can produce are limited by our available resources and by technology. And if we want to increase our production of one good, we must decrease our production of something else-we face a tradeoff You are now going to study the limits to production.

The production possibilities frontier (PPF) is the boundary between those combinations of goods and services that can be produced and those that cannot. Toillustrate the PPF, we look at a model economy in which the quantities produced of only two goods change, while the quantities produced of all the other goods and services remain the same.
Let's look at the production possibilities frontier for cola and pizza, which represent any pair of goods or services.

## Production Possibilities frontier

The production possibilities frontier for cola and pizza shows the limits to the production of these twogoods, given the total resources and technology avail- able to produce them. Figure 2.1 shows this production possibilities frontier. The table lists combinations of the quantities of pizza and cola that can be produced in a month and the figure graphs these combinations. The $x$-axis shows the quantity of pizzas produced, and the $y$-axis shows the quantity of cola produced.
The PPP illustrates scarcity because the points out- side the frontier are unattainable. These points describe wants that can't be satisfied.
We can produce at any point inside the PPF or on the PPF. These points are attainable. For example, we can produce 4 million pizzas and 5 million cans ofcola. Figure 2.1 shows this combination as point $E$ on the graph and as possibility $E$ in the table.
Moving along the PPF from point $E$ to point $D$ (possibility $D$ in the table) we produce more cola andless pizza: 9 million cans of cola and 3 million pizzas. Or moving in the opposite direction from point $E$ to point $F$ (possibility $F$ in the table), we produce morepizza and less cola: 5 million pizzas and no cola.

## Production Efficiency

We achieve production efficiency if we produce goods and services at the lowest possible cost. This outcome occurs at all the points on the PPR. At points inside the PPP, production is inefficient because we are giving up more than necessary of one good to produce a given quantity of the other good.
For example, at point $Z$ in Fig. 2.1, we produce 3 million pizzas and 5 million cans of cola, but we have enough resources to produce 3 million pizzas and 9 million cans of cola. Our pizzas cost more cola than necessary. We can get them for a lower cost. Onlywhen we produce on the PPP do we incur the lowest possible cost of production. Production inside the PPP is inefficient because resources are either unused or misallocated or both. Resources are unused when they
are idle but could be working. For example, we might leave some of the factories idle or some workers unemployed.
Resources are misallocated when they are assigned to tasks for which they are not the best match. For example, we might assign skilled pizza chefs to work in a cola factory and skilled cola workers to cook pizza in a pizzeria. We could get more pizzas andmore cola if we reassigned these workers to the tasks that more closely match their skills.

## Tradeoff Along the PFF

Achoice along the PPPinvolves a tradeoff. Tradeoffs like that between cola and pizza arise in every imaginable real-world situation in which a choice must be made. At any given time, we have a fixed amount of labor, land, capital, and entrepreneurship and a given state of technology. We can employ these resources and technology to produce goods and services, but we are limited in what we can produce.
When doctors want to spend more on AIDS and cancer research, they face a tradeoff: more medical research for less of some other things. When Congress wants to spend more on education and healthcare, it faces a tradeoff: more education and healthcare for less national defense or homeland security. When anenvironmental group argues for less logging, it is suggesting a tradeoff: greater conservation of endangeredwildlife for less paper. When you want a higher grade on your next test, you face a tradeoff: spend moretime studying and less leisure or sleep time.
All the tradeoffs you've just considered involve a cost - an opportunity cost.

## Opportunity Cost

The opportunity cost of an action is the highest-valued alternative forgone. The PPP makes this idea preciseand enables us to calculate opportunity cost. Alongthe PPP, there are only two goods, so there is only one alternative forgone: some quantity of the other good. To produce more pizzas we must produce less cola. The opportunity cost of producing an additional pizza is the cola we must forgo. Similarly, theopportunity cost of producing an additional can of cola is the quantity of pizza we must forgo.
In Fig. 2.1, if we move from point $C$ to point $D$, we produce an additional 1 million pizzas but 3 mil- lion fewer cans of cola. The additional 1 million pizzas cost 3 million cans of cola. Or 1 pizza costs 3 cans of cola. Similarly, if we move from $D$ to $C$, we produce an additional 3 million cans of cola but 1 million fewer pizzas. The additional 3 million cans of cola cost 1 million pizzas. Or 1 can of cola costs $1 / 3$ of a pizza.

Opportunity Cost Is a Ratio Opportunity cost is a ratio. It is the decrease in the quantity produced of one good divided by the increase in the quantity produced of another good as we move along the production possibilities frontier.
Because opportunity cost is a ratio, the opportunity cost of producing an additional can of cola is equal to the inverse of the opportunity cost of producing an additional pizza. Check this proposition by returning to the calculations we've just done. In the move from $C$ to $D$, the opportunity cost of a pizza is 3 cans of cola. And in the move from $D$ to C , the opportunity cost of a can of cola is 113 of a pizza. So the opportunity cost of pizza is the inverse of the opportunity cost of cola.
figure 2.1 Production Possibilities
Frontier


The table lists six production possibilities for cola and pizzas. Row $A$ tells us that if we produce no pizzas, the maximum quantity of cola we can produce is 15 million cans. Points $A, B, C, D, E$, and $F$ in the figure represent the rows of the table. The curve passing through these points is the production possibilities frontier (PPF).

The PPF separates the attainable from the unattainable. Production is possible at any point inside the orange area or on the frontier. Points outside the frontier are unattainable. Points inside the frontier, such as point $Z$, are inefficient because resources are wasted or misallocated. At such points, it is possible to use the available resources to produce more of either or both goods.

Increasing Opportunity Cost The opportunity cost of a pizza increases as the quantity of pizzas produced increases. The outward-bowed shape of the PPP reflects increasing opportunity cost. When we produce a large quantity of cola and a small quantity of pizza- between points $A$ and $B$ in Fig. 2.1 -the frontier has a gentle slope. An increase in the quantity of pizzas costs a small decrease in the quantity of cola-the opportunity cost of a pizza is a small quantity of cola.
When we produce a large quantity of pizzas and a small quantity of cola-between points $E$ and PinFig. 2.1-the frontier is steep. A given increase in the quantity of pizzas costs a large decrease in the quantity of cola, so the opportunity cost of a pizza is a large quantity of cola.

## Using Resources Efficiently

We achieve production efficiency at every point on the PPP, but which of these points is best? The answer is the point on the PPF at which goods and services are produced in the quantities that provide the greatest possible benefit. When goods and services are produced at the lowest possible cost and in the quantities that provide the greatest possible benefit, we have achieved allocative efficiency.
The questions that we raised when we reviewed the four big issues in Chapter 1 are questions about allocative efficiency. To answer such questions, we must measure and compare costs and benefits.

## The PFF and Marginal Cost

The marginal cost of a good is the opportunity costof producing one more unit of it. We calculate marginal cost from the slope of the PPF. As the quantity of pizzas produced increases, the PPF gets steeper and the marginal cost of a pizza increases. Figure 2.2 illustrates the calculation of the marginal cost of a pizza.
Begin by finding the opportunity cost of pizza in blocks of 1 million pizzas. The cost of the first mil-lion pizzas is 1 million cans of cola; the cost of thesecond million pizzas is 2 million cans of cola; thecost of the third million pizzas is 3 million cans of cola, and so on. The bars in part (a) illustrate thesecalculations.
The bars in part (b) show the cost of an average pizza in each of the 1 million pizza blocks. Focus on the third million pizzas-the move from $C$ to $D$ in part (a). Over this range, because 1 million pizzas cost 3 million cans of cola, one of these pizzas, on average, costs 3 cans of cola-the height of the bar in part (b).
Next, find the opportunity cost of each additional pizza-the marginal cost of a pizza. The marginal cost of a pizza increases as the quantity of pizzas produced increases. The marginal cost at point $C$ is less than it is at point $D$. On average over the range from $C$ to $D$, the marginal cost of a pizza is 3 cans of cola. But it exactly equals 3 cans of cola only in the middle of the range between $C$ and $D$.
The middle dot in part (b) indicates that the marginal cost of a pizza is 3 cans of cola when 2.5 million pizzas are produced. Each dot in part (b) is interpretedin the same way. The curve that passes through these dots, labeled $M C$, is the marginal cost curve. It shows the marginal cost of a pizza at each quantity of pizzas as we move along the PPF.

FIGURE 2.2 The PPF and Marginal Cost

(a) PPF and opportunity cost

(b) Marginal cost

Marginal cost is calculated from the slope of the PPF. As the quantity of pizzas produced increases, the PPF gets steeper and the marginal cost of a pizza increases. The bars in part (a) show the opportunity cost of pizza in blocks of 1 million pizzas. The bars in part (b) show the cost of an average pizza in each of these 1 million blocks. The red curve, MC, shows the marginal cost of a pizza at each point along the PPF. This curve passes through the center of each of the bars in part (b).

## Preferences and Marginal Benefit

The marginal benefit from a good or service isthe benefit received from consuming one more unit of it. This benefit is subjective. It depends on people's preferences-people's likes and dislikes and the intensity of those feelings.
Marginal benefit and preferences stand in sharp contrast to marginal costand production possibilities. Preferences describe what people like and want and the production possibilities describe the limits or constraints on what is feasible.
We need a concrete way of illustrating preferences that parallels the way we illustrate the limits to production using the PPP.
The device that we use to illustrate preferences is the marginal benefit curve, which is a curve that shows the relationship between the marginal benefit from a good and the quantity consumed of that good. Note that the marginal benefit curve is unrelated to the PPP and cannot be derived from it.

We measure the marginal benefit from a good or service by the most that people are willing to pay for an additional unit of it. The idea is that you are willing to pay less for a good than it is worth to you but you are not willing to pay more: The most you are willing to pay for something is its marginal benefit.
It is a general principle that the more we have of any good or service, the smaller is its marginal benefit and the less we are willing to pay for an additional unit of it. This tendency is so widespread and strong that we call it a principle - the principle of decreasing marginal benefit. The basic reason why marginal benefit decreases is that we like variety. The more we consume of any one good or service, the more we tire of it and would prefer to switch to something else.

Think about your willingness to pay for a pizza. If pizza is hard to come by and you can buy only a few slices a year, you might be willing to pay a high price to get an additional slice. But if pizza is all you've eaten for the past few days, you are willing to pay almost nothing for another slice.
You've learned to think about cost as opportunity cost, not as a dollar cost. You can think about marginal benefit and willingness to pay in the same way. The marginal benefit, measured by what you are willing to pay for something, is the quantity of other goods and services that you are willing to forgo. Let's continue with the example of cola and pizza and illustrate preferences this way.

Figure 2.3 illustrates preferences as the willingness to pay for pizza in terms of cola. In row $A$, with 0.5 million pizzas available, people are willing to pay 5 cans of cola per pizza. As the quantity of pizzas increases, the amount that people are willing to pay for a pizza falls. With 4.5 million pizzas available, people are willing to pay only 1 can of cola per pizza. Let's now use the concepts of marginal cost and marginal benefit to describe allocative efficiency.

FIGURE 2.3 Preferences and the Marginal Benefit Curve


The smaller the quantity of pizzas available, the more cola people are willing to give up for an additional pizza. With 0.5 million pizzas available, people are willing to pay 5 cans of cola per pizza. But with 4.5 million pizzas, people are willing to pay only 1 can of cola per pizza. Willingness to pay measures marginal benefit. A universal feature of people's preferences is that marginal benefit decreases.

## Allocative Efficiency

At any point on the PPP, we cannot produce more of one good without giving up some other good. At the best point on the PPE we cannot produce more of one good without cola, and the marginal benefit from a pizza is 4 cans of cola. Because someone values an additional pizza more highly than it costs to produce, we can get more value from our resources by moving some of them out of producing cola and into producing pizza. Now suppose we produce 3.5 million pizzas. The marginal cost of a pizza is now 4 cans of cola, but

FIGURE 2.4 Efficient Use of Resources

(b) Marginal benefit equals marginal cost

The greater the quantity of pizzas produced, the smaller is the marginal benefit (MB) from pizza-the less cola people are willing to give up to get an additional pizza. But the greater the quantity of pizzas produced, the greater is the marginal cost ( MC ) of a pizza-the more cola people must give up to get an additional pizza. When marginal benefit equals marginal cost, resources are being used efficiently.
the marginal benefit from a pizza is only 2 cans of cola. Because the additional pizza costs more to produce than anyone thinks it is worth, we can get more value from our resources by moving some of them away from producing pizza and into producing cola.

Suppose we produce 2.5 million pizzas. Marginal cost and marginal benefit are now equal at 3 cans of cola. This allocation of resources between pizzas and cola is efficient. If more pizzas are produced, the forgone cola is worth more than the additional pizzas. If fewer pizzas are produced, the forgone pizzas are worth more than the additional cola.

## Economic Growth

During the past 30 years, production per person in the United States has doubled. The expansion of production possibilities is called economic growth. Economic growth increases our standard ofliving, but it doesn't overcome scarcity and avoid opportunity cost. To make our economy grow, we face a tradeoff: the faster we make production grow, the greater is the opportunity cost of economic growth.

## The Cost of Economic Growth

Economic growth comes from technological change and capital accumulation. Technological change isthe development of new goods and of better ways of producing goods and services. Capital accumulation is the growth of capital resources, including human capital. Technological change and capital accumulation have vastly expanded our production possibilities. We can produce automobiles that provide us with more transportation than was available when we had onlyhorses and carriages. We can produce satellites thatprovide global communications on a much larger scale than that available with the earlier cable technology. But if we use our resources to develop new technologies and produce capital, we must decrease our production of consumption goods and services. New technologies and new capital have an opportunity cost. Let's look at this opportunity cost.

Instead of studying the PPP of pizzas and cola, we'll hold the quantity of cola produced constant and examine the PPP for pizzas and pizza ovens. Figure 2.5 shows this $P P P$ as the blue curve $P P P$. If we devote no resources to producing pizza ovens, we produce at point $A$. If we produce 3 million pizzas, we can produce 6 pizza ovens at point $B$. If we produce no pizza, we can produce 10 ovens at point $C$.
The amount by which our production possibilities expand depends on the resources we devote to technological change and capital accumulation. If we devote no resources to this activity (point A), our PPPremains the blue curve PPPo in Fig. 2.5. If we cut the current pizza production and produce 6 ovens (point B), then in the future, we'll have more capital and our $P P P$ will rotate outward to the position shown by the red curve $P P P_{1}$. The fewer resources we use for producing pizza and the more resources we use for producing ovens, the greater is the expansion of our future production possibilities.

Economic growth brings enormous benefits in the form of increased consumption in the future, but economic growth is not free and it doesn't abolish scarcity.
In Fig. 2.5, to make economic growth happen we must use some resources to produce new ovens, which leaves fewer resources to produce pizzas. To move to $B^{\prime}$ in the future, we must move from $A$ to $B$ today. The opportunity cost of more pizzas in the future is fewer pizzas today. Also, on the new PPP, we still face a tradeoff and opportunity cost.

The ideas about economic growth that we have explored in the setting of the pizza industry also apply to nations. Hong Kong and the United States provide a striking case study.
figure 2.5 Economic Growth


PPFo shows the limits to the production of pizzas and pizza ovens, with the production of all other goods and services remaining the same. If we devote no resources to producing pizza ovens and produce 5 million pizzas, our production possibilities will remain the same at $P P F_{0}$. But if we decrease pizza production to 3 million and produce 6 ovens, at point $B$, our production possibilities expand. After one period, the PPF rotates outward to $P P F_{1}$ and we can produce at point $B^{\prime}$, a point outside the original $P P F_{0}$. We can rotate the PPF outward, but we cannot avoid opportunity cost. The opportunity cost of producing more pizzas in the future is fewer pizzas today.

## ECONOMICS IN ACTION

## Hong Kong Catching Up to the United States

In 1963, the production possibilities per person inthe United States were more than four times those in Hong Kong (see the figure). The United States devotes one fifth of its resources to accumulating capital, and in 1963, the United States was at point A on its PPP. Hong Kong devotes one third of its resources to accumulating capital, and in 1963, Hong Kong was at point $A$ on its PPF.

Since 1963, both economies have experienced economic growth, but because Hong Kong devotes abigger fraction of its resources to accumulating capital, its production possibilities have expanded more quickly.
By 2013, production possibilities per person in Hong Kong equaled those in the United States. If Hong Kong continues to devote more resources to accumulating capital than the United States does (at point B on its 2013 PPP), Hong Kong will continue to grow more rapidly. But if Hong Kong decreases its capital accumulation (moving to point D on its 2013 PPP), then its rate of economic growth will slow.
Hong Kong is typical of the fast-growing Asian economies, which include Taiwan, Thailand, South Korea, China, and India. Production possibilities

## A Nation's Economic Growth

The experiences of the United States and Hong Kong make a striking example of the effects of our choices about consumption and capital accumulation on the rate of economic growth.
If an economy devotes all its factors of production to producing consumption goods and services and none to advancing technology and accumulating capital, its production possibilities in the future will be the same as they are today. To expand production possibilities in the future, a nation or an economy must devote fewer resources to producing current consumption goods and services and some resources to accumulating capital and developing new technologies. As production possibilities expand, consumption in the future can increase. The decrease in today's consumption is the opportunity cost of tomorrow's increase in consumption.


Economic Growth in the United States and Hong Kong

## Gains From Trade

People can produce for themselves all the goods andservices that they consume, or they can produce one good or a few goods and trade with others. Producingonly one good or a few goods is called specialization. We are going to learn how people gain by specializingin the production of the good in which they have a comparative advantage and trading with others.

## Comparative Advantage and Absolute Advantage

A person has a comparative advantage in an activity ifthat person can perform the activity at a lower opportunity cost than anyone else. Differences in opportunity costs arise from differences in individual abilities and from differences in the characteristics of other resources. No one excels at everything. Although no one excels at everything, some people excel and can outperform others in a large number of activities - perhaps even in all activities. A person who is more productive than others has an absolute advantage.
Absolute advantage involves comparing productivities - production per hour - whereas comparative advantage involves comparing opportunity costs.
A person who has an absolute advantage does not have a comparative advantage in every activity. Because ability and resources vary from one person to another, people have different opportunity costs of producing various goods. These differences in opportunity cost are the source of comparative advantage.
Let's explore the idea of comparative advantage by looking at two smoothie bars: one operated by Liz and the other operated by Joe.

Joe's Smoothie Bar Joe produces smoothies and salads in a small, low-tech bar. He has only one blender, and it's a slow, old machine that keeps stopping. Even if Joe uses all his resources to produce smoothies; he can produce only 6 an hour-see Table 2.1. But Joe is good at making salads, and if he uses all his resources in this activity, he can produce 30 salads an hour. Joe's ability to make smoothies and salads is the same regardless of how he splits an hour between thetwo tasks. He can make a salad in 2 minutes or asmoothie in 10 minutes. For each additional smoothieJoe produces, he must decrease his production of salads by 5. And for each additional salad he produces, he must decrease his production of smoothies by $1 / 5$ of a smoothie. So Joe's opportunity cost of producing 1 smoothie is 5 salads, and Joe's opportunity cost of producing 1 salad is $1 / 5$ of a smooth i.e. Joe's customers buy smoothies and salads in equal quantities. So Joe spends 50 minutes of each hour making smoothies and 10 minutes of each hour making salads. With this division of his time, Joe produces 5 smoothies and 5 salads an hour.
Figure 2.6(a) illustrates the production possibilities at Joe's smoothie bar-Joe's PPF. Joe's PPF is linear (not outward bowed) because his ability to produce salads and smoothies is the same no matter how he divides his time between the two activities. Joe's opportunity cost of a smoothie is constant - it is the same at all quantities of smoothiesproduced.

Liz's Smoothie Bar Liz also produces smoothies and salads but in a high-tech bar that is much more productive than Joe's. Liz can turn out either a smoothie or a salad every 2 minutes see Table 2.2. If Liz spends all her time making smoothies, she can produce 30 an hour. And if she spends all her time making salads, she can also produce 30 an hour.

Liz's ability to make smoothies and salads, like Joe's, is the same regardless of how she divides her time between the two tasks. She can make a salad in 2 minutes or a smoothie in 2 minutes. For each additional smoothie Liz produces, she must decrease her production of salads by 1. And for each additional salad she produces, she must decrease her production of smoothies by 1 . So Liz's opportunity cost of producing 1 smoothie is 1 salad. Liz's opportunity cost of producing 1 salad is 1 smoothie.
Liz's customers buy smoothies and salads in equal quantities, so she splits her time equally between the two items and produces 15 smoothies and 15 salads an hour.
Figure 2.6(b) illustrates the production possibilities at Liz's smoothie bar-Liz's PPP. Like Joe's, Liz's PPP is linear because her ability to produce salads and smoothies is the same no matter how she divides her time between the two activities. Liz's opportunity cost of a smoothie is 1 salad at all quantities of smoothies produced.

| TABLE 2.2 | Liz'sProduction Possibilities <br> Minutes to <br> produce 1 | Quantity <br> per hour |
| :--- | :---: | :---: |
| Item | 2 | 30 |
| Smoothies | 2 | 30 |

Joe's Comparative Advantage In which of the twoactivities does Joe have a comparative advantage? To answer this question, first recall the definition of comparative advantage. A person has a comparative advantage when that person's opportunity cost of producing a good is lower than another person's opportunity cost of producing that same good. Joe's opportunity cost of producing a salad is only $1 / 5$ of a smoothie, while Liz's opportunity cost of producing a salad is 1 smoothie. So Joe has a comparative advantage in producingsalads.

Liz's Comparative Advantage If Joe has a comparative advantage in producing salads; Liz must have a comparative advantage in producing smoothies. Check the numbers. For Joe, a smoothie costs 5 salads, and for Liz, a smoothie costs only 1 salad. So Liz has a comparative advantage in making smoothies.

## Achieving the Gains from Trade

Liz and Joe run into each other one evening in a singles bar. After a few minutes of getting acquainted, Liz tells Joe about her amazing smoothie business.
Her only problem, she tells Joe, is that she would like to produce more because potential customers leave when her lines get too long.
Joe doesn't want to risk spoiling a blooming relationship by telling Liz about his own struggling business, but he takes the risk. Joe explains to Liz that hespends 50 minutes of every hour making 5 smoothies and 10 minutes making 5 salads. Liz's eyes pop. "Have I got a deal for you!" she exclaims.

FIGURE 2.6 The Production Possibilities Frontiers

(a) Joe

Joe can produce 30 salads per hour, 1 every two minutes, if he produces no smoothies. Or, he can produce 6 smoothies per hour, 1 every 10 minutes, if he produces no salads. Joe's customers buy equal quantities of salads and smoothies, so Joe produces 5 of each. His opportunity cost of a smoothie is 5 salads.

(b) Liz

Liz can produce 30 salads or 30 smoothies per hour, 1 of either item every two minutes. Liz's customers buy equal quantities of salads and smoothies, so she produces 15 of each. Liz's opportunity cost of a smoothie is 1 salad.

Liz's Proposal Here's the deal that Liz sketches on a paper napkin. Joe stops making smoothies and allocates all his time to producing salads; Liz stops making salads and allocates all her time to producing smoothies. That is, they both specialize in producingthe good in which they have a comparative advantage. Together they produce 30 smoothies and 30 salads-see Table 2.3(b).
They then trade. Liz suggests trading at a price of 2 salads per smoothie. For her, that is a good deal because she can produce a smoothie at a cost of 1 salad and sell it to Joe for 2 salads. It is also a good deal for Joe because he can produce a salad at a cost of $1 / 5$ of a smoothie and sell it to Liz for $1 / 2$ a smoothie.
Liz explains that any price above 1 salad per smoothie is good for her and any price below 5 saladsper smoothie is good for Joe, so a price of 2 salads per smoothie lets them both gain, as she now describes.

At the proposed price, Liz offers to sell Joe 10 smoothies in exchange for 20 salads. Equivalently, Joe sells Liz 20 salads in exchange for 10 smoothies. [see Table 2.3(c)] After this trade, Joe has 10 salads-the 30 he produces minus the 20 he sells to Liz. He also has the 10 smoothies that he buys from Liz. So Joe now has increased the quantities of smoothies and salads that he can sell to his customers-see Table 2.3(d).
Liz has 20 smoothies - the 30 she produces minus the 10 she sells to Joe. She also has the 20 salads that she buys from Joe. Liz has increased the quantities of smoothies and salads that she can sell to her customers - see Table 2.3(d). Both Liz and Joe gain 5 smoothies and 5 salads an hour-see Table 2.3(e).

Illustrating Liz's Idea To illustrate her idea, Liz grabs a fresh napkin and draws the graphs in Fig. 2.7. First, she sketches Joe's PPP in part (a) and shows the point at which he is producing before they meet.
Recall that he is producing 5 smoothies and 5 salads an hour at point $A$.
She then sketches her own PPF in part (b), and marks the point $A$ at which she is producing 15 smoothies and 15 salads an hour.

TABLE 2.3 Liz and Joe Gain from Trade

| (a) Before trade | Liz | Joe |
| :--- | :---: | :---: |
| Smoothies | 15 | 5 |
| Salads | 15 | 5 |
| (b) Specialization | Liz | Joe |
| Smoothies | 30 | 0 |
| Salads | 0 | 30 |
| (c) Trade | Liz | Joe |
| Smoothies | sell 10 | buy 10 |
| Salads | buy 20 | sell 20 |
| (d) After trade | Liz | Joe |
| Smoothies |  | 20 |
| Salads | 20 | 10 |
| (e) Gains from trade | Liz | Joe |
| Smoothies | +5 | +5 |
| Salads | +5 | +5 |

She then shows what happens when they each specialize in producing the good in which they have a comparative advantage. Joe specializes in producing salads and produces 30 salads and no smoothies at point Bon his PPP.
Liz specializes in producing smoothies and produces 30 smoothies and no salads at point $B$ on her PPE. They then trade smoothies and salads at a price of 2 salads per smoothie or $1 / 2$ a smoothie per salad.The "Trade line" that Liz draws on each part of the figure illustrates the tradeoff that each faces at theproposed price.
Liz now shows Joe the amazing outcome of her idea. After specializing and trading, Joe gets 10 smoothies and 10 salads at point C - a gain of 5 smoothies and 5 salads. He moves to a point
outside his PPP. And Liz gets 20 smoothies and 20 salads at point C -also a gain of 5 smoothies and 5 salads-and moves to a point outside her PPE.
Despite Liz being more productive than Joe, both gain from specializing at producing the good in which they have a comparative advantage and trading.

FIGURE 2.7 The Gains from Trade

(a) Joe

Initially, Joe produces at point $A$ on his PPF in part (a), and Liz produces at point $A$ on her PPF in part (b). Joe's opportunity cost of producing a salad is less than Liz's, so Joe has a comparative advantage in producing salads. Liz's opportunity cost of producing a smoothie is less than Joe's, so Liz has a comparative advantage in producing smoothies.

If Joe specializes in making salads, he produces 30 salads and no smoothies at point $B$ on his PPF. If Liz specializes in

(b) Liz
making smoothies, she produces 30 smoothies and no salads at point $B$ on her PPF. They exchange salads for smoothies along the red "Trade line." Liz buys salads from Joe for less than her opportunity cost of producing them. Joe buys smooth ies from Liz for less than his opportunity cost of producing them. Each goes to point $C$ - a point outside his or her PPF. With specialization and trade, Joe and Liz gain 5 smoothies and 5 salads each with no extra resources.

## Economic Coordination

For 7 billion people to specialize and produce millions of different goods and services, individual choices must somehow be coordinated. Two competing coordination systems have been used: centraleconomic planning and markets (see At Issue, p. 46). Central economic planning works badly because economic planners don't know people's production possibilities and preferences, so production ends up inside the PPF and the wrong things are produced. Decentralized coordination works best but to do so it needs four complementary social institutions. They are

- Firms
- Markets
- Property rights
- Money


## Firms

A firm is an economic unit that hires factors of production and organizes them to produce and sell goods and services. Firms coordinate a huge amount of economic activity. For example, Wal-Mart buys or rents largebuildings, equips them with storage shelves and checkout lanes, and hires labor. Wal-Mart directs thelabor and decides what goods to buy and sell.
But Sam Walton would not have become one of the wealthiest people in the world if Wal-Mart produced everything that it sells. He became rich by specializing in providing retail services and buyingfrom other firms that specialize in producing goods(just as Liz and Joe did). This trade needs markets.

## Markets

In ordinary speech, the word market means a placewhere people buy and sell goods such as fish, meat, fruits, and vegetables.
In economics, a market is any arrangement that enables buyers and sellers to get information and to do business with each other. An example is the world oil market, which is not a place but a network of producers, consumers, wholesalers, and brokers who buy and sell oil. In the world oil market, decision makers make deals by using the Internet. Enterprising individuals and firms, each pursuing their own self-interest, have profited by making markets - by standing ready to buy or sell items in which they specialize. But markets can work only when property rights exist.

## Property Rights

The social arrangements that govern the ownership, use, and disposal of anything that people value are called property rights. Realproperty includes land and buildings-the things we call property in ordinary speech-and durable goods such as plant and equipment. Financialproperty includes stocks and bonds and money in the bank. Intellectualproperty is the intangible product of creative effort. This type of property includes books, music, computer programs, and inventions of all kinds and is protected by copyrights and patents. Where property rights are enforced, people have the incentive to specialize and produce the goods and services in which they have a comparative advantage. Where people can steal the production of others, resources are devoted not to production but to protecting possessions.

## Money

Money is any commodity or token that is generally acceptable as a means of payment. Liz and Joe don't need money. They can exchange salads andsmoothies. In principle, trade in markets can exchange any item for any other item. But you can perhaps imagine how complicated life would be if we exchanged goods for other goods. The "invention" of money makes trading in markets much more efficient.

## Circular Flows Through Markets

Trading in markets for goods and services and factors of production creates a circular flow of expenditures and incomes. Figure 2.8 on page 83 shows the circular flows.

Households specialize and choose the quantities of labor, land, capital, and entrepreneurial services to sell or rent to firms. Firms choose the quantities of factors of production to hire. These (anti-clockwise) flows go through thefactor markets. Households choose the quantities of goods and services to buy, and firms choose the quantities to produce. These (anti-clockwise) flows go through the goods markets. Households receive incomes and make expenditures on goods and services (clockwise flows). How do markets coordinate all these decisions?

## Coordinating Decisions

Markets coordinate decisions through price adjustments. Suppose that some people who want to buy hamburgers are not able to do so. To make buying and selling plans the same, either more hamburgers must be offered for sale or buyers must scale down their appetites (or both). A rise in the price of a hamburger produces this outcome. It encourages producers to offer more hamburgers for sale and encourages some people to change their lunch plans. When the price is right, buying plans and selling plans match. Alternatively, suppose that more hamburgers are available than people want to buy. In this case, more hamburgers must be bought or fewer hamburgers must be offered for sale (or both). A fall in the price of hamburger achieves this outcome. Itencourages people to buy more hamburgers and it encourages firms to produce a smaller quantity of hamburgers.

FIGURE 2.8 Circular Flows in the Market Economy


Households and firms make economic choices and markets coordinate these choices.

Households choose the quantities of labor, land, capital, and entrepreneurial services to sell or rent to firms in exchange for wages, rent, interest, and profits. Households also choose how to spend their incomes on the various types of goods and services available.

Firms choose the quantities of factors of production to hire and the quantities of goods and services to produce.

Goods markets and factor markets coordinate these choices of households and firms.

The counterclockwise red flows are real flows-the flow of factors of production from households to firms and the flow of goods and services from firms to households.

The clockwise green flows are the payments for the red flows. They are the flow of incomes from firms to households and the flow of expenditure on goods and services from households to firms.

## ECONOMIC ANALYSIS

- The table shows some data on U.S. production and imports in 2000 and 2013.

| U.S. production and imports | $\mathbf{2 0 0 0}$ | $\mathbf{2 0 1 3}$ |
| :--- | :---: | :---: |
| Oil and gas (million barrels per day) | 16 | 21 |
| Other goods and services (units per day) | 34 | 42 |
| Oil and gas imports (million barrels per day) | 13 | 8 |

- The news story provides data on oil production and the table adds together the production of oil and natural gas, both of which come from shale formations.
- "Fracking" has increased the production of oil and gas from 16 million barrels per day in 2000 to 21 million barrels per day in 2013, an increase of 31 percent.
- The consumption of oil and gas has remained at 29 million barrels per day, so the increase in production has decreased the imports of oil and gas from 13 million barrels per day in 2000 to 8 million barrels per day in 2013.
- "Fracking" is not the only advance in technology and productivity. Other advances in technology and investment include such items as cellphone networks, Internet services, and industrial robots.
- These advances in technology and investment in capital have increased the production of other goods and services from 34 units in 2000 to 42 units in 2013, an increase of 24 percent. (Think of these "units" as giant shopping carts of goods and services each one of which costs $\$ 1$ billion in the prices of 2013.)
- We can use the data in the table to make a graph of the U.S. PPF, with other goods and services (on the $y$-axis), and oil and gas (on the $x$-axis). Figure 1 shows the PPF for 2000 and for 2013 and marks the points on the PPFs at which the U.S. economy produced.
- Figure 2 also shows the PPF for 2013 and explains why we still import some oil and gas.
- We could easily produce the 29 million barrels per day that we consume. But to increase production from 21 million barrels per day to 29 million barrels, we would incur a higher opportunity cost of energy.
- Recall that the slope of the PPF measures opportunity cost. So the slope of the PPF at point $A$, the production point in 2013, measures the opportunity cost of producing oil and gas in 2013.
- If our production of oil and gas is efficient, we produce only the quantity at which the opportunity cost equals the cost of buying it from other countries.


Figure 1 PPF for Oil and Gas and Other Goods and Services: 2000 and 2013


Figure 2 Production, Imports, and Consumption of Oil and Gas in 2013

- The red "Trade line" in Fig. 2 shows our import possibilities. By importing 8 million barrels per day, we can consume 29 million barrels at point $B$ outside our PPF.
- The opportunity cost of producing more oil and gas in the United States is greater than the opportunity cost of importing it, so it is efficient to limit our production and import some energy.
- With further expansion of "fracking," we might eventually stop importing oil and gas and even start to export it.


## PART TWO: HOW MARKETS WORK

## CHAPTER 3: Demand and Supply

After studying this chapter, you will be able to:

- Describe a competitive market and think about a price as an opportunity cost
- Explain the influences on demand Explain the influences on supply
- Explain how demand and supply determine prices and quantities bought and sold
- Use the demand and supply model to make predictions about changes in prices and quantities
A disease that kills banana tress is jumping continents and left unchecked will bring a big drop in banana production. What will happen to the price of bananas if the disease isn't contained? The demand and supply model answers this question.
This model that you're about to study is the main tool of economics. It explains how prices are determined and how they guide the use of resources to influence What, How, and For Whom goods andservices are produced. Economics in the News at the end of the chapter answers the question about the price of bananas.


## Markets and Prices

A market has two sides: buyers and sellers. There are markets for goods such as apples and hiking boots, for services such as haircuts and tennis lessons, for factors of production such as computer programmers and earthmovers, and for other manufactured inputs suchas memory chips and auto parts. There are also markets for money such as Japanese yen and for financialsecurities such as Yahoo! stock. Only our imaginationlimits what can be traded in markets.
Some markets are physical places where buyers and sellers meet and where an auctioneer or a broker helps to determine the prices. Examples of this type of market are the New York Stock Exchange and the wholesale fish, meat, and produce markets. Some markets are groups of people spread around the world who never meet and know little about eachother but are connected through the Internet or by telephone and fax. Examples are the e-commerce markets and the currency markets.
But most markets are unorganized collections of buyers and sellers. You do most of your trading in this type of market. Markets vary in the intensity of competition that buyers and sellers face. Inthis chapter, we're going to study a competitive market - a market that has manybuyers and many sellers, so no single buyer or seller can influence the price. Producers offer items for sale only if the price is high enough to cover their opportunity cost. And consumers respond to changing opportunity cost byseeking cheaper alternatives to expensive items. We are going to study how people respond to prices and the forces that determine prices. But to pursue these tasks, we need to understand the relationship between a price and an opportunity cost.
In everyday life, the price of an object is the number of dollars that must be given up in exchange for it. Economists refer to this price as the money price.

The opportunity cost of an action is the highest-valued alternative forgone. If, when you buy a cup of coffee, the highest-valued thing you forgo is somegum, then the opportunity cost of the coffee is the quantity of gum forgone. We can calculate the quantity of gum forgone from the money prices of the coffee and the gum.
If the money price of coffee is $\$ 1$ a cup and the money price of gum is $50 ¢$ a pack, then the opportunity cost of one cup of coffee is two packs of gum. Tocalculate this opportunity cost, we divide the price of a cup of coffee by the price of a pack of gum and findthe ratio of one price to the other. The ratio of one price to another is called a relative price, and a relative price is an opportunity cost.
We can express the relative price of coffee in terms of gum or any other good. The normal way of expressing a relative price is in terms of a "basket" ofall goods and services. To calculate this relative price, we divide the money price of a good by the money price of a "basket" of all goods (called a price index). The resulting relative price tells us the opportunity cost of the good in terms of how much of the "basket" we must give up to buy it.
The demand and supply model that we are about to study determines relative prices, and the word "price" means relative price. When we predict that a price will fall, we do not mean that its money price will fall - although it might. We mean that its relative price will fall. That is, its price will fall relative to the average price of other goods and services.

## Demand

If you demand something, then you

1. Want it.
2. Can afford it.
3. Plan to buy it.

Wants are the unlimited desires or wishes that people have for goods and services. How many timeshave you thought that you would like something "if only you could afford it" or "if it weren't so expensive"? Scarcity guarantees that many - perhaps most - of our wants will never be satisfied. Demand reflects a decision about which wants to satisfy.
The quantity demanded of a good or service is the amount that consumers plan to buy during a given time period at a particular price. The quantity demanded is not necessarily the same as the quantity actually bought. Sometimes the quantity demanded exceeds the amount of goods available, so the quantity bought is less than the quantity demanded. The quantity demanded is measured as an amount per unit of time. For example, suppose that you buy one cup of coffee a day. The quantity of coffee that you demand can be expressed as 1 cup per day, 7 cups per week, or 365 cups per year. Many factors influence buying plans, and one of them is the price. We look first at the relationship between the quantity demanded of a good and its price. To study this relationship, we keep all other influences on buying plans the same and we ask: How, other things remaining the same, does the quantity demanded of a good change as its price changes? The law of demand provides the answer.

## The Law of Demand

The law of demand states:
Other things remaining the same, the higher the price of a good, the smaller is the quantity demanded; and the lower the price of a good, the greater is the quantity demanded.
Why does a higher price reduce the quantity demanded? For two reasons:

- Substitution effect
- Income effect

Substitution Effect When the price of a good rises, other things remaining the same, its relative price- its opportunity cost-rises. Although each good is unique, it has substitutesother goods that can beused in its place. As the opportunity cost of a good rises, the incentive to economize on its use and switch to a substitute becomes stronger.

Income Effect When a price rises, other things remaining the same, the price rises relative to income. Faced with a higher price and an unchanged income, people cannot afford to buy all the things they previously bought. They must decrease the quantities demanded of at least some goods and services.
Normally, the good whose price has increased will be one of the goods that people buy less $0 £$

To see the substitution effect and the income effect at work, think about the effects of a change in the price of an energy bar. Several different goods are substitutes for an energy bar. For example, an energydrink could be consumed instead of an energy bar. Suppose that an energy bar initially sells for $\$ 3$ and then its price falls to $\$ 1.50$. People now substitute energy bars for energy drinks-the substitution effect. And with a budget that now has some slack from the lower price of an energy bar, people buy even more energy bars - the income effect. The quantity of energy bars demanded increases for these two reasons.
Now suppose that an energy bar initially sells for $\$ 3$ and then the price doubles to $\$ 6$. People now buy fewer energy bars and more energy drinks-the substitution effect. And faced with a tighter budget, people buy even fewer energy bars - the income effect. The quantity of energy bars demanded decreases for these two reasons.

## Demand Curve and Demand Schedule

You are now about to study one of the two most used curves in economics: the demand curve. You are also going to encounter one of the most critical distinctions: the distinction between demand and quantity demanded.
The term demand refers to the entire relationship between the price of a good and the quantity demanded of that good. Demand is illustrated by the demand curve and the demand schedule. The termquantity demanded refers to a point on a demand curve - the quantity demanded at a particular price.

Figure 3.1 shows the demand curve for energy bars. A demand curve shows the relationship betweenthe quantity demanded of a good and its price whenall other influences on consumers' planned purchases remain the same.

The table in Fig. 3.1 is the demand schedule for energy bars. A demand schedule lists the quantities demanded at each price when all the other influences on consumers' planned purchases remain the same. For example, if the price of a bar is $50 c t$, the quantity demanded is 22 million a week. If the price is $\$ 2.50$, the quantity demanded is 5 million a week. The other rows of the table show the quantities demanded at prices of $\$ 1.00$, $\$ 1.50$, and $\$ 2.00$.
We graph the demand schedule as a demand curve with the quantity demanded on the $x$ axis and the price on the $y$-axis. The points on the demand curve labeled $A$ through $E$ correspond to the rows of the demand schedule. For example, point $A$ on the graph shows a quantity demanded of 22 million energy bars a week at a price of $50 ¢$ a bar.

Willingness and Ability to Pay Another way of looking at the demand curve is as a willingness-and-ability-to-pay curve. The willingness and ability to pay is a measure of marginal benefit. If a small quantity is available, the highest price that someone is willing and able to pay for one more unit is high. But as the quantity available increases, the marginal benefit of each additional unit falls and the highest price that someone is willing and able to pay also falls along the demand curve.
In Fig. 3.1, if only 5 million energy bars are avail- able each week, the highest price that someone is willing to pay for the 5 millionth bar is $\$ 2.50$. But if 22 million energy bars are available each week, some- one is willing to pay 50ct for the last bar bought.

## A Change in Demand

When any factor that influences buying plans changes, other than the price of the good, there is a change in demand. Figure 3.2 illustrates an increase in demand. When demand increases, the demand curve shifts right- ward and the quantity demanded at each price is greater. For example, at $\$ 2.50$ a bar, the quantity demanded on the original (left) demand curve is 5 million energy bars a week. On the new (right) demand curve, at $\$ 2.50$ a bar, the quantity demanded is 15 million bars a week.
Look closely at the numbers in the table and check that the quantity demanded at each price is greater.

A change in any influence on buying plans other than the price of the good itself results in a new demand schedule and a shift of the demand curve. A change in income changes the demand for energy bars. At a price of $\$ 1.50$ a bar, 10 million bars a week are demanded at the original income (row $C$ of the table) and 20 million bars a week are demanded at the new higher income (row $\mathrm{C}^{\prime}$ ). A rise in income increases the demand for energy bars. The demand curve shifts rightward, as shown by the shift arrow and the resulting curve.
figure 3.1 The Demand Curve


|  | Price <br> (dollars per bar) | Quantity demanded <br> (millions of bars <br> per week) |
| :---: | :---: | :---: |
| A | 0.50 | 22 |
| B | 1.00 | 15 |
| C | 1.50 | 10 |
| D | 2.00 | 7 |
| E | 2.50 | 5 |

The table shows a demand schedule for energy bars. At a price of $50 \$$ a bar, 22 million bars a week are demanded; at a price of $\$ 1.50$ a bar, 10 million bars a week are demanded. The demand curve shows the relationship between quantity demanded and price, other things remaining the same. The demand curve slopes downward: As the price falls, the quantity demanded increases.

The demand curve can be read in two ways. For a given price, the demand curve tells us the quantity that people plan to buy. For example, at a price of $\$ 1.50$ a bar, people plan to buy 10 million bars a week. For a given quantity, the demand curve tells us the maximum price that consumers are willing and able to pay for the last bar available. For example, the maximum price that consumers will pay for the 15 millionth bar is $\$ 1.00$.
figure 3.2 An Increase in Demand


| Original demand schedule Original income |  |  | New demand schedule New higher income |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | Price (dollars per bar) | Quantity demanded (millions of bars per week) |  | Price (dollars per bar) | Quantity demanded (millions of bars per week) |
| A | 0.50 | 22 | $A^{\prime}$ | 0.50 | 32 |
| B | 1.00 | 15 | $B^{\prime}$ | 1.00 | 25 |
| $c$ | 1.50 | 10 | $C^{\prime}$ | 1.50 | 20 |
| D | 2.00 | 7 | $D^{\prime}$ | 2.00 | 17 |
| E | 2.50 | 5 | $E^{\prime}$ | 2.50 | 15 |

Six main factors bring changes in demand. They are changes in:

- The prices of related goods
- Expected future prices
- Income
- Expected future income and credit
- Population
- Preferences

Prices of Related Goods The quantity of energy bars that consumers plan to buy depends in part on the prices of substitutes for energy bars. A substitute is a good that can be used in place of another good. For example, a bus ride is a substitute for a train ride; a hamburger is a substitute for a hot dog; and an energy drink is a substitute for an energy bar. If the price of a substitute for an energy bar rises, people buy less of the substitute and more energy bars. For example, if the price of an energy drink rises, people buy fewer energy drinks and more energy bars. The demand for energy bars increases.
The quantity of energy bars that people plan to buy also depends on the prices of complements with energy bars. A complement is a good that is used in conjunction with another good. Hamburgers and fries are complements, and so are energy bars and exercise. If the price of an hour at the gym falls, people buy more gym time and more energy bars.

Expected Future Prices If the expected future price of a good rises and if the good can be stored, the opportunity cost of obtaining the good for future use is lower today than it will be in the future when people expect the price to be higher. So people retimetheir purchases - they substitute over time. They buy moreof the good now before its price is expected torise (and less afterward), so the demand for the good today increases. If the expected future price of a good falls, the opportunity cost of buying the good today is high relative to what it is expected to be in the future. So people retime their purchases. They buy less of the good now before its price is expected to fall, so the demand for the good decreases today and increases in the future. Because people expect computer prices to keep falling, the current demand for computers is less (and the future demand is greater) than it otherwise would be.

Income Consumers' income influences demand. When income increases, consumers buy more of most goods; and when income decreases, consumers buy less of most goods. Although an increase in income leads to an increase in the demand for most goods, itdoes not lead to an increase in the demand for allgoods. A normal good is one for which demand increases as income increases. An inferior good is one for which demand decreases as income increases. Asincomes increase, the demand for air travel (a normalgood) increases and the demand for long-distance bustrips (an inferior good) decreases.

Expected Future Income and Credit When expected future income increases or credit becomes easier to get, demand for a good might increase now. For example, a salesperson gets the news that she willreceive a big bonus at the end of the year, so she goes into debt and buys a new car right now, rather than waiting until she receives the bonus.

Population Demand also depends on the size and the age structure of the population. The larger the population, the greater is the demand for all goods and services; the smaller the population, the smaller is the demand for all goods and services. Also, the larger the proportion of the population in a given age group, the greater is the demand for the goods and services used by that age group.

For example, during the 1990s, a decrease in the college-age population decreased the demand for college places. During those same years, the number ofAmericans aged 85 years and over increased by morethan 1 million. As a result, the demand for nursing home services increased.

Preferences Demand depends on preferences. Preferences determine the value that people place on each good and service. Preferences depend on such things as the weather, information, and fashion. Forexample, greater health and fitness awareness has shifted preferences in favor of energy bars, so the demand for energy bars has increased. Table 3.1 summarizes the influences on demand and the direction of those influences.

## A Change in the Quantity Demanded Versus a Change in Demand

Changes in the influences on buying plans bring either a change in the quantity demanded or a change in demand. Equivalently, they bring either a movement along the demand curve or a shift of the demand curve. The distinction between a change in the quantity demanded and a change in demand is the same as that between a movement along the demand curve and a shift of the demand curve. A point on the demand curve shows the quantity demanded at a given price, so a movement along the demand curve shows a change in the quantity demanded. The entire demand curve shows demand, so a shift of the demand curve shows a change in demand. Figure 3.3 illustrates these distinctions.

Movement Along the Demand Curve If the price of the good changes but no other influence on buying plans changes; we illustrate the effect as a movement along the demand curve. A fall in the price of a good increases the quantity demanded of it. In Fig. 3.3, we illustrate the effect of a fall in price as a movement down along the demand curve $D_{0}$.

A Shift of the Demand Curve If the price of a good remains constant but some other influence on buying plans changes, there is a change in demand for that good. We illustrate a change in demand as a shift of the demand curve. For example, if more people work out at the gym, consumers buy more energy bars regardless of the price of a bar. That is what a rightward shift of the demand curve shows-more energybars are demanded at each price. In Fig. 3.3, there is a change in demand and the demand curve shifts when any influence on buying plans changes, other than the price of the good.
Demand increases and the demand curve shifts rightward (to the red demand curve $\mathrm{D}_{1}$ ) if the price of asubstitute rises, the price of a complement falls, the expected future price of the good rises, income increases (for a normal good), expected future income or credit increases, or the population increases. Demand decreases and the demand curve shifts leftward (to the red demand curve $D_{2}$ ) if the price of a substitute falls, the price of a complementrises, the expected future price of the good falls, income decreases (for a normal good), expected future income or credit decreases, or the populationdecreases. (For an inferior good, the effects of changes in income are in the opposite direction to those described above.)

## TABLE 3.1 The Demand for Energy Bars

## The Law of Demand

The quantity of energy bars demanded

| Decreases if: | Increases if: |
| :--- | :--- |
| - The price of an energy | - The price of an energy |
| bar rises | bar falls |

## Changes in Demand

The demand for energy bars

| Decreases if: | Increases |
| :---: | :---: |
| - The price of a substitute falls | - The price of a substitute rises |
| - The price of a complement rises | - The price of a complement falls |
| The expected future price of an energy bar falls | The expected future price of an energy bar rises |
| - Income falls* | - Income rises* |
| - Expected future income falls or credit becomes harder to get* | - Expected future income rises or credit becomes easier to get* |
| The population decreases | The population increases |
| *An energy bar is a normal good. |  |

## Supply

If a firm supplies a good or service, the firm

1. Has the resources and technology to produce it.
2. Can profit from producing it.
3. Plans to produce it and sell it.

A supply is more than just having the resources and the technology to produce something. Resources and technology are the constraints that limit what ispossible.
Many useful things can be produced, but they are not produced unless it is profitable to do so. Supply reflects a decision about which technologically feasible items to produce.
The quantity supplied of a good or service is the amount that producers plan to sell during a given time period at a particular price. The quantity sup-plied is not necessarily the same amount as the quantity actually sold. Sometimes the quantity sup-plied is greater than the quantity demanded, so thequantity sold is less than the quantity supplied.

FIGURE 3.3 A Change in the Quantity Demanded Versus a Change in Demand


When the price of the good changes, there is a movementalong the demand curve and $a$ change in the quantity demanded. When any other influence on buying plans changes, there is a shift of the demand curve and a change in demand. An increase in demand shifts the demand curve rightward (from $D_{0}$ to $D_{1}$ ). A decrease in demand shifts the demand curve leftward (from $D_{0}$ to $D_{2}$ ).

Like the quantity demanded, the quantity sup- plied is measured as an amount per unit of time. For example, suppose that GM produces 1,000 cars a day. The quantity of cars supplied by GM can be expressed as 1,000 a day, 7,000 a week, or 365,000 a year. Without the time dimension, we cannot tell whether a particular quantity is large or small. Many factors influence selling plans, and again one of them is the price of the good. We look first at the relationship between the quantity supplied of a good and its price. Just as we did when we studied demand, to isolate the relationship between the quantities supplied of a good and its price, we keep all other influences on selling plans the same and ask: How does the quantity supplied of a good change as its price changes when other things remain the same?
The law of supply provides the answer.
The Law of Supply states:
Other things remaining the same, the higher the price of a good, the greater is the quantity supplied; and the lower the price of a good, the smaller is the quantity supplied.

Why does a higher price increase the quantity sup- plied? It is because marginal cost increases. As the quantity produced of any good increases, the marginal cost of producing the good increases. (See Chapter 2, Pg. 73 to review marginal cost.
It is never worth producing a good if the price received for the good does not at least cover the marginal cost of producing it. When the price of a good rises, other things remaining the same, producers are willing to incur a higher marginal cost, so they increase production. The higher price brings forth an increase in the quantity supplied. Let's now illustrate the law of supply with a supply curve and a supply schedule.

## Supply Curve and Supply Schedule

You are now going to study the second of the two most used curves in economics: the supply curve. You're also going to learn about the critical distinction between supply and quantity supplied.
The term supply refers to the entire relationship between the price of a good and the quantity supplied of it. Supply is illustrated by the supply curveand the supply schedule. The term quantity supplied refers to a point on a supply curve - the quantity supplied at a particular price.
Figure 3.4 shows the supply curve of energy bars. A supply curve shows the relationship between the quantity supplied of a good and its price when all otherinfluences on producers' planned sales remain the same. The supply curve is a graph of a supply schedule.
The table in Fig. 3.4 sets out the supply schedule for energy bars. A supply schedule lists the quantities supplied at each price when all the other influences onproducers' planned sales remain the same. For example, if the price of an energy bar is $50 ¢$, the quantity supplied is zero - in row $A$ of the table. If the price of an energy bar is $\$ 1.00$, the quantity supplied is 6 million energy bars a week - in row $B$. The other rows of the table show the quantities supplied at prices of $\$ 1.50, \$ 2.00$, and $\$ 2.50$.
To make a supply curve, we graph the quantity supplied on the $x$-axis and the price on the $y$-axis. The points on the supply curve labeled $A$ through $E$ correspond to the rows of the supply schedule. For example, point $A$ on the graph shows a quantity supplied of zero at a price of $50 \zeta$ an energy bar. Point $E$ shows a quantity supplied of 15 million bars at $\$ 2.50$ an energybar.

Minimum Supply Price The supply curve can be interpreted as a minimum-supply-price curve - a curve that shows the lowest price at which someone is willing to sell. This lowest price is the marginal cost. If a small quantity is produced, the lowest price at which someone is willing to sell one more unit is low.
But as the quantity produced increases, the marginal cost of each additional unit rises, so the lowest price at which someone is willing to sell an additional unit rises along the supply curve.
In Fig. 3.4, if 15 million bars are produced each week, the lowest price at which someone is willing to sell the 15 millionth bar is $\$ 2.50$. But if 10 million bars are produced each week, someone is willing to accept $\$ 1.50$ for the last bar produced.
figure 3.4 The Supply Curve


The table shows the supply schedule of energy bars. For example, at a price of $\$ 1.00,6$ million bars a week are supplied; at a price of $\$ 2.50,15$ million bars a week are supplied. The supply curve shows the relationship between the quantity supplied and the price, other things remaining the same. The supply curve slopes upward: As the price of a good increases, the quantily supplied increases.

A supply curve can be read in two ways. For a given price, the supply curve tells us the quantity that producers plan to sell at that price. For example, at a price of $\$ 1.50$ a bar, producers are planning to sell 10 million bars a week. For a given quantity, the supply curve tells us the minimum price at which producers are willing to sell one more bar. For example, if 15 million bars are produced each week, the lowest price at which a producer is willing to sell the 15 millionth bar is $\$ 2.50$.

## A Change in Supply

When any factor that influences selling plans other than the price of the good changes, there is a changein supply. Six main factors bring changes in supply. They are changes in

- The prices of factors of production
- The prices of related goods produced
- Expected future prices
- The number of suppliers
- Technology
- The state of nature

Prices of Factors of Production The prices of the factors of production used to produce a good influence its supply. To see this influence, think about the sup- ply curve as a minimum-supply-price curve. If the price of a factor of production rises, the lowest price that a producer is willing to accept for that good rises, so supply decreases. For example, during 2008, as the price of jet fuel increased, the supply of air travel decreased. Similarly, a rise in the minimumwage decreases the supply of hamburgers.

Prices of Related Goods Produced The prices of related goods that firms produce influence supply. Forexample, if the price of an energy drink rises, firmsswitch production from bars to drinks. The supply ofenergy bars decreases. Energy bars and energy drinksare substitutes in production- good $s$ that can be produced by using the same resources. If the price of beef rises, the supply of cowhide increases. Beef and cow- hide are complements in productiongoods that must be produced together.

Expected Future Prices If the expected future price of a good rises, the return from selling the good in the future increases and is higher than it is today. So sup- ply decreases today and increases in the future.

The Number of Suppliers The larger the number of firms that produce a good, the greater is the supply of the good. As new firms enter an industry, the supply in that industry increases. As firms leave an industry, the supply in that industry decreases.

Technology The term "technology" is used broadly to mean the way that factors of production are used to produce a good. A technology change occurs when a new method is discovered that lowers the cost of producing a good. For example, new methods used in the factories that produce computer chips have lowered the cost and increased the supply of chips.

The State of Nature The state of nature includes all the natural forces that influence production. It includes the state of the weather and, more broadly the natural environment. Good weather can increase the supply of many agricultural products and bad weather can decrease their supply. Extreme natural events such as earthquakes, tornadoes, and hurricanes can also influence supply.

FIGURE 3.5 An Increase in Supply


| Original supply schedule Old technology |  |  | New supply schedule New technology |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | Price \|dollars per bar) | Quantity supplied (millions of bars per week) |  | Price (dollars per bar) | Quantity supplied (millions of bars per week) |
| A | 0.50 | 0 | $A^{\prime}$ | 0.50 | 7 |
| B | 1.00 | 6 | $B^{\prime}$ | 1.00 | 15 |
| C | 1.50 | 10 | $C^{\prime}$ | 1.50 | 20 |
| D | 2.00 | 13 | $D^{\prime}$ | 2.00 | 25 |
| E | 2.50 | 15 | E' | 2.50 | 27 |

## A Change in the Quantity Supplied Versus a Change in Supply

A change in any influence on selling plans other than theprice of the good itself results in a new supply schedule and a shift of the supply curve. For example, a new, cost-saving technology for producing energy bars changes the supply of energy bars. At a price of $\$ 1.50$ a bar, 10 million bars a week are supplied when producers use the old technology (row C of the table) and 20 million energy bars a week aresupplied when producers use the new technology (row $\mathrm{C}^{\prime}$ ). An advance in technology increases the supply of energybars. The supply curve shifts rightward, as shown by the shift arrow and the resulting curve.

## TABLE 3.2 The Supply of Energy Bars

## The Law of Supply

The quantity of energy bars supplied
Decreases if: Increases if:

- The price of an energy - The price of an energy bar falls bar rises


## Changes in Supply

The supply of energy bars

Decreases if:

- The price of a factor of production used to produce energy bars rises
- The price of a substitute in production rises
- The price of a complement in production falls
- The expected future price of an energy bar rises
- The number of suppliers of bars decreases
- A technology change decreases energy bar production
- A natural event decreases energy bar production

Increases if:

- The price of a factor of production used to produce energy bars falls
- The price of a substitute in production falls
- The price of a complement in production rises
- The expected future price of an energy bar falls
- The number of suppliers of bars increases
- A technology change increases energy bar production
- A natural event increases energy bar production

Figure 3.6 illustrates and summarizes these distinctions. If the price of the good changes and other things remain the same, there is a change in the quantity supplied of that good. If the price of the good falls, the quantity supplied decreases and there is a movement down along the supply curve $S_{0}$. If the price of the good rises, the quantity supplied increases and there is a movement up along the supply curve $S_{0}$. When any other influence on selling plans changes, the supply curve shifts and there is a change insupply. If supply increases, the supply curve shifts rightward to $\mathrm{S}_{1}$ supply decreases, the supply curve shifts leftward to $\mathrm{S}_{2}$.

FIGURE 3.6 A Change in the Quantity Supplied Versus a Change in Supply


When the price of the good changes, there is a movement along the supply curve and a change in the quantity supplied, shown by the blue arrows on supply curve $S_{0}$. When any other influence on selling plans changes, there is a shiff of the supply curve and a change in supply. An increase in supply shifts the supply curve rightward (from $S_{0}$ to $S_{1}$ ), and a decrease in supply shifts the supply curve leffward (from $S_{0}$ to $S_{2}$ ).

## Market Equilibrium

We have seen that when the price of a good rises, thequantity demanded decreases and the quantity sup-plied increases. We are now going to see how the price adjusts to coordinate buying plans and selling plans and achieve equilibrium in the market.
An equilibrium is a situation in which opposing forces balance each other. Equilibrium in a market occurs when the price balances buying plans and selling plans. The equilibrium price is the price at which the quantity demanded equals the quantity supplied. The equilibrium quantity is the quantity bought and sold at the equilibrium price. A market moves toward its equilibrium because

- Price regulates buying and selling plans.
- Price adjusts when plans don't match.


## Price as a Regulator

The price of a good regulates the quantities demanded and supplied. If the price is too high, the quantity supplied exceeds the quantity demanded. If the price is too low, the quantity demanded exceeds the quantity supplied. There is one price at which the quantity demanded equals the quantity supplied. Let's work out what that price is. Figure 3.7 shows the market for energy bars. The table shows the demand schedule (from Fig. 3.1) and the supply schedule (from Fig. 3.4). If the price is $50 ¢$ a bar, the quantity demanded is 22 million bars a week but no bars are supplied. There is a shortage of 22 million bars a week. The final column of the table shows this shortage. At a price of $\$ 1.00$ a bar, there is still a shortage but only of 9 million bars a week.
If the price is $\$ 2.50$ a bar, the quantity supplied is 15 million bars a week but the quantity demanded is only 5 million. There is a surplus of 10 million bars a week.
The one price at which there is neither a shortage nor a surplus is $\$ 1.50$ a bar. At that price, the quantity demanded equals the quantity supplied: 10 mil-lion bars a week. The equilibrium price is $\$ 1.50$ a bar, and the equilibrium quantity is 10 million bars a week.
Figure 3.7 shows that the demand curve and the supply curve intersect at the equilibrium price of $\$ 1.50$ a bar. At each price above $\$ 1.50$ a bar, there is a surplus of bars. For example, at $\$ 2.00$ a bar. The surplus is 6 million bars a week, as shown by the blue arrow. At each price below $\$ 1.50$ a bar, there is a shortage of bars. For example, at $\$ 1.00$ a bar, the shortage is 9 million bars a week, as shown by the red arrow.

## Price Adjustments

You've seen that if the price is below equilibrium, there is a shortage; and if the price is above equilibrium, there is a surplus. But can we count on the price to change and eliminate a shortage or a surplus? We can, because such price changes are beneficial to both buyers and sellers. Let's see why the price changes when there is a shortage or a surplus.

Shortage Forces the Prices Up Suppose the price of an energy bar is \$1. Consumers plan to buy 15 million bars a week, and producers plan to sell 6 million bars a week. Consumers can't force producers to sell more than they plan, so the quantity that is actually offered for sale is 6 million bars a week. In this situation, powerful forces operate to increase the price and move it toward the equilibrium price. Some producers, noticing lines of unsatisfied consumers, raise the price.
Some producers increase their output. As producers push the price up, the price rises toward its equilibrium. The rising price reduces the shortage because it decreases the quantity demanded and increases the quantity supplied. When the price has increased to the point at which there is no longer a shortage, the forces moving the price stop operating and the price comes to rest at its equilibrium.

FIGURE 3.7 Equilibrium


| Price <br> (dollars <br> per bar) | Quantity <br> demanded | Quantity <br> supplied | Shortage (-) <br> or surplus ( + ) |
| :---: | :---: | :---: | :---: |
|  |  |  |  |

The table lists the quantity demanded and the quantity supplied as well as the shoriage or surplus of bars at each price. If the price is $\$ 1.00$ a bar, 15 million bars a week are demanded and 6 million bars are supplied. There is a shortage of 9 million bars a week, and the price rises.

If the price is $\$ 2.00 \mathrm{a}$ bar, 7 million bars a week are demanded and 13 million bars are supplied. There is a surplus of 6 million bars a week, and the price falls.

If the price is $\$ 1.50$ a bar, 10 million bars a week are demanded and 10 million bars are supplied. There is neither a shortage nor a surplus, and the price does not change. The price at which the quantity demanded equals the quantity supplied is the equilibrium price, and 10 million bars a week is the equilibrium quantity.

A Surplus Forces the Price Down Suppose the price of a bar is $\$ 2$. Producers plan to sell 13 million bars a week, and consumers plan to buy 7 million bars a week. Producers cannot force consumers to buy more than they plan, so the quantity that is actually bought is 7 million bars a week. In this situation, powerful forces operate to lower the price and move it toward the equilibrium price. Some producers, unable to sell the quantities of energy bars they planned to sell, cut their prices. In addition, some producers scale back production. As producers cut the price, the price falls toward its equilibrium. The falling price decreases the surplus because it increases the quantity demanded and decreases the quantity supplied. When the price has fallen to the point at which there is no longer a surplus, the forces moving the price stop operating and the price comes to rest at its equilibrium.

The Best Deal Available for Buyers and Sellers When the price is below equilibrium, it is forced upward. Why don't buyers resist the increase andrefuse to buy at the higher price? The answer is because they value the good more highly than its current price and they can't satisfy their demand at the current price. In some markets - for example, the markets that operate on eBay - the buyers might even be the ones who force the price up by offering to pay a higher price.
When the price is above equilibrium, it is bid downward. Why don't sellers resist this decrease and refuse to sell at the lower price? The answer is because their minimum supply price is below the current price and they cannot sell all they would like to at thecurrent price. Sellers willingly lower the price to gainmarket share.
At the price at which the quantity demanded and the quantity supplied are equal, neither buyers nor sellers can do business at a better price. Buyers pay the highest price they are willing to pay for the last unit bought, and sellers receive the lowest price at which they are willing to supply the last unit sold.
When people freely make offers to buy and sell and when demanders try to buy at the lowest possible price and suppliers try to sell at the highest possible price, the price at which trade takes place is the equilibrium price - the price at which the quantity demanded equals the quantity supplied. The price coordinates the plans of buyers and sellers, and no one has an incentive to change it.

## Predicting Changes in Price and Quantity

The demand and supply model that we have just studied provides us with a powerful way of analyzinginfluences on prices and the quantities bought and sold. According to the model, a change in price stems from a change in demand, a change in supply, or achange in both demand and supply. Let's look first atthe effects of a change in demand.

## An increase in demand

If more people join health clubs, the demand for energy bars increases. The table in Fig. 3.8 (on page 106) shows theoriginal and new demand schedules for energy bars as well as the supply schedule of energy bars.
The increase in demand creates a shortage at the original price, and to eliminate the shortage the price must rise.

Figure 3.8 shows what happens. The figure shows the original demand for and supply of energy bars. The original equilibrium price is $\$ 1.50$ an energy bar, and the equilibrium quantity is 10 million energy bars a week. When demand increases, the demand curve shifts rightward. The equilibrium price rises to $\$ 2.50$ an energy bar, and the quantity supplied increases to 15 million energy bars a week, as high- lighted in the figure. There is an increase in the quantity supplied but no change in supply-a movement along, but no shift of, the supply curve.

## A Decrease in Demand

We can reverse this change in demand. Start at a price of $\$ 2.50$ a bar with 15 million energy bars a week being bought and sold, and then work out what hap-pens if demand decreases to its original level. Such a decrease in demand might arise if people switch to energy drinks (a substitute for energy bars). The decrease in demand shifts the demand curve leftward. The equilibrium price falls to $\$ 1.50$ a bar, the quantity supplied decreases, and the equilibrium quantity decreases to 10 million bars a week. We can now make our first two predictions:

1. When demand increases, the price rises and the quantity increases.
2. When demand decreases, the price falls and the quantity decreases.

## An Increase in Supply

When Nestle (the producer of PowerBar) and otherenergy bar producers switch to a new cost-saving technology, the supply of energy bars increases. Figure 3.9 shows the new supply schedule (the same one that was shown in Fig. 3.5). What are the new equilibrium price and quantity? The price falls to $\$ 1.00$ a bar, and the quantity increases to 15 million bars a week. You can see why by looking at the quantities demanded and supplied at the old price of $\$ 1.50$ a bar. The new quantity supplied at that price is 20 million bars a week, and there is a surplus. The price falls. Only when the price is $\$ 1.00$ a bar does the quantity supplied equal the quantity demanded.
Figure 3.9 illustrates the effect of an increase in supply. It shows the demand curve for energy bars and the original and new supply curves. The initial equilibrium price is $\$ 1.50$ a bar, and the equilibrium quantity is 10 million bars a week. When sup- ply increases, the supply curve shifts rightward. The equilibrium price falls to $\$ 1.00$ a bar, and the quantity demanded increases to 15 million bars a week, highlighted in the figure. There is an increase in the quantity demanded but no change in demand - a movement along, but no shift of, the demand curve.

## Decrease in Supply

Start out at a price of $\$ 1.00$ a bar with 15 million bars a week being bought and sold. Then suppose that the cost of labor or raw materials rises and the supply of energy bars decreases. The decrease in sup- ply shifts the supply curve leftward. The equilibrium price rises to $\$ 1.50$ a bar, the quantity demanded decreases, and the equilibrium quantity decreases to 10 million bars a week.
We can now make two more predictions:

1. When supply increases, the price falls and the quantity increases.
2. When supply decreases, the price rises and the quantity decreases.

HGURE 3.9 The Effects of a Change in Supply


| Price <br> (dollars <br> per bar) | Quantity <br> demanded <br> (millions of bars <br> per week) | Quantity supplied <br> (millions of bars per week) |  |
| :---: | :---: | :---: | :---: |
| 0.50 | 22 | Original | New |
| 1.00 | 15 | 0 | 7 |
| 1.50 | 10 | 6 | 15 |
| 2.00 | 7 | 10 | 20 |
| 2.50 | 5 | 15 | 25 |

Initially, the supply of energy bars is shown by the blue supply curve. The equilibrium price is $\$ 1.50$ a bar, and the equilibrium quantity is 10 million bars a week. When the new cost-saving technology is adopted, the supply of energy bars increases and the supply curve shifts rightward to become the red curve.

At $\$ 1.50$ a bar, there is now a surplus of 10 million bars a week. The price of an energy bar falls to a new equilibrium of $\$ 1.00$ a bar. As the price falls to $\$ 1.00$, the quantity demanded increases - shown by the blue arrow on the demand curve--to the new equilibrium quantity of 15 million bars a week. Following an increase in supply, the quantity demanded increases but demand does not change-the demand curve does not shiff.

You've now seen what happens to the price and the quantity when either demand or supply changes while the other one remains unchanged. In real markets, both demand and supply can change together. When this happens, to predict the changes in price and quantity, we must combine the effects that you've just seen. That is your final task in this chapter.

## Changes in Both Demand and Supply

You now know how a change in demand or a change in supply changes the equilibrium price and quantity. But sometimes, events occur that change both demand and supply. When both demand and supply change, we find the resulting change in the equilibrium price and equilibrium quantity by combining the separate cases you've just studied. Four cases need to be considered. Both demand and supply might increase or decrease, and demand or supply might increase and the other decrease.

Both Demand and Supply Change in the Same Direction When demand and supply change in the same direction, the equilibrium quantity changes in that same direction, but to predict whether the price rises or falls, we need to know the magnitudes of the changes in demand and supply. If demand increases by more than supply increases, the price rises. But if supply increases by more than demand increases, the price falls.
Figure 3.10(a) shows the case when both demand and supply increase and by the same amount. The equilibrium quantity increases. But because the increase in demand equals the increase in supply, neither a shortage nor a surplus arises so the price doesn't change. A bigger increase in demand would have created a shortage and a rise in the price; a bigger increase in supply would have created a surplus and a fall in the price. Figure 3.10(b) shows the case when both demand and supply decrease by the same amount. Here the equilibrium quantity decreases and again the price might either rise or fall.

Both Demand and Supply Change in Opposite Directions When demand and supply change in opposite directions, we can predict how the price changes, but we need to know the magnitudes of the changes in demand and supply to say whether the equilibrium quantity increases or decreases.
If demand changes by more than supply, the equilibrium quantity changes in the same direction as the change in demand. But if supply changes by more than demand, the equilibrium quantity changes in the same direction as the change in supply.

FIGURE 3.10 The Effects of Changes in Both Demand and Supply in the Same Direction


(a) Increase in both demand and supply

An increase in demand shifts the demand curve rightward to become the new demand curve and an increase in supply shifts the supply curve righ1ward to become the new supply curve. The price might rise or fall, but the quantity increases.
(b) Decrease in both demand and supply

A decrease in demand shifts the demand curve leftward to become the new demand curve and a decrease in supply shifts the supply curve leftward to become the new supply curve. The price might rise or fall, but the quantity decreases.
figure 3.11 The Effects of Changes in Both Demand and Supply in Opposite Directions

(a) Decrease in demand; increase in supply

(b) Increase in demand; decrease in supply
(a) Decrease in demand; increase in supply A decrease in demand shifts the demand curve leftward to become the new demand curve and an increase in
supply shifts the supply curve rightward to become the new supply curve. The price falls, but the quantity might increase or decrease.
(b) Increase in demand; decrease in supply

An increase in demand shifts the demand curve rightward to become the new demand curve and a decrease in supply shifts the supply curve leftward to become the new supply curve. The price rises, but the quantity might increase or decrease.

## ECONOMIC ANALYSIS

- In the market for bananas, a decrease in world production would decrease supply.
- A decrease in the supply of bananas would raise their price, decrease the equilibrium quantity, and decrease the quantity of bananas demanded.
- We can see the likely price increase by looking at previous events in the banana market.
- Figure 1 shows the price of bananas since 2004. You can see that there was a big temporary jump in the price in 2008.
- That jump in price was not caused by a decrease in banana production because as Fig. 2 shows, banana production has increased every year since 2004 except for 2012.
- What happened in 2008 ? The answer is a spike in the price of oil.
- Transporting bananas from plantations in Central and South America to your neighborhood grocery store uses a lot of fuel. So when the cost of fuel increased in 2008, the cost of delivering bananas increased, and the U.S. consumer price of bananas increased.
- A decrease in supply caused by the TR4 disease would have a similar effect on the banana market to what happened in 2008.
- Figure 3 illustrates this effect. The supply of bananas decreases from $S_{N}$ (normal) to $S_{D}$ (disease), the price rises, the equilibrium quantity decreases, and the quantity of bananas demanded decreases.


Figure 1 The Price of Bananas: 2004-2014


Figure 2 Banana Production: 2004-2012


Figure 3 The Market for Bananas


Figure 1 The Price of Bananas: 2004-2014


Figure 2 Banana Production: 2004-2012


Figure 3 The Market for Bananas

## MATHEMATICAL NOTE

## Demand, Supply, and Equilibrium

## Demesnd Curve

The law of demand says that as the price of a good or service falls, the quantity demanded of that good or service increases. We can illustrate the law of demand by drawing a graph of the demand curve or writing down an equation. When the demand curve is a straight line, the following equation describes it:

$$
P=a-b Q_{D}
$$

where $P$ is the price and $Q_{D}$ is the quantity demanded. The $a$ and $b$ are positive constants.

The demand equation tells us three things:

1. The price at which no one is willing to buy the $\operatorname{good}\left(Q_{D}\right.$ is zero). That is, if the price is $a$, then the quantity demanded is zero. You can see the price $a$ in Fig. 1. It is the price at which the demand curve hits the $y$-axis-what we call the demand curve's " $y$-intercept."
2. As the price falls, the quantity demanded increases. If $Q_{D}$ is a positive number, then the price $P$ must be less than $a$. As $Q_{D}$ gets larger, the price $P$ becomes smaller. That is, as the quantity increases, the maximum price that buyers are willing to pay for the last unit of the good falls.
3. The constant $b$ tells us how fast the maximum price that someone is willing to pay for the good falls as the quantity increases. That is, the constant $b$ tells us about the steepness of the demand curve. The equation tells us that the slope of the demand curve is $-b$.

## Seppaly Curve

The law of supply says that as the price of a good or service rises, the quantity supplied of that good or service increases. We can illustrate the law of supply by drawing a graph of the supply curve or writing down an equation. When the supply curve is a straight line, the following equation describes it:

$$
P=c+d Q_{s}
$$

where $P$ is the price and $Q_{S}$ is the quantity supplied. The $c$ and $d$ are positive constants.

The supply equation tells us three things:

1. The price at which sellers are not willing to supply the good ( $Q_{S}$ is zero). That is, if the price is $c$, then no one is willing to sell the good. You can see the price $c$ in Fig. 2. It is the price at which the supply curve hits the $y$-axis-what we call the supply curve's " $y$-intercept."
2. As the price rises, the quantity supplied increases. If $Q_{S}$ is a positive number, then the price $P$ must be greater than $c$. As $Q_{S}$ increases, the price $P$ becomes larger. That is, as the quantity increases, the minimum price that sellers are willing to accept for the last unit rises.
3. The constant $d$ tells us how fast the minimum price at which someone is willing to sell the good rises as the quantity increases. That is, the constant $d$ tells us about the steepness of the supply curve. The equation tells us that the slope of the supply curve is $d$.


Figure 2 Supply Curve

## Market Equilibrium

Demand and supply determine market equilibrium. Figure 3 shows the equilibrium price ( $P^{*}$ ) and equilibrium quantity $\left(Q^{*}\right)$ at the intersection of the demand curve and the supply curve.

We can use the equations to find the equilibrium price and equilibrium quantity. The price of a good adjusts until the quantity demanded $Q_{D}$ equals the quantity supplied $Q_{S}$. So at the equilibrium price ( $P^{*}$ ) and equilibrium quantity ( $Q^{*}$ ),

$$
Q_{D}=Q_{S}=Q^{*}
$$

To find the equilibrium price and equilibrium quantity, substitute $Q^{*}$ for $Q_{D}$ in the demand equation and $Q^{*}$ for $Q_{S}$ in the supply equation. Then the price is the equilibrium price $\left(P^{*}\right)$, which gives

$$
\begin{aligned}
& P^{*}=a-b Q^{*} \\
& P^{*}=c+d Q^{*} .
\end{aligned}
$$

Notice that

$$
a-b Q^{*}=c+d Q^{*}
$$

Now solve for $Q^{*}$ :

$$
\begin{aligned}
a-c & =b Q^{*}+d Q^{*} \\
a-c & =(b+d) Q^{*} \\
Q^{*} & =\frac{a-c}{b+d}
\end{aligned}
$$

To find the equilibrium price $P^{*}$, substitute for $Q^{*}$ in either the demand equation or the supply equation.


Figure 3 Market Equilibrium

Using the demand equation, we have

$$
\begin{aligned}
& P^{*}=a-b\left(\frac{a-c}{b+d}\right) \\
& P^{*}=\frac{a(b+d)-b(a-c)}{b+d} \\
& P^{*}=\frac{a d+b c}{b+d} .
\end{aligned}
$$

Alternatively, using the supply equation, we have

$$
\begin{aligned}
& P^{*}=c+d\left(\frac{a-c}{b+d}\right) \\
& P^{*}=\frac{c(b+d)+d(a-c)}{b+d} \\
& P^{*}=\frac{a d+b c}{b+d} .
\end{aligned}
$$

## An Example

The demand for ice-cream cones is

$$
P=800-2 Q_{D} .
$$

The supply of ice-cream cones is

$$
P=200+1 Q_{s} .
$$

The price of a cone is expressed in cents, and the quantities are expressed in cones per day.

To find the equilibrium price $\left(P^{*}\right)$ and equilibrium quantity $\left(Q^{*}\right)$, substitute $Q^{*}$ for $Q_{D}$ and $Q_{S}$ and $P^{*}$ for $P$. That is,

$$
\begin{aligned}
& P^{*}=800-2 Q^{*} \\
& P^{*}=200+1 Q^{*}
\end{aligned}
$$

Now solve for $Q^{*}$ :

$$
\begin{aligned}
800-2 Q^{*} & =200+1 Q^{*} \\
600 & =3 Q^{*} \\
Q^{*} & =200 .
\end{aligned}
$$

And

$$
\begin{aligned}
P^{*} & =800-2(200) \\
& =400 .
\end{aligned}
$$

The equilibrium price is $\$ 4$ a cone, and the equilibrium quantity is 200 cones per day.

## CHAPTER 4: Elasticity

After studying this chapter, you will be able to:

- Define, calculate, and explain the factors that influence the price elasticity of demand
- Define, calculate, and explain the factors that influence the income elasticity of demand and the cross elasticity of demand
- Define, calculate, and explain the factors that influence the elasticity of supply

In 2013, coffee production soared and the price tumbled. Despite the increased production, coffee producers saw their revenue fall. How does the quantity of coffee produced influence the price of coffee and the revenue of coffee producers?
To answer this and similar questions, we use the neat tool that you study in this chapter: elasticity.
At the end of the chapter, in Economics in the News, we use the concept of elasticity to answer the question about the market for coffee. But we begin by explaining elasticity in another familiar setting: the market for pizza.

## Price Elasticity of Demand

Youknow that when supply decreases, the equilibrium price rises and the equilibrium quantity decreases. But does the price rise by a large amount and the quantity decrease by a little? Or does the price barely rise and the quantity decrease by a large amount?
The answer depends on the responsiveness of the quantity demanded of a good to a change in its price. If the quantity demanded is not very responsive to a change in the price, the price rises a lot and the equilibrium quantity doesn't change much. If the quantity demanded is very responsive to a change in the price, the price barely rises and the equilibrium quantity changes a lot.
You might think about the responsiveness of the quantity demanded of a good to a change in its price in terms of the slope of the demand curve. If the demand curve is steep, the quantity demanded of the good isn't very responsive to a change in the price. If the demand curve is almost flat, the quantity demanded is very responsive to a change in the price.
But the slope of a demand curve depends on the units in which we measure the price and the quantity - we can make the curve steep or almost flat justby changing the units in which we measure the price and the quantity. Also we often want to compare thedemand for different goods and services and quantities of these goods are measured in unrelated units. For example, a pizza producer might want to compare the demand for pizza with the demand for softdrinks. Which quantity demanded is more responsive to a price change? This question can't be answered by comparing the slopes of two demandcurves. The units of measurement of pizza and soft drinks are unrelated. But the question can be answered with a measure of responsiveness that is independent of units of measurement. Elasticity is such a measure.

The price elasticity of demand is a units-free measure of the responsiveness of the quantity demanded of agood to a change in its price when all other influences on buying plans remain the same.

## Calculating Price Elasticity of Demand

We calculate the price elasticity of demand by using the formula:

$$
\begin{gathered}
\text { Percentage change } \\
\text { Price elasticity of } \\
\text { demand }
\end{gathered}=\frac{\text { in quantity damanded }}{\text { Percentage change in price }}
$$

To calculate the price elasticity of demand for pizza, we need to know the quantity demanded of pizza at two different prices, when all other influences on buying plans remain the same. Figure 4.1 zooms in on a section of the demand curve for pizza and shows how the quantity demanded responds to a small change in price.
Initially, the price is $\$ 20.50$ a pizza and 9 pizzas an hour are demanded-the original point. The price then falls to $\$ 19.50$ a pizza, and the quantity demanded increases to 11 pizzas an hour-the newpoint. When the price falls by $\$ 1$ a pizza, the quantity demanded increases by 2 pizzas an hour.
To calculate the price elasticity of demand, we express the change in price as a percentage of the average price and the change in the quantity demanded as a percentage of the average quantity. By using the average price and average quantity, we calculate the elasticity at a point on the demand curve midway between theoriginal point and the new point.
The original price is $\$ 20.50$ and the new price is
$\$ 19.50$, so the price change is $\$ 1$ and the average price is $\$ 20$ a pizza. Call the percentage change in the price $\% \Delta P$, then

$$
\% \Delta P=\Delta P / P_{a v e} \times 100=(\$ 1 / \$ 20) \times 100=5 \%
$$

The original quantity demanded is 9 pizzas and the new quantity demanded is 11 pizzas, so the quantity change is 2 pizzas and the average quantity demanded is 10 pizzas. Call the percentage change in the quantity demanded of $Q$, then

$$
\% \Delta Q=\Delta Q / Q_{a v e} \times 100=(2 / 10) \times 100=20 \% .
$$

The price elasticity of demand equals the percentage change in the quantity demanded (20 percent) divided by the percentage change in price ( 5 percent) and is 4 . That is,

$$
\begin{aligned}
\text { Price elasticity of demand } & =\frac{\% \Delta Q}{\% \Delta P} \\
& =\frac{20 \%}{5 \%}=4
\end{aligned}
$$

Average Price and Quantity Notice that we use the average price and average quantity. We do this because it gives the most precise measurement of elasticity-at the midpoint between the original price and the new price. If the price falls from $\$ 20.50$ to $\$ 19.50$, the $\$ 1$ price change is 4.9 percent of $\$ 20.50$. The change in quantity of 2 pizzas is 22.2 percentof 9 pizzas, the original quantity. So if we use these numbers, the price elasticity of demand is 22.2 divided by 4.9 , which equals 4.5 .

FIGURE 4.1 Calculating the Elasticity of Demand


The elasticity of demand is calculated by using the formula:*

$$
\begin{aligned}
\text { Price elasticity of demand } & =\frac{\begin{array}{c}
\text { Percentage change } \\
\text { in quantity demanded }
\end{array}}{\text { Percentage change in price }} \\
& =\frac{\% \Delta Q}{\% \Delta P} \\
& =\frac{\Delta Q / Q_{\text {ove }}}{\Delta P / P_{\text {ove }}} \\
& =\frac{2 / 10}{1 / 20}=4
\end{aligned}
$$

This calculation measures the elasticity at an average price of $\$ 20$ a pizza and an average quantity of 10 pizzas an hour.

[^0]But if the price rises from $\$ 19.50$ to $\$ 20.50$, the $\$ 1$ price change is 5.1 percent of $\$ 19.50$. The change in quantity of 2 pizzas is 18.2 percent of 11 pizzas, the original quantity. So if we use these numbers, the price elasticity of demand is 18.2 divided by 5.1 , which equals 3.6 . By using percentages of the average price and aver-age quantity, we get the same value for the elasticity regardless of whether the price falls from $\$ 20.50$ to $\$ 19.50$ or rises from $\$ 19.50$ to $\$ 20.50$.

Percentages and Proportions Elasticity is the ratio of two percentage changes, so when we divide one percentage change by another, the 100 s cancel. A percent- age change is a proportionate change multiplied by 100 . The proportionate change in price is $\Delta \mathrm{P} / \mathrm{P}_{\text {ave }}$ and the proportionate change in quantity demanded is

$$
\Delta Q / Q_{\text {ave }} \text {. So if we divide } \Delta Q / Q_{\text {ave by }} \Delta P / P_{\text {ave }}
$$

we get the same answer as we get by using percentage changes.
A Units-free Measure Now that you've calculated a price elasticity of demand, you can see why it is aunits-free measure. Elasticity is a units-free measure because the percentage change in each variable is independent of the units in which the variable is measured. The ratio of the two percentages is a number without units.

Minus Sign and Elasticity When the price of a good rises, the quantity demanded decreases. Because a positive change in price brings a negative change in the quantity demanded, the price elasticity of demand is a negative number. But it is the magnitude, or absolute value, of the price elasticity of demand thattells us how responsive the quantity demanded is. Soto compare price elasticities of demand, we use the magnitude of the elasticity and ignore the minus sign.

## Inelastic and Elastic Demand

If the quantity demanded remains constant when the price changes, then the price elasticity of demand is zero and the good is said to have a perfectly inelastic • demand. One good that has a very low price elasticity of demand (perhaps zero over some price range) is insulin. Insulin is of such importance to some diabetics that if the price rises or falls, they do not change the quantity they buy.
If the percentage change in the quantity demanded equals the percentage change in the price, then the price elasticity equals 1 and the good is said to have aunit elastic demand. Between perfectly inelastic demand and unit elastic demand is a general case in which the percentage change in the quantity demanded is less than the percentage change in the price. In this case, the price elasticity of demand is between zero and 1 and the good is said to have an inelastic demand. Food and shelter are examples of goods with inelastic demand.
If the quantity demanded changes by an infinitely large percentage in response to a tiny price change, then the price elasticity of demand is infinity and thegood is said to have a perfectly elastic demand. An example of a good that has a very high elasticity of demand (almost infinite) is a soft drink from two campus machines located side by side. If the twomachines offer the
same soft drinks for the same price, some people buy from one machine and some from the other. But if one machine's price is higher than the other's, by even a small amount, no one buys from the machine with the higher price. Drinks from the two machines are perfect substitutes. Thedemand for a good that has a perfect substitute is perfectly elastic.
Between unit elastic demand and perfectly elastic demand is another general case in which the percentage change in the quantity demanded exceeds the percentage change in price. In this case, the price elasticity of demand is greater than 1 and the good is said to have an elastic demand. Automobiles and furniture are examples of goods that have elastic demand. Figure 4.2 on page 124 shows three demand curves that cover the entire range of possible elasticities of demand that you've just reviewed. In Fig. 4.2(a), the quantity demanded is constant regardless of the price, so thisdemand is perfectly inelastic. In Fig. 4.2(b), the percentage change in the quantity demanded equals the percentage change in price, so this demand is unit elastic. In Fig. 4.2(c), the price is constant regardless of the quantity demanded, so this figure illustrates a perfectly elastic demand.
You now know the distinction between elastic and inelastic demand. But what determines whether the demand for a good is elastic or inelastic?

## The Factors That Determine the Elasticity of Demand

The elasticity of demand for a good depends on

- The closeness of substitutes
- The proportion of income spent on the goodThe time elapsed since the price change

Closeness of Substitutes The closer the substitutes for a good, the more elastic is the demand for it. Oil as fuel or raw material for chemicals has no close substitutes so the demand for oil is inelastic. Plastics are close substitutes for metals, so the demand for metals is elastic.
The degree of substitutability depends on how narrowly (or broadly) we define a good. For example, a smartphone has no close substitutes, but an Apple iPhone is a close substitute for a Samsung Galaxy. So the elasticity of demand for smartphones is lower than the elasticity of demand for an iPhone or a Galaxy.
In everyday language we call goods such as food and shelter necessities and goods such as exoticvacations luxuries. A necessity has poor substitutes, so it generally has an inelastic demand. A luxury usuallyhas many substitutes, one of which is not buying it. So a luxury generally has an elastic demand.

Proportion of Income Spent on the Good Other things remaining the same, the greater the proportion of income spent on a good, the more elastic (or less inelastic) is the demand for it. Think about your own elasticity of demand for chewing gum and housing. If the price of gum rises, you consume almost as much as before. Yourdemand for gum is inelastic. If apartment rents rise, you look for someone to share with. Your demand for housing is not as inelastic as your demand for gum. Why the difference? Housing takes a big chunk of your budget, and gum takes little. You barely notice the higher price of gum, while the higher rent puts your budget under severe strain.

Time Elapsed Since Price Change The longer the time that has elapsed since a price change, the more elastic is demand. When the price of oil increased by 400 percent during the 1970s, people barely changedthe quantity of oil and gasoline they bought. But gradually, as more efficient auto and airplane engineswere developed, the quantity bought decreased. The demand for oil became more elastic as more time elapsed following the huge price hike.

## Elasticity Along a Linear Demand Curve

Elasticity of demand is not the same as slope. And agood way to see this fact is by studying a demand curve that has a constant slope but a varying elasticity.
The demand curve in Fig. 4.3 is linear, which means that it has a constant slope. Along this demand curve, a $\$ 5$ rise in the price brings a decrease of 10 pizzas an hour.
But the price elasticity of demand is not constant along this demand curve. To see why, let's calculate some elasticities.

FIGURE 4.3 Elasticity Along a Linear Demand Curve


On a linear demand curve, demand is unit elastic of the midpoint (elasticity is 1), elastic above the midpoint, and inelastic below the midpoint.
At the midpoint of the demand curve, the price is $\$ 12.50$ and the quantity is 25 pizzas per hour. If the price rises from $\$ 10$ to $\$ 15$ a pizza, the quantity demanded decreases from 30 to 20 pizzas an hour and the average price and average quantity are at the mid- point of the demand curve. So

$$
\begin{aligned}
\text { Price elasticity of demand } & =\frac{10 / 25}{5 / 12.50} \\
& =1
\end{aligned}
$$

That is, at the midpoint of a linear demand curve,the price elasticity of demand is 1. At prices above the midpoint, the price elasticity of demand is greater than 1: Demand is elastic. To seethat demand is elastic, let's calculate the elasticitywhen the price rises from $\$ 15$ to $\$ 25$ a pizza. You can see that quantity demanded decreases from 20 to zeropizzas an hour. The average price is $\$ 20$ a pizza, and the average quantity is 10 pizzas. Putting these numbers into the elasticity formula:

$$
\begin{aligned}
\text { Price elasticity of demand } & =\frac{\Delta Q / Q_{\text {ave }}}{\Delta P / P_{a v e}} \\
& =\frac{20 / 10}{10 / 20} \\
& =4
\end{aligned}
$$

That is, the price elasticity of demand at an averageprice of $\$ 20$ a pizza is 4.
At prices below the midpoint, the price elasticity of
demand is less than 1: Demand is inelastic. For example, if the price rises from zero to $\$ 10$ a pizza, the quantity demanded decreases from 50 to 30 pizzas an hour. The average price is now $\$ 5$ and the average quantity is 40 pizzas an hour. So

$$
\begin{aligned}
\text { Price elasticity of demand } & =\frac{20 / 40}{10 / 5} \\
& =1 / 4
\end{aligned}
$$

That is, the price elasticity of demand at an averageprice of $\$ 5$ a pizza is $1 / 4$.

## Total Revenue and Elasticity

The total revenue from the sale of a good equals the price of the good multiplied by the quantity sold. When a price changes, total revenue also changes. But a cut in the price does not always decrease total revenue. The change in total revenue depends on the elasticity of demand in the following way:

- If demand is elastic, a 1 percent price cut increasesthe quantity sold by more than 1 percent and total revenue increases.
- If demand is inelastic, a 1 percent price cutincreases the quantity sold by less than 1 percent and total revenue decreases.
- If demand is unit elastic, a 1 percent price cutincreases the quantity sold by 1 percent and total revenue does not change.
In Fig. 4.4(a), over the price range $\$ 25$ to $\$ 12.50$ a pizza, demand is elastic. At a price of $\$ 12.50$ a pizza, demand is unit elastic. Over the price range from $\$ 12.50$ a pizza to zero, demand is inelastic.

FIGURE 4.4 Elasticity and Total Revenue

(b) Total revenue

When demand is elastic, in the price range from $\$ 25$ to $\$ 12.50$, a decrease in price (part a) brings an increase in total revenue (part b). When demand is inelastic, in the price range from $\$ 12.50$ to zero, a decrease in price (part a) brings a decrease in total revenue (part b). When demand is unit elastic, at a price of $\$ 12.50$ (part a), total revenue is at a maximum (part b).

Figure 4.4(b) shows total revenue. At a price of $\$ 25$, the quantity sold is zero, so total revenue is zero. At a price of zero, the quantity demanded is 50 pizzas an hour and total revenue is again zero. A price cut in the elastic range brings an increase in total revenue - the percentage increase in the quantity demanded is greater than the percentage decrease in price. A price cut in the inelastic range brings a decrease in total revenue - the percentage increase in the quantity demanded is less than the percentage decrease in price. At unit elasticity, total revenue is at a maximum.

## ECONOMICS IN ACTION

## Elastic and Inelastic Demand

The real-world price elasticities of demand in the table range from 1.52 for metals, the item with the most elastic demand in the table, to 0.05 for oil, the item with the most inelastic demand in the table. The demand for food is also inelastic.

Oil and food, which have poor substitutes and inelastic demand, might be classified as necessities. Furniture and motor vehicles, which have good substitutes and elastic demand, might be classified as luxuries.

Price Elasticities of Demand
Good or Service
Elasticity

| Good or Service | Elasticity |
| :--- | :---: |
| Elastic Demand |  |
| Metals | 1.52 |
| Electrical engineering products | 1.39 |
| Mechanical engineering products | 1.30 |
| Furniture | 1.26 |
| Motor vehicles | 1.14 |
| Instrument engineering products | 1.10 |
| Transportation services | 1.03 |
| Inelastic Demand |  |
| Gas, electricity, and water | 0.92 |
| Chemicals | 0.89 |
| Clothing | 0.64 |
| Banking and insurance services | 0.56 |
| Housing services | 0.55 |
| Agricultural and fish products | 0.42 |
| Books, magazines, and newspapers | 0.34 |
| Food | 0.12 |
| Cigarettes | 0.11 |
| Soft drinks | 0.05 |
| Oil | 0.05 |

## Price Elasticities of Demand for Food

The price elasticity of demand for food in the United States is estimated to be 0.12 . This elasticity is an average over all types of food. The demand for most food items is inelastic, but there is a wide range of elasticities as the figure below shows for a range of fruits, vegetables, and meats.

The demand for grapes and the demand for beef are elastic. The demand for oranges is unit elastic. These food items, especially grapes and beef, have many good substitutes. Florida winter tomatoes have closer substitutes than tomatoes in general, so the demand for the Florida winter variety is more elastic (less inclastic) than the demand for tomatoes.

Carrots and cabbage, on which we spend a very small proportion of income, have an almost zero elastic demand.


Price Elasticities of Demand for Food

Figure 4.4 shows how we can use this relationship between elasticity and total revenue to estimate elasticity using the total revenue test. The
total revenue test is a method of estimating the price elasticity of demand by observing the change in total revenue that results from a change in the price, when all other influences on the quantity sold remain the same.

- If a price cut increases total revenue, demand is elastic.
- If a price cut decreases total revenue, demand is inelastic.
- If a price cut leaves total revenue unchanged, demand is unit elastic


## More Elasticities of Demand

Suppose the economy is expanding and people are enjoying rising incomes. You know that a change in income changes demand. So this increased prosperity brings an increase in the demand for most types of goods and services. By how much will a rise in income increase the demand for pizza? This question is answered by the income elasticity of demand.

## Income Elasticity of Demand

The income elasticity of demand is a measure of the responsiveness of the demand for a good or service to a change in income, other things remaining the same. It tells us by how much a demand curve shifts at a given price.
The income elasticity of demand is calculated by using the formula:

$$
\begin{gathered}
\text { Income elasticity } \\
\text { of demand }
\end{gathered}=\frac{\text { in quantity demanded }}{\text { Percentage change in income }} .
$$

Income elasticities of demand can be positive or negative and they fall into three interesting ranges:

- Positive and greater than 1 ( normal good, income elastic)
- Positive and less than 1 (normal good, income inelastic)
- Negative (inferior good)

Income Elastic Demand Suppose that the price of pizza is constant and 9 pizzas an hour are bought. Then incomes rise from $\$ 975$ to $\$ 1,025$ a week. No other influence on buying plans changes and the quantity of pizzas sold increases to 11 an hour.
The change in the quantity demanded is +2 pizzas.
The average quantity is 10 pizzas, so the quantity demanded increases by 20 percent. The change in income is $+\$ 50$ and the average income is $\$ 1,000$, so incomes increase by 5 percent. The income elasticity of demand for pizza is

$$
\frac{20 \%}{5 \%}=4
$$

The demand for pizza is income elastic. The percentage increase in the quantity of pizza demandedexceeds the percentage increase in income.

## ECONOMICS IN ACTION

## Necessities and Luxuries

The demand for a necessity such as food or clothing is income inelastic, while the demand for a luxury such as airline and foreign travel is income elastic. But what is a necessity and what is a luxury depends on the level of income. For people with a low income, food and clothing can be luxuries. So the level of income has a big effect on income elasticities of demand.

The figure shows this effect on the income elasticity of demand for food in 10 countries. In countries with low incomes, such as Tanzania and India, the income elasticity of demand for food is high. In countries with high incomes, such as Canada, the income elasticity of demand for food is low. That is, as income increases, the income elasticity of demand for food decreases. Low-income consumers spend a larger percentage of any increase in income on food than do high-income consumers.


Income Elasticities in $\mathbf{1 0}$ Countries

Income Inelastic Demand If the income elasticity of demand is positive but less than I, demand is incomeinelastic. The percentage increase in the quantity demanded is positive but less than the percentage increase in income.
Whether demand is income elastic or income inelastic has an important implication for the percentage of income spent on a good. If the demand for a good is income elastic, the
percentage of income spent on that good increases as income increases. And if the demand for a good is income inelastic, the percentage of income spent on that good decreases as income increases.

ECONOMICS IN ACTION Income Elastic and Inelastic Demand
The table shows some real-world income elasticites of demand and confirms that a necessity such as food or clothing is income inelastic, while the demand for a luxury such as airline travel is income elastic.

## Some Real-World Income Elasticities of Demand

Income Elastic Demand
Airline travel 5.82
Movies 3.41
Foreign travel 3.08
Electricity 1.94
Restaurant meals 1.61
Local buses and trains 1.38
Haircuts 1.36
Automobiles $\quad 1.07$

Income Inelastic Demand
Tobacco 0.86
Alcoholic drinks 0.62
Furniture 0.53
Clothing 0.51
Newspapers and magazines 0.38
Telephone 0.32
Food 0.14

Inferior Goods If the income elasticity of demand is negative, the good is an inferior good. The quantity demanded of an inferior good and the amount spent on it decrease when income increases. Goods in this category include small motorcycles, potatoes, and rice. Low-income consumers buy these goods and spend a large percentage of their incomes on them.

## Cross Elasticity of Demand

The burger shop next to your pizzeria has just raised its price of a burger. You know that pizzas and burgers are substitutes. You also know that when the price of a substitute for pizza rises, the demand for pizza increases. But how big is the influence of the price of a burger on the demand for pizza?
You know, too, that pizza and soft drinks are complements. And you know that if the price of a complement of pizza rises, the demand for pizza decreases. So you wonder, by how much will a rise in the price of asoft drink decrease the demand for your pizza? To answer this question, you need to know about the cross elasticity of demand for pizza. Let's examine this elasticity measure.
We measure the influence of a change in the price of a substitute or complement by using the concept of the cross elasticity of demand. The cross elasticity of demand is a measure of the responsiveness of the demand for a good to a change in the price of a substitute or complement, other things remaining the same. We calculate the cross elasticity of demand by using the formula:

## Percentage change

$\begin{gathered}\text { Cross elasticity } \\ \text { of demand }\end{gathered}=\frac{\text { in quantity demanded }}{\begin{array}{c}\text { Percentage change in price of } \\ \text { a substitute or complement }\end{array}}$.
The cross elasticity of demand can be positive or negative. If the cross elasticity of demand is positive, demand and the price of the other good change in the same direction, so the two goods are substitutes. If the cross elasticity of demand is negative, demand and the price of the other good change in opposite directions, so the two goods are complements.

Substitutes Suppose that the price of pizza is constant and people buy 9 pizzas an hour. Then the priceof a burger rises from $\$ 1.50$ to $\$ 2.50$. No other influence on buying plans changes and the quantity of pizzas bought increases to 11 an hour.
The change in the quantity demanded at the cur- rent price is +2 pizzas-the new quantity, 11 pizzas, minus the original quantity, 9 pizzas. The average quantity is 10 pizzas. So the quantity of pizzas demanded increases by 20 percent. That is,

$$
\Delta Q^{\prime} Q_{\text {ave }} \times 100=(+2 / 10) \times 100=+20 \%
$$

The change in the price of a burger, a substitute for pizza, is $+\$ 1$-the new price, $\$ 2.50$, minus theoriginal price, $\$ 1.50$. The average price is $\$ 2$ a burger. So the price of a burger rises by 50 percent. That is,

$$
\Delta P / P_{\text {ave }} \times 100=(+\$ 1 / \$ 2) \times 100=+50 \%
$$

So the cross elasticity of demand for pizza with respect to the price of a burger is

$$
\frac{+20 \%}{+50 \%}=0.4
$$

Figure 4.5 illustrates the cross elasticity of demand.
Because pizza and burgers are substitutes, when the price of a burger rises, the demand for pizza increases. The demand curve for pizza shifts rightward from $D_{0}$ to $D_{1}$. Because a rise in the price of a burger brings an increase in the demand for pizza, the cross elasticity of demand for pizza with respect to the price of a burger is positive. Both the price and the quantity change in the same direction.

Complements Now suppose that the price of pizza is constant and 11 pizzas an hour are bought. Then the price of a soft drink rises from $\$ 1.50$ to $\$ 2.50$. No other influence on buying plans changes and the quantity of pizzas bought falls to 9 an hour.

FIGURE 4.5 Cross Elasticity of Demand


A burger is a substitute for pizza. When the price of a burger rises, the demand for pizza increases and the demand curve for pizza shifts rightward from $D_{0}$ to $D_{1}$. The cross elasticity of demand is positive.

A soft drink is a complement of pizza. When the price of a soft drink rises, the demand for pizza decreases and the demand curve for pizza shiffs leffward from $D_{0}$ to $D_{2}$. The cross elasticity of demand is negative.

The change in the quantity demanded is the opposite of what we've just calculated: The quantity of pizzas demanded decreases by 20 percent (-20\%).
The change in the price of a soft drink, a rise of $\$ 1$ from $\$ 1.50$ to $\$ 2.50$, is the same as the change in the price of a burger that we've just calculated. That is, the price rises by 50 percent (+50\%). So the cross elasticity of demand for pizza with respect to the price of a soft drink is

$$
\frac{-20 \%}{+50 \%}=-0.4
$$

Because pizza and soft drinks arecomplements, when the price of a soft drink rises, the demand for pizza decreases.
In Fig. 4.5, when the price of soft drinks rises the demand curve for pizza shifts leftward from $D_{0}$ to $D_{2}$. Because a rise in the price of a soft drink brings a decrease in the demand for pizza, the cross elasticity of demand for pizza with respect to the price of a softdrink is negative. The price and quantity change inopposite directions.
The magnitude of the cross elasticity of demand determines how far the demand curve shifts. The larger the cross elasticity (absolute value), the greater is the change in demand and the larger is the shift in the demand curve.
If two items are close substitutes, such as two brands of spring water, the cross elasticity is large. If two items are close complements, such as movies and popcorn, the cross elasticity is large.
If two items are somewhat unrelated to each other, such as newspapers and orange juice, the cross elasticity is small - perhaps even zero.

## Elasticity of Supply

You know that when demand increases, the equilibrium price rises and the equilibrium quantity increases. But does the price rise by a large amount and the quantity increase by a little? Or does theprice barely rise and the quantity increase by a large amount?
The answer depends on the responsiveness of the quantity supplied to a change in the price. If thequantity supplied is not very responsive to price, then an increase in demand brings a large rise in the price and a small increase in the equilibrium quantity. If the quantity supplied is highly responsive to price, then an increase in demand brings a small rise in the price and a large increase in the equilibrium quantity.
The problems that arise from using the slope of the supply curve to indicate responsiveness are thesame as those we considered when discussing theresponsiveness of the quantity demanded, so we usea units - free measure - the elasticity of supply.

## Calculating the Elasticity of Supply

The elasticity of supply measures the responsiveness of the quantity supplied to a change in the price of a good when all other influences on selling plans remain the same. It is calculated by using the formula:

$$
\begin{aligned}
& \begin{array}{c}
\text { Percentage change in } \\
\text { of supply }
\end{array}
\end{aligned}=\frac{\text { quantity supplied }}{\text { Percentage change in price }} .
$$

We use the same method that you learned when youstudied the elasticity of demand. (Refer back to p. 122 to check this method.)

Elastic and Inelastic Supply If the elasticity of supply is greater than 1 , we say that supply is elastic; and ifthe elasticity of supply is less than I, we say that supply is inelastic.
Suppose that when the price rises from $\$ 20$ to $\$ 21$, the quantity supplied increases from 10 to 20 pizzas per hour. The price rise is $\$ 1$ and the average price is $\$ 20.50$, so the price rises by 4.9 percent of the average price. The quantity increases from 10 to 20 pizzas an hour, so the increase is 10 pizzas, the average quantity is 15 pizzas, and the quantity increases by 67 percent. The elasticity of supply is equal to 67 percent divided by 4.9 percent, which equals 13.67. Because the elasticity of supply exceeds 1 (in this case by a lot), supply is elastic. In contrast, suppose that when the price rises from $\$ 20$ to $\$ 30$, the quantity of pizza supplied increases from 10 to 13 per hour. The price rise is $\$ 10$ and theaverage price is $\$ 25$, so the price rises by 40 percent of the average price. The quantity increases from 10to 13 pizzas an hour, so the increase is 3 pizzas, theaverage quantity is 11.5 pizzas an hour, and the quantity increases by 26 percent. The elasticity of supply is equal to 26 percent divided by 40 percent, which equals 0.65 . Now, because the elasticity of supply is less than I, supply is inelastic. Figure 4.6 shows the range of elasticities of supply.
If the quantity supplied is fixed regardless of the price, the supply curve is vertical and the elasticity of supply is zero. Supply is perfectly inelastic. This case is shown in Fig. 4.6(a). A special intermediate case occurs when the percentage change in price equals the percentage change in quantity. Supply is then unit elastic. This case is shown in Fig. 4.6(b). No matter how steep the supply curve is, if it is linear and passes through the origin, supply is unit elastic. If there is a price at which sellers are willing to offer any quantity for sale, the supply curve is horizontal and the elasticity of supply is infinite. Supply is perfectly elastic. This case is shown in Fig. 4.6(c).

## The Factors That Influence the Elasticity of Supply

The elasticity of supply of a good depends on

- Resource substitution possibilities
- Time frame for the supply decision

Resource Substitution Possibilities Some goods and services can be produced only by using unique or rare productive resources. These items have a low, perhaps even a zero, elasticity of supply. Other goods and services can be produced by using commonly available resources that could be allocated toa wide variety of alternative tasks. Such items have a high elasticity of supply.
A van Gogh painting is an example of a good with a vertical supply curve and a zero elasticity of supply. At the other extreme, wheat can be grown on land that is almost equally good for growing corn, so it is just as easy to grow wheat as corn. The opportunity cost of wheat in terms of forgone corn is almost constant. As a result, the supply curve of wheat is almost horizontal and its elasticity of supply is very large.
Similarly, when a good is produced in many differentcountries (for example, sugar and beef), the supply ofthe good is highly elastic.
The supply of most goods and services lies between these two extremes. The quantity produced can be increased but only by incurring a higher cost. If a higher price is offered,
the quantity supplied increases. Such goods and services have an elasticity of supply between zero andinfinity.

FIGURE 4.6 Inelastic and Elastic Supply

(a) Perfectly inelastic supply

(b) Unit elastic supply

(c) Perfectly elastic supply

Each supply illustrated here has a constant elasticity. The supply curve in part (a) illustrates the supply of a good that has a zero elasticity of supply. Each supply curve in part (b) illustrates the supply of a good with a unit elasticity of
supply. All linear supply curves that pass through the origin illustrate supplies that are unit elastic. The supply curve in part (c) illustrates the supply of a good with an infinite elasticity of supply.

Time Frame for the Supply Decision To study the influence of the amount of time elapsed since a price change; we distinguish three time frames of supply:

- Momentary supply
- Short-run supply
- Long-run supply

Momentary Supply When the price of a good changes, the immediate response of the quantity supplied is determined by the momentary supply of that good.
Some goods, such as fruits and vegetables, have a perfectly inelastic momentary supply-a verticalsupply curve. The quantities supplied depend on crop-planting decisions made earlier. In the case of oranges, for example, planting decisions have to be made many years in advance of the crop being available. Momentary supply is perfectly inelastic because, on a given day, no matter what the price of oranges, producers cannot change their output. They have picked, packed, and shipped their crop to market, and the quantity available for that day is fixed.
In contrast, some goods have a perfectly elastic momentary supply. Long-distance phone calls are an example. When many people simultaneously make a call, there is a big surge in the demand for telephone cables, computer switching, and satellite time. The quantity supplied increases, but the price remains constant. Long-distance carriers monitor fluctuations in demand and reroute calls to ensure that the quantity supplied equals the quantity demanded without changing the price.

Short-Run Supply The response of the quantity sup- plied to a price change when only some of the possible adjustments to production can be made is determined by short-run supply. Most goods have an inelastic short-run supply. To increase output in the short run, firms must work their labor force overtime and perhaps hire additional workers. To decrease their output in the short run, firms either lay off workers or reduce their hours of work. With the passage of time, firms can make more adjustments, perhaps training additional workers or buying additional tools and other equipment.

For the orange grower, if the price of oranges falls, some pickers can be laid off and oranges left on the trees to rot. Or if the price of oranges rises, the grower can use more fertilizer and improved irrigation to increase the yield of the existing trees. But an orange grower can't change the number of trees producing oranges in the short run.

Long-Run Supply The response of the quantity supplied to a price change after all the technologically possible ways of adjusting supply have been exploited is deter- mined by longrun supply. For most goods and services, long-run supply is elastic and perhaps perfectly elastic. For the orange grower, the long run is the time it takes new tree plantings to grow to full maturity - about 15 years. In some cases, the long-run adjustment occurs only after a completely new production plant has been built and workers have been trained to operate it typically a process that might take several years.

## ECONOMIC ANALYSIS

- The table below summarizes data provided by the International Coffee Organization, which supplements and updates some of the information in this news article.


## Summary of Coffee Data

| Year | Quantity produced <br> (millions of bags) | Price <br> (U.s. cents per pound) |
| :---: | :---: | :---: |
| 2012 | 134 | 135 |
| 2013 | 145 | 100 |

- Figure 1 provides yet more data and shows that after falling through 2013, the price of coffee climbed steeply in 2014.
- The price of coffee fluctuates because the supply of coffee fluctuates. And the price flucutates by much more than the quantity of coffee produced because the demand for coffee is inelastic.
- The news headline provides the first clue that demand is inelastic: When the price falls, revenue shrinks. This information enables us to use the total revenue test, "If a price cut decreases total revenue, then demand is inelastic."
- We can estimate the price elasticity of demand for coffee, assuming that demand didn't change, by using the events in the market in 2012 and 2013.
- Figure 2 illustrates the global market for coffee in these two years. The demand curve for coffee is $D$ and in 2012 , the supply curve of coffee was $S_{12}$. The equilibrium price was 135 cents per pound, and the equilibrium quantity was 134 million bags.
- In 2013, supply increased to $S_{13}$, the price fell to 100 cents per pound, and the quantity increased to 145 million bags.
- Figure 3 focuses on the demand curve and summarizes the elasticity calculation. The price fell by 35 cents, which is 30 percent of the average price of 117 cents. The quantity demanded increased by 11 million bags, which is 7.9 percent of the average quantity.
- The price elasticity of demand is 7.9 percent/ 30 percent, which equals 0.26 . A 1 percent fall in price brings a 0.26 percent increase in the quantity demanded. And a 1 percent increase in the quantity brings a fall in price equal to $1 / 0.26$, or almost 4 percent.
- When demand is inelastic, a small percentage change in the supply brings a large percentage change in the price.


Figure 1 The Coffee-Price Roller Coaster


Figure 2 The Market for Coffee: 2012 and 2013


Figure 3 The Price Elasticity of Demand for Coffee

## CHAPTER 5: Efficiency and Equity

After studying this chapter, you will be able to:

- Describe the alternative methods of allocating scarce resources
- Explain the connection between demand and marginal benefit and define consumer surplus; and explain the connection between supply and marginal cost and define producer surplus
- Explain the conditions under which markets are efficient and inefficient
- Explain the main ideas about fairness and evaluate claims that markets result in unfair outcomes
Every day, millions of people make self-interested choices to drive to work rather than take the bus or train. The outcome of these choices is gridlock and alot of lost time. Are we using our highways and our time efficiently?
One way of eliminating traffic jams is to make people pay for road use - to make all the highwaystoll-roads. Rich people can easily pay a toll, but poor people can't afford to pay. Would tolls be fair?
We'll answer these questions about highway use in Economics in the News at the end of the chapter. But first, we examine the efficiency and fairness of alternative ways of allocating scarce resources.


## Resource Allocation Methods

If resources were abundant, and not scarce, we would not need to allocate them among alternative uses. But resources are scarce: They must be allocated some-how. Our goal is to discover how resources might be allocated efficiently and fairly. So what are the alternative methods of allocating scarce resources?
Eight alternative methods that might be used are

- Marketprice
- Command
- Majority rule
- Contest
- First-come, first-served
- Lottery
- Personal characteristics
- Force

Let's briefly examine each method.

## Market Price

When a market price allocates a scarce resource, thepeople who are willing and able to pay that price get the resource. Two kinds of people decide not to paythe market price: those who can afford to pay but choose not to buy and those who are too poor andsimply can't afford to buy.

For many goods and services, distinguishing between those who choose not to buy and those who can't afford to buy doesn't matter. But for a few items, it does matter. For example, poor people can't afford to pay school fees and doctors' fees. Because poor people can't afford items that most people consider to be essential, these items are usually allocated by one of the other methods.

## Command

A command system allocates resources by the order (command) of someone in authority. In the U.S. economy, the command system is used extensively inside firms and government departments. For example, if you have a job, most likely someone tells you what to do. Your labor is allocated to specific tasks by a command.
A command system works well in organizations in which the lines of authority and responsibility are clear and it is easy to monitor the activities being per-
formed. But a command system works badly when the range of activities to be monitored is large and when it is easy for people to fool those in authority. North Korea uses a command system and it works so badly that it even fails to deliver an adequate supply of food.

## Majority Rule

Majority rule allocates resources in the way that a majority of voters choose. Societies use majority rule to elect representative governments that make some of the biggest decisions. For example, majority rule decides the tax rates that end up allocating scarce resources between private use and public use. And majority rule decides how tax dollars are allocated among competing uses such as education and healthcare. Majority rule works well when the decisions being made affect large numbers of people and self-interest must be suppressed to use resources mosteffectively.

## Contest

A contest allocates resources to a winner (or a groupof winners). Sporting events use this method. SerenaWilliams competes with other tennis professionals, and the winner gets the biggest payoff. But contestsare more general than those in a sports arena, though we don't normally call them contests. For example, Bill Gates won a contest to provide the world's personal computer operating system.
Contests do a good job when the efforts of the "players" are hard to monitor and reward directly. When a manager offers everyone in the company the opportunity to win a big prize, people are motivated to work hard and try to become the winner. Only a few people end up with a big prize, but many people work harder in the process of trying to win. The total output produced by the workers is much greater than it would be without the contest.

## First-Come, First-Served

A first-come, first-served method allocates resources to those who are first in line. Many casual restaurants won't accept reservations. They use first-come, first-served to allocate their scarce tables. Highway space is allocated in this way too: The first to arrive at the on-ramp
gets the road space. If too many vehicles enter the highway, the speed slows and people wait in line for some space to become available.
First-come, first-served works best when, as in the above examples, a scarce resource can serve just one user at a time in a sequence. By serving the user whoarrives first, this method minimizes the time spent waiting for the resource to become free.

## Lottery

Lotteries allocate resources to those who pick the winning number, draw the lucky cards, or come uplucky on some other gaming system. State lotteriesand casinos reallocate millions of dollars' worth ofgoods and services every year.
But lotteries are more widespread than jackpots and roulette wheels in casinos. They are used to allocate landing slots to airlines at some airports and places in the New York and Boston marathons, and have been used to allocate fishing rights and the electro- magnetic spectrum used by cellphones.
Lotteries work best when there is no effective way to distinguish among potential users. of a scarce resource.

## Personal Characteristics

When resources are allocated on the basis of personal characteristics, people with the "right" characteristics get the resources. Some of the resources that matter most to you are allocated in this way. For example, you will choose a marriage partner on the basis of personal characteristics. But this method can also be used in unacceptable ways. Allocating the best jobs to white, Anglo-Saxon males and dis-criminating against visible minorities and women is an example.

## Force

Force plays a crucial role, for both good and ill, in allocating scarce resources. Let's start with the ill. War, the use of military force by one nation against another, has played an enormous rolehistorically in allocating resources. The economic supremacy of European settlers in the Americas and Australia owes much to the use of this method. Theft, the taking of the property of others without their consent, also plays a large role. Both large-scale organized crime and small-scale petty crime collectively allocate billions of dollars' worth of resourcesannually.
But force plays a crucial positive role in allocating resources. It provides the state with an effective method of transferring wealth from the rich to the poor, and it provides the legal framework in which voluntary exchange in markets takes place.
A legal system is the foundation on which our market economy functions. Without courts toenforce contracts, it would not be possible to do business. But the courts could not enforce contracts without the ability to apply force if necessary. The state provides the ultimate force that enables the courts to do their work.
More broadly, the force of the state is essential to uphold the principle of the rule of law. This principle is the bedrock of civilized economic (and social and political) life. With the rule of law upheld, people can go about their daily economic lives with the assurance that
their property will be protected-that they can sue for violations against their property (and be sued if they violate the property of others).
Free from the burden of protecting their property and confident in the knowledge that those with whom they trade will honor their agreements, people can get on with focusing on the activity in which they have a comparative advantage and trading for mutual gain.

## Benefit, Cost, and Surplus

Resources are allocated efficiently and in the social interest when they are used in the ways that people value most highly. You saw in Chapter 2 that this outcome occurs when the quantities produced are at the point on the PPF at which marginal benefit equals marginal cost (see pp. 73-75). We're now going to see whether competitive markets produce the efficient quantities. We begin on the demand side of a market.

## Demand, Willingness to Pay, and Value

In everyday life, we talk about "getting value formoney." When we use this expression, we are distinguishing between value and price. Value is what weget, and price is what we pay. The value of one more unit of a good or service is its marginal benefit. We measure marginal benefit by the maximum price that is willingly paid for another unit of the good or service. But willingness to paydetermines demand. A demand curve is a marginal benefit curve.
In Fig. 5.1(a), Lisa is willing to pay $\$ 1$ for the 30th slice of pizza and $\$ 1$ is her marginal benefit from that slice. In Fig. 5.1(b), Nick is willing to pay $\$ 1$ for the 10th slice of pizza and $\$ 1$ is his marginalbenefit from that slice. But at what quantity is the market willing to pay $\$ 1$ for the marginal slice? Theanswer is provided by the market demand curve.

HGURE 5.1 Individual Demand, Market Demand, and Marginal Social Benefit

(a) Lisa's demand

(b) Nick's demand

(c) Market demand

At a price of $\$ 1$ a slice, the quantity demanded by Lisa is 30 slices and the quantity demanded by Nick is 10 slices, so the quantity demanded by the market is 40 slices. Lisa's demand
curve in part (a) and Nick's demand curve in part (b) sum horizontally to the market demand curve in part (c). The market demand curve is the marginal social benefit $(M S B)$ curve.

## Individual Demand and Market Demand

The relationship between the price of a good and the quantity demanded by one person is called individual demand. And the relationship between the price of a good and the quantity demanded by all buyers is called market demand.
The market demand curve is the horizontal sum of the individual demand curves and is formedby adding the quantities demanded by all the individuals at each price.
Figure 5.1(c) illustrates the market demand for pizza if Lisa and Nick are the only people in the market. Lisa's demand curve in part (a) and Nick's demand curve in part (b) sum horizontally to the market demand curve in part (c).

## Consumer Surplus

We don't always have to pay as much as we are willing to pay. We get a bargain. When people buy something for less than it is worth to them, they receive a consumer surplus. Consumer surplus is the excess of the benefit received from a good over the amount paid for it. We can calculate consumer surplus as themarginal benefit (or value) of a good minus its price, summed over the quantity bought.
Figure $5.2(\mathrm{a})$ shows Lisa's consumer surplus from pizza when the price is $\$ 1$ a slice. At this price, shebuys 30 slices a month because the 30th slice is worth exactly $\$ 1$ to her. But Lisa is willing to pay $\$ 2$ for the10th slice, so her marginal benefit from this slice is $\$ 1$ more than she pays for it-she receives a surplus of $\$ 1$ on the 10 th slice.
Lisa's consumer surplus is the sum of the surpluses on all of the slices she buys. This sum is the area of the triangle-the area below the demand curve andabove the market price line. The area of this triangle is equal to its base ( 30 slices) multiplied by its height ( $\$ 1.50$ ) divided by 2, which is $\$ 22.50$. The area of the rectangle in Fig. 5.2(a) shows what Lisa pays for 30 slices of pizza. Figure $5.2(b)$ shows Nick's consumer surplus, and part (c) shows the consumer surplus for the market. The consumer surplus for the market is the sum of the consumer surpluses of Lisa and Nick.
All goods and services have decreasing marginal benefit, so people receive more benefit from their consumption than the amount they pay.

## Supply and Marginal Cost

Your next task is to see how market supply reflects marginal cost. The connection between supply and cost closely parallels the related ideas about demand and benefit that you've just studied. Firms are in business to make a profit. To do so, they must sell their output for a price that exceeds the cost of production. Let's investigate the relationship between cost and price.

## Supply, Cost, and Minimum Supply-Price

Firms make a profit when they receive more from the sale of a good or service than the cost of producing it.
Just as consumers distinguish between value andprice, so producers distinguish between cost and price. Cost is what a firm gives up when it produces a good or service and price is what a firm receives when it sells the good or service.

FIGURE 5.2 Demand and Consumer Surplus


Lisa is willing to pay $\$ 2$ for her 10th slice of pizza in part (a). At a market price of $\$ 1$ a slice, Lisa receives a surplus of $\$ 1$ on the 10 th slice. The green triangle shows her consumer surplus on the 30 slices she buys at $\$ 1$ a slice. The
green triangle in part (b) shows Nick's consumer surplus on the 10 slices that he buys at $\$ 1$ a slice. The green area in part (c) shows the consumer surplus for the market. The blue rectangles show the amounts spent on pizza.

The cost of producing one more unit of a good or service is its marginal cost. Marginal cost is the mini-mum price that producers must receive to induce them to offer one more unit of a good or service for sale. But the minimum supply-price determines supply. A supply curve is a marginal cost curve.
In Fig. 5.3(a), Maria is willing to produce the $100^{\text {th }}$ pizza for $\$ 15$, her marginal cost of that pizza. In Fig. 5.3(b), Max is willing to produce the 50th pizza for $\$ 15$, his marginal cost. What quantity is this market willing to produce for $\$ 15$ a pizza? The answer is provided by the market supplycurve.

## Individual Supply and Market Supply

The relationship between the price of a good and thequantity supplied by one producer is called individual supply. And the relationship between the price of a good and the quantity supplied by all producers is called market supply.

The market supply curve is the horizontal sum of the individual supply curves and is formed byadding the quantities supplied by all the producersat each price. Figure 5.3(c) illustrates the market supply of pizzas if Maria and Max are the only producers.
Maria's supply curve in part (a) and Max's supply curve in part (b) sum horizontally to the market supply curve in part (c).
At a price of $\$ 15$ a pizza, Maria supplies 100 pizzas and Max supplies 50 pizzas, so the quantity sup- plied by the market at $\$ 15$ a pizza is 150 pizzas.
For Maria and Max, their supply curves are their marginal cost curves. For society, the market supplycurve is its marginal cost curve. We call the society'smarginal cost marginal social cost. So the market supply curve is also the marginal social cost (MSC) curve.

FIGURE 5.3 Individual Supply, Market Supply, and Marginal Social Cost

(a) Maria's supply

(b) Max's supply

(c) Market supply

At a price of $\$ 15$ a pizza, the quantity supplied by Maria is 100 pizzas and the quantity supplied by Max is 50 pizzas, so the quantity supplied by the market is 150 pizzas. Maria's
supply curve in part (a) and Max's supply curve in part (b) sum horizontally to the market supply curve in part (c). The market supply curve is the marginal social cost (MSC) curve.

## Producer Surplus

When price exceeds marginal cost, the firm receives a producer surplus. Producer surplus is the excess of the amount received from the sale of a good or service over the cost of producing it. Wecalculate producer surplus as the price received minus the marginal cost (or minimum supply-price), summed over the quantity sold.
FIGURE 5.4 Supply and Producer Surplus

(a) Maria's producer surplus

(b) Max's producer surplus

(c) Market producer surplus

Figure 5.4(a) shows Maria's producer surplus from pizza when the price is $\$ 15$ a pizza. At this price, shesells 100 pizzas a month because the $100^{\text {th }}$ pizza costs her $\$ 15$ to produce. But Maria is willing to produce the 50th pizza for her marginal cost, which is $\$ 10$, so she receives a surplus of $\$ 5$ on this pizza.
Maria's producer surplus is the sum of the surpluses on the pizzas she sells. This sum is the area of the triangle - the area below the market price and above the supply curve. The
area of this triangle is equal to its base (100) multiplied by its height (\$10) divided by 2, which is $\$ 500$.
The shaded area below the supply curve in Fig. 5.4(a) shows what it costs Maria to produce 100 pizzas.
The area of the triangle in Fig. 5.4(b) shows Max's producer surplus and the area above the curve in Fig. 5.4(c) shows the producer surplus for the market The producer surplus for the market is the sum of the producer surpluses of Maria and Max.

Maria is willing to produce the 50th pizza for $\$ 10$ in part(a). At a market price of $\$ 15$ a pizza, Maria gets a surplus of $\$ 5$ on the 50th pizza. The triangle shows her producer surplus on the 100 pizzas she sells at $\$ 15$ each. The triangle in part (b) shows Max's producer surplus on the 50 pizzas that he sells at $\$ 15$ each. The area above the curve in part (c) shows producer surplus for the market. The shaded areas show the cost of producing the pizzas sold.

## Is the Competitive Market Efficient?

Figure 5.5(a) shows the market for pizza. The market forces that you studied in Chapter 3 (pp. 104-105)pull the pizza market to its equilibrium price of $\$ 15$ a pizza and equilibrium quantity of 10,000 pizzas a day. Buyers enjoy a consumer surplus and sellers enjoy a producer surplus, but is this competitive equilibrium efficient?

## Efficiency of Competitive Equilibrium

You've seen that the market demand curve for agood or service tells us the marginal social benefit from it. You've also seen that the market supply curve of a good or service tells us the marginal social cost of producing it.
Equilibrium in a competitive market occurs when the quantity demanded equals the quantity supplied at the intersection of the demand curve and the sup- ply curve. At this intersection point, marginal social benefit on the demand curve equals marginal social cost on the supply curve. This equality is the condition for allocative efficiency. So in equilibrium, a competitive market achieves allocative efficiency.

Figure 5.5 illustrates the efficiency of competitive equilibrium. The demand curve and the supply curve intersect in part (a) and marginal social benefit equals marginal social cost in part (b). If production is less than 10,000 pizzas a day, the marginal pizza is valued more highly than it costs toproduce. If production exceeds 10,000 pizzas a day, the marginal pizza costs more to produce than the value that consumers place on it. Only when 10,000 pizzas a day are produced is the marginal pizza worth exactly what it costs.
The competitive market pushes the quantity of pizzas produced to its efficient level of 10,000 a day. If production is less than 10,000 pizzas a day, a short- age raises the price, which increases production. If production exceeds 10,000 pizzas a day, a surplus of pizzas lowers the price, which decreases production. So a competitive pizza market is efficient. Figure 5.5(a) also shows the consumer surplus and producer surplus. The sum of consumer surplus and producer surplus is called total surplus. When the efficient quantity is produced, total surplus is maximized.Buyers and sellers acting in their self-interest end up promoting the social interest.

## FIGURE 5.5 An Efficient Market for Pizza


(a) Equilibrium and surpluses


## (b) Efficiency

Competifive equilibrium in part (a) occurs when the quantity demanded equals the quantity supplied. Resources are used efficiently in part (b) when marginal social benefit, MSB, equals marginal social cost, MSC. Total surplus, which is the sum of consumer surplus (the green triangle) and producer surplus (the blue triangle) is maximized.

The efficient quantity in part (b) is the same as the equilibrium quantity in part (a). The competitive pizza market produces the efficient quantity of pizzas.

## Market Failure

Markets are not always efficient, and when a market is inefficient, we call the outcome market failure. In a market failure, either too little (underproduction) or too much (overproduction) of an item is produced.

Underproduction In Fig. 5.6(a), the quantity of pizzas produced is 5,000 a day. At this quantity, consumers are willing to pay $\$ 20$ for a pizza that costs only $\$ 10$ to produce. The quantity produced is inefficient - there is underproduction-and total surplus is smaller than itsmaximum possible level.
We measure the scale of inefficiency by deadweight loss, which is the decrease in total surplus that results from an inefficient level of production. The triangle in Fig. 5.6(a) shows the deadweight loss.

Overproduction In Fig. 5.6(b), the quantity of pizzas produced is 15,000 a day. At this quantity, consumers are willing to pay only $\$ 10$ for a pizza that costs $\$ 20$ to produce. By producing the 15,000th pizza, $\$ 10$ of resources are wasted. Again, the gray triangle shows the deadweight loss, which reduces the total surplus to less than its maximum. Inefficient production creates a deadweight loss that is borne by the entire society: It is a social loss.

- Price and quantity regulations
- Taxes and subsidies
- Externalities
- Public goods and common resources Monopoly
- High transactions costs

Price and Quantity Regulations Aprice regulation, either a price cap or a price floor, blocks the price adjustments that balance the quantity demanded and the quantity supplied and lead to underproduction. A quantity regulation that limits the amount that a farm is permitted to produce also leads to underproduction.

Taxes and Subsidies Taxes increase the prices paid by buyers, lower the prices received by sellers, and lead
to underproduction. Subsidies, which are payments by the government to producers, decrease the pricespaid by buyers, increase the prices received by sellers, and lead to overproduction.

Externalities An externality is a cost or a benefit that affects someone other than the seller or the buyer. Anexternal cost arises when an electric utility burns coal and emits carbon dioxide. The utility doesn't con-sider the cost of climate change when it decides howmuch power to produce. The result is overproduction. An external benefit arises when an apartment owner installs a smoke detector and decreases herneighbor's fire risk. She doesn't consider the benefit toher neighbor when she decides how many detectors to install. The result is underproduction.

Figure 5.6 Underproduction and Overproduction

(a) Underproduction

(b) Overproduction

If 5,000 pizzas a day are produced, in part (a), total surplus (the sum of the green and blue areas) is smaller than its maxi- mum by the amount of the deadweight loss (the triangle). At all quantities below 10,000 pizzas a day, the benefit from one more pizza exceeds its cost. If 15,000 pizzas a day are produced, in part (b), total surplus is also smaller than its maximum by the amount of the deadweight loss. At all quantities in excess of 10,000 pizzas a day, the cost of one more pizza exceeds its benefit.

Public Goods and Common Resources Apublic good is a good or service from which everyone benefits and no one can be excluded. National defense is an example. A competitive market would underproduce national defense because everyone would try tofree ride on everyone else.
A common resource is owned by no one but is available to be used by everyone. Atlantic salmon is anexample. It is in everyone's self-interest to ignore the costs they impose on others when they decide how much of a common resource to use: It is overused.

Monopoly A monopoly is a firm that is the sole provider of a good or service. Local water supplyand cable television are supplied by firms that are monopolies. The monopoly's selfinterest is to maximize its profit, and because it has no competitors, itproduces too little and charges too high a price: It underproduces.

High Transaction Costs When you buy your first house, you will also buy the services of an agent and a lawyer to do the transaction. Economists call the costs of the services that enable a market to bring buyers and sellers together transactions costs. It is costly to operate any market, but some markets are so costly to operate that they simply don't. For example, there's no market in time slots on a local tennis court. Instead, the court uses first-come, first-served: You hang around until the court becomes vacant and "pay" with your waiting time. When transactions costs are high, the market might underproduce.
You now know the conditions under which resource allocation is efficient. You've seen how a competitive market can be efficient, and you've seen some obstacles to efficiency. Can alternative allocation methods improve on the market?

## Alternatives to the Market

When a market is inefficient, can one of the alternative nonmarket methods that we described at the beginning of this chapter do a better job? Sometimes it can.
Often, majority rule might be used in an attempt to improve the allocation of resources. But majority rule has its own shortcomings. A group that pursues the self-interest of its members can become the majority. For example, a price or quantity regulation that creates inefficiency is almost always the result of a self-interested group becoming the majority and imposing costs on the minority. Also, with majority rule, votes must be translated into actions by bureaucrats who have their own agendas based on their self-interest.
Managers in firms issue commands and avoid the transactions costs that they would incur if they went to a market every time they needed a job done.
First-come, first-served works best in some situations. Think about the scene at a busy ATM. Insteadof waiting in line people might trade places at a "market" price. But someone would need to ensure that trades were honored. At a busy ATM, first-come, first-served is the most efficient arrangement.
There is no one efficient mechanism that allocates all resources efficiently. But markets, when supplemented by other mechanisms such as majority rule, command systems, and first-come, first-served, do anamazingly good job.

## Is the Competitive Market Fair?

When a natural disaster strikes, such as a severe winter storm or a hurricane, the prices of many essential items jump. The reason prices jump is that thedemand and willingness to pay for these items hasincreased, but the supply has not changed. So thehigher prices achieve an efficient allocation of scarce resources. News reports of these price hikes almost never talk about efficiency. Instead, they talk aboutequity or fairness. The claim that is often made isthat it is unfair for profit-seeking dealers to cheat the victims of natural disaster. Similarly, when low-skilled people work for a wage that is below what most would regard as a "living wage," the media and politicians talk of employers taking unfair advantage of their workers. How do we decide whether something is fair or unfair? You know when you think something is unfair, but how do you know? What are the principles of fairness?
Philosophers have tried for centuries to answer this question. Economists have offered their answers too. But before we look at the proposed answers, you should know that there is no universally agreed upon answer.
Economists agree about efficiency. That is, they agree that it makes sense to make the economic pie as large as possible and to produce it at the lowest possible cost. But they do not agree about equity. That is,they do not agree about what are fair shares of theeconomic pie for all the people who make it. The reason is that ideas about fairness are not exclusively economic ideas. They touch on politics, ethics, andreligion. Nevertheless, economists have thought aboutthese issues and have a contribution to make. Let'sexamine the views of economists on this topic.
To think about fairness, think of economic life as a game - a serious game. All ideas about fairness can be divided into two broad groups. They are

- It's not fair if the result isn't fair.
- It's not fair if the rules aren't fair.


## It's Not Fair if the Result is Not Fair

The earliest efforts to establish a principle of fairness were based on the view that the result is what matters. The general idea was that it is unfair if people's incomes are too unequal. For example, it is unfair that a bank president earns millions of dollars a year while a bank teller earns only thousands of dollars. It is unfair that a store owner makes a larger profit and her customers pay higher prices in the aftermath of a winter storm.
During the nineteenth century, economists thought they had made an incredible discovery: Efficiency requires equality of incomes. To make the economic pie as large as possible, it must be cut into equal pieces, one for each person. This idea turns out to be wrong. But there is a lesson in the reason that it is wrong, so this idea is worth a closer look.

Utilitarianism The nineteenth-century idea that only equality brings efficiency is called utilitarian- ism. Utilitarianism is a principle that states that we should strive to achieve "the greatest happiness forth greatest number." The people who developed this idea were known as utilitarians. They includedthe most eminent thinkers, such as Jeremy Benthamand John Stuart Mill.

Utilitarians argued that to achieve "the greatest happiness for the greatest number," income must be transferred from the rich to the poor up to the point of complete equality-to the point at which there areno rich and no poor.
They reasoned in the following way: First, everyone has the same basic wants and a similar capacity toenjoy life. Second, the greater a person's income, thesmaller is the marginal benefit of a dollar. The millionth dollar spent by a rich person brings a smallermarginal benefit to that person than the marginalbenefit that the thousandth dollar spent brings to a poorer person. So by transferring a dollar from the millionaire to the poorer person, more is gained than is lost. The two people added together are better off Figure 5.7 illustrates this utilitarian idea. Tom and Jerry have the same marginal benefit curve, MB. (Marginal benefit is measured on the same scale of 1 to 3 for both Tom and Jerry.) Tom is at point $A$. He earns $\$ 5,000$ a year, and his marginal benefit from a dollar is 3 units. Jerry is at point B. He earns $\$ 45,000$ a year, and his marginal benefit from a dollar is 1 unit. If a dollar is transferred from Jerry to Tom, Jerry loses 1 unit of marginal benefit and Tom gains 3 units. So together, Tom and Jerry are better off - they are sharing the economic pie more efficiently. If a second dollar is transferred, the same thing happens:
Tom gains more than Jerry loses. And the same istrue for every dollar transferred until they both reach point C. At point C, Tom and Jerry have $\$ 25,000$ each and a marginal benefit of 2 units. Now they are sharing the economic pie in the most efficient way. It brings the greatest happiness to Tom and Jerry.

The Big Tradeoff One big problem with the utilitarian ideal of complete equality is that it ignores the costs of making income transfers. Recognizing the costs of making income transfers leads to what is called the big tradeoff, which is a tradeoff between efficiency and fairness. The big tradeoff is based on the following facts. Income can be transferred from people with high incomes to people with low incomes only by taxing the high incomes. Taxing people's income from employment makes them work less. It results in thequantity of labor being less than the efficient quantity. Taxing people's income from capital makes them save less. It results in the quantity of capital being less than the efficient quantity. With smaller quantities ofboth labor and capital, the quantity of goods and services produced is less than the efficient quantity. The economic pie shrinks.

The tradeoff is between the size of the economic pie and the degree of equality with which it is shared. The greater the amount of income redistribution through income taxes, the greater is the inefficiency - the smaller is the economic pie.
There is a second source of inefficiency. A dollar taken from a rich person does not end up as a dollar in the hands of a poorer person. Some of the dollar isspent on administration of the tax and transfer system. The cost of tax-collecting agencies, such as the Internal Revenue Service (IRS), and welfare-administering agencies, such as the Centers for Medicare \& Medicaid Services (CMS), must be paid with some of the taxes collected. Also, taxpayers hire accountants, auditors, and lawyers to help them ensure that they pay the correct amount of taxes. These activities use skilled labor and capital resources that could otherwise be used to produce goods and services that people value.

FIGURE 5.7 Utilitarian Fairness


Tom earns $\$ 5,000$ and has 3 units of marginal benefit at point $A$. Jerry earns $\$ 45,000$ and has 1 unit of marginal benefit at point $B$. If income is transferred from Jerry to Tom, Jerry's loss is less than Tom's gain. Only when each of them has $\$ 25,000$ and 2 units of marginal benefit (at point $C$ ) can the sum of their total benefit increase no further.
When all these costs are taken into account, taking a dollar from a rich person does not give a dollar to a poor person. It is possible that with high taxes, people with low incomes might end up being worse off. Suppose, for example, that highly taxed entrepreneurs decide to work less hard and shut down some of their businesses. Low-income workers get fired and must seek other, perhaps even lower-paid, work. Today, because of the big tradeoff, no one says that fairness requires complete equality of incomes.

Make the Poorest as Well Off as Possible A new solution to the big-tradeoff problem was proposed byphilosopher John Rawls in a classic book entitled ATheory of justice, published in 1971. Rawls says that, taking all the costs of income transfers into account, the fair distribution of the economic pie is the one that makes the poorest person as well off as possible.The incomes of rich people should be taxed, and afterpaying the costs of administering the tax and transfersystem, what is left should be transferred to the poor.But the taxes must not be so high that they make theeconomic pie shrink to the point at which the poorestperson ends up with a smaller piece. A bigger share of a smaller pie can be less than a smaller share of a bigger pie. The goal is to make the piece enjoyed by the poorest person as big as possible. Most likely, this piece will not be an equal share.
The "fair results" idea requires a change in the results after the game is over. Some economists say that these changes are themselves unfair and propose a different way of thinking about fairness.

## It's Not fair if the Rules Aren't Fair

The idea that it's not fair if the rules aren't fair is based on a fundamental principle that seems to be hardwired into the human brain: the symmetry principle. The symmetry principle is the requirement that people in similar situations be treated similarly. It is the moral principle that lies at the center of all thebig religions and that says, in some form or other, "Behave toward other people in the way you expect them to behave toward you."
In economic life, this principle translates into equality of opportunity. But equality of opportunity to do what? This question is answered by the philosopher Robert Nozick in a book entitled Anarchy, State, and Utopia, published in 1974.
Nozick argues that the idea of fairness as an out- come or result cannot work and that fairness must be based on the fairness of the rules. He suggests thatfairness obeys two rules:

1. The state must enforce laws that establish and protect private property.
2. Private property may be transferred from one person to another only by voluntary exchange.

The first rule says that everything that is valuable must be owned by individuals and that the state must ensure that theft is prevented. The second rule says that the only legitimate way a person can acquire property is to buy it in exchange for some- thing else that the person owns. If these rules, which are the only fair rules, are followed, then the result is fair. It doesn't matter how unequally the economic pie is shared, provided that the pie is made by people, each one of whom voluntarily provides services in exchange for the share of the pie offered in compensation.
These rules satisfy the symmetry principle. If these rules are not followed; the symmetry principle is broken. You can see these facts by imagining a world inwhich the laws are not followed.
First, suppose that some resources or goods are not owned. They are common property. Theneveryone is free to participate in a grab to use them. The strongest will prevail. But when the strongest prevails, the strongest effectively owns theresources or goods in question and prevents others from enjoying them.
Second, suppose that we do not insist on voluntary exchange for transferring ownership of resources from one person to another. The alternative is involuntarytransfer. In simple language, the alternative is theft.
Both of these situations violate the symmetry principle. Only the strong acquire what they want. Theweak end up with only the resources and goods that the strong don't want.
In a majority-rule political system, the strong are those in the majority or those with enough resources to influence opinion and achieve a majority.
In contrast, if the two rules of fairness are followed, everyone, strong and weak, is treated in a similar way. All individuals are free to use their resources and human skills to create things that arevalued by themselves and others and to exchange the fruits of their efforts with all others. This set of arrangements is the only one that obeys the symmetry principle.

Fair Rules and Efficiency If private property rights are enforced and if voluntary exchange takes place in a competitive market with none of the obstacles described above (p. 152), resources will be allocatedefficiently.
According to the Nozick fair-rules view, no matter how unequal is the resulting distribution of income and wealth, it will be fair.
It would be better if everyone were as well offas those with the highest incomes, but scarcity prevents that outcome and the best attainable outcome is the efficient one.

Hurricane Katrina shut down electricity supplies over a wide area and increased the demand for portable generators. What is the fair way to allocate the available generators?
If the market price is used, the outcome is efficient. Sellers and buyers are better off and no one is worse off. But people who own generators make alarger profit and the generators go to those who wantthem most and can afford them. Is that fair? On the Nozick rules view, the outcome is fair. On the fair outcome view, the outcome might be considered unfair. But what are the alternatives? They are command; majority rule; contest; first-come, first-served; lottery; personal characteristics; and force. Except by chance, none of these methods delivers an allocation of generators that is either fair or efficient. It is unfair in the rules view because the distribution involves involuntary transfers of resources among citizens. It is unfair in the results view because the poorest don't end up being made as well off as possible.

## ECONOMIC ANAIYSIS

- The discovery that adding a highway does not ease congestion points to congestion pricing as a solution. Let's see how a congestion price works.
- A highway has the marginal social cost curve MSC in the figures. The highway can carry only 10,000 vehicles per hour with no congestion.
- At off-peak times, the demand curve and marginal benefit curve is $D_{0}=M S B_{0}$; and the peak time demand curve and marginal benefit curve is $D_{p}=M S B_{p}$.
- Figure 1 illustrates inefficient road use. At off-peak, the outcome is efficient but at the peak demand time, at a zero price, 40,000 vehicles per hour enter the road. The marginal social cost is $\$ 6$ per vehicle-hour and there is a deadweight loss (of time and gasoline) shown by the gray triangle.
- Figure 2 illustrates efficient road use at the peak period. Imposing a congestion charge of $\$ 3$ per vehicle-hour brings an equilibrium at 25,000 vehicles per hour, which is the efficient quantity. Total surplus, the sum of consumer surplus (green) plus producer surplus (blue), is maximized.
- Congestion charges would be paid by all road users, regardless of whether they are rich or poor, but they don't have to leave the poor worse off.
- Revenue raised from congestion charges can be redistributed to low-income households if there is a fairness problem.
- So long as road users pay the marginal social cost of their decision, road use is efficient.


Figure 2 Efficient Rush-Hour Road Use

- London has a simple system of congestion pricing that imposes a charge when a vehicle enters a central congestion zone.
- Singapore has the world's most sophisticated congestion pricing with the price displayed on gantries (see photo), and the price rises as congestion increases and falls as congestion eases.
- Advances in technology make congestion pricing an attractive alternative to congestion.


Figure 1 Inefficient Rush-Hour Road Use

## CHAPTER 6: Government actions in markets

After studying this chapter, you will be able to:

- Explain how a rent ceiling creates a housing shortage
- Explain how a minimum wage law creates unemployment
- Explain the effects of a tax
- Explain the effects of production quotas and subsidies
- Explain how markets for illegal goods work

In New York City, where food servers and grocery clerks earn the minimum wage of $\$ 7.25$ an hour, a budget one-bedroom apartment rents for $\$ 1,000$ a month. For the lowest paid, that leaves $\$ 160$ a month for food, clothing, and other necessities. What can governments do to help these people?
This chapter explains the effects of minimum wage laws and rent ceilings, and Economics in the News at the end of the chapter looks at some recent changes in state minimum wages. The chapter also explains the effects of taxes, production quotas and subsidies, and laws that make trading in some things illegal.

## A Housing Market with a Rent Ceiling

We spend more of our income on housing than on any other good or service, so it isn't surprising that rents can be a political issue. When rents are high, or when they jump by a large amount, renters might lobby the government for limits on rents.
A government regulation that makes it illegal to charge a price higher than a specified level is called a price ceiling or price cap.
The effects of a price ceiling on a market depend crucially on whether the ceiling is imposed at a level that is above or below the equilibrium price. A price ceiling set above the equilibrium price has no effect. The reason is that the price ceiling does not constrain the market forces. The force of the law and the market forces are not in conflict. But a price ceiling below the equilibrium price has powerful effects on a market. The reason is that the priceceiling attempts to prevent the price from regulating the quantities demanded and supplied. The force of the law and the market forces are inconflict.
When a price ceiling is applied to a housing market, it is called a rent ceiling. A rent ceiling set below the equilibrium rent creates

- A housing shortage
- Increased search activity
- A black market


## A Housing Shortage

At the equilibrium price, the quantity demanded equals the quantity supplied. In a housing market, when the rent is at the equilibrium level, the quantity of housing supplied equals the quantity of housing demanded and there is neither a shortage nor a surplus of housing.

But at a rent set below the equilibrium rent, the quantity of housing demanded exceeds the quantity of housing supplied - there is a shortage. So if a rentceiling is set below the equilibrium rent, there will be a shortage of housing.
When there is a shortage, the quantity available is the quantity supplied, and somehow this quantity must be allocated among the frustrated demanders. One way in which this allocation occurs is through increased search activity.

## Increased Search Activity

The time spent looking for someone with whom to do business is called search activity. In some markets, such as the housing market, people spend a lot of time checking the alternatives available before making a choice. When a price is regulated and there is a shortage, search activity increases. In the case of a rent-controlled housing market, frustrated would-be renters scan the newspapers, not only for housing ads but also for death notices! Any information about newly available housing is useful, and apartment seekers race to be first on the scene when news of a possible supplier breaks. The opportunity cost of a good is equal not only to its price but also to the value of the search time spent finding the good. So the opportunity cost of housing is equal to the rent (a regulated price) plus the time and other resources spent searching for the restricted quantity available. Search activity is costly. It uses time and other resources, such as phone calls, automobiles, and gasoline that could have been used in other productive ways. A rent ceiling controls only the rent portion of the cost of housing. The cost of increased search activitymight end up making the full cost of housing higherthan it would be without a rent ceiling.

## A Black Market

A rent ceiling also encourages illegal trading in a black market, an illegal market in which the equilibrium price exceeds the price ceiling. Black markets occur in rentcontrolled housing and many other markets. For example, scalpers run black markets in tickets for big sporting events and rock concerts.
The level of a black market rent depends on how tightly the rent ceiling is enforced. With loose enforcement, the black market rent is close to the unregulated rent. But with strict enforcement, the black market rent is equal to the maximum price that a renter is willing to pay.
Figure 6.1 illustrates the effects of a rent ceiling. The demand curve for housing is $D$ and the supply curve is S . A rent ceiling is imposed at $\$ 800$ a month. Rents that exceed $\$ 800$ a month are in the gray-shaded illegal region in the figure. You can see that the equilibrium rent, where the demand and supply curves intersect, is in the illegal region.
At a rent of $\$ 800$ a month, the quantity of housing supplied is 60,000 units and the quantity demanded is 100,000 units. So with a rent of $\$ 800$ a month, there is a shortage of 40,000 units of housing.
To rent the 60,000th unit, someone is willing to pay $\$ 1,200$ a month. They might pay this amount by incurring search costs that bring the total cost of housing to \$1,200 a month, or they might pay a black market price of $\$ 1,200$ a month. Either way, they end up incurring a cost that exceeds what the equilibrium rent would be in an unregulated market.

## FIGURE 6.1 A Rent Ceiling



A rent above the rent ceiling of $\$ 800$ a month is illegal. At a rent of $\$ 800$ a month,the quantity of housing supplied is 60,000 units. Frustratedrenters spend time searching for housing and they make deals with landlords in a black market. Someone is willingto pay \$1,200 a month for the 60,000th unit.

## Inefficiency of a Rent Ceiling

A rent ceiling set below the equilibrium rent results in an inefficient underproduction of housing services. The marginal social benefit of housing exceeds its marginal social costand a deadweight loss shrinks the producer surplus and consumer surplus (Chapter 5, pp. 150152).

Figure 6.2 shows this inefficiency. The rent ceiling ( $\$ 800$ per month) is below the equilibrium rent ( $\$ 1,000$ per month) and the quantity of housing supplied ( 60,000 units) is less than the efficient quantity ( 80,000 units).
Because the quantity of housing supplied (the quantity available) is less than the efficient quantity, there is a deadweight loss, shown by the gray triangle. Producer surplus shrinks to the blue triangle and consumer surplus shrinks to the green triangle. The rectangle represents the potential loss from increased search activity. This loss is borne by consumers and the full loss from the rent ceiling is the sum of the deadweight lossand the increased cost of search.

FIGURE 6.2 The Inefficiency of a Rent Ceiling


Without a rent ceiling, the market produces an efficient 80,000 units of housing at a rent of $\$ 1,000$ a month. A rent ceiling of $\$ 800$ a month decreases the quantity of housing supplied to 60,000 units. Producer surplus and consumer surplus shrink and a deadweight loss arises. The red rectangle represents the cost of resources used in increased search activity. The full loss from the rent ceiling equals the sum of the shaded rectangle and triangle.

## Are Rent Ceilings Fair?

Rent ceilings might be inefficient, but don't they achieve a fairer allocation of scarce housing? Let's explore this question.
Chapter 5 (pp. 154-156) reviews two key ideas about fairness. According to the fair-rules view, anything that blocks voluntary exchange is unfair, so rent ceilings are unfair. But according to the fair-result view, a fair outcome is one that benefits the less well off So according to this view, the fairest outcome is the one that allocates scarce housing to the poorest. To see whether rent ceilings help to achieve a fairer outcome in this sense, we need to consider how the market allocates scarce housing resources in the face of a rent ceiling. Blocking rent adjustments doesn't eliminate scarcity. Rather, because it decreases the quantity of housing available, it creates an even bigger challenge for the housing market. Somehow, the market must ration a smaller quantity of housing and allocate that housing among the people who demand it.
When the rent is not permitted to allocate scarce housing, what other mechanisms are available, andare they fair? Some possible mechanisms are:

- A lottery
- First-come, first-served
- Discrimination

A lottery allocates housing to those who are lucky, not to those who are poor. First-come, first- served (a method used to allocate housing in England after World War II) allocates
housing tothose who have the greatest foresight and who get their names on a list first, not to the poorest. Discrimination allocates scarce housing based on the views and selfinterest of the owner of the housing. In the case of public housing, what counts is the self-interest of the bureaucracy that administers the allocation.
In principle, self-interested owners and bureaucrats could allocate housing to satisfy some criterion of fairness, but they are not likely to do so. Discrimination based on friendship, family ties, and criteria such as race, ethnicity, or sex is more likely to enter the equation. We might make such discrimination illegal, but we cannot prevent it from occurring. It is hard, then, to make a case for rent ceilings on the basis of fairness. When rent adjustments are blocked, other methods of allocating scarce housing resources operate that do not produce a fair outcome.

## ECONOMICS IN ACTION

Rent Control Winners: The Rich and Famous
New York, San Francisco, London, and Paris, four ofthe world's great cities, have rent ceilings in some part of their housing markets. Boston had rent ceilings formany years but abolished them in 1997. Many other U.S. cities do not have, and have never had, rent ceilings. Among them are Atlanta, Baltimore, Chicago, Dallas, Philadelphia, Phoenix, and Seattle.
To see the effects of rent ceilings in practice we can compare the housing markets in cities with ceilings with those without ceilings. We learn two main les-sons from such a comparison. First, rent ceilings definitely create a housing short- age. Second, they do lower the rents for some but raise them for others. A survey conducted in 1997 showed that the rents of housing units actually available for rent were 2.5 times the average of all rents in New York but equal to the average rent in Philadelphia. The winners from rent ceilings are the families that have lived in a city for a long time. In New York, these families include some rich and famous ones. The voting power of the winners keeps the rent ceilings in place. Mobile newcomers are the losers in a city with rent ceilings. The bottom line is that, in principle and in practice, rent ceilings are inefficient and unfair.

## A Labor Market with a Minimum Wage

For each one of us, the labor market is the market that influences the jobs we get and the wages we earn. Firms decide how much labor to demand, and the lower the wage rate, the greater is the quantity of labor demanded. Households decide how much labor to supply, and the higher the wage rate, the greater is the quantity of labor supplied. The wage rate adjusts to make the quantity of labor demanded equal to thequantity supplied.
When wage rates are low, or when they fail to keep up with rising prices, labor unions might turn to governments and lobby for a higher wage rate.
A government regulation that makes it illegal to charge a price lower than a specified level is called a price floor.
The effects of a price floor on a market depend crucially on whether the floor is imposed at a level that is above or below the equilibrium price.
A price floor set below the equilibrium price has no effect. The reason is that the price floor does not constrain the market forces. The force of the law andthe market forces are not in conflict. But a price floor set above the equilibrium price has powerful effects on a market. The reason is
that the price floor attempts to prevent the price from regulating the quantities demanded and supplied. The force of the law and the market forces are in conflict.
When a price floor is applied to a labor market, it is called a minimum wage. A minimum wage imposed at a level that is above the equilibrium wage createsunemployment. Let's look at the effects of aminimum wage.

## Minimum Wage Brings Unemployment

At the equilibrium price, the quantity demanded equals the quantity supplied. In a labor market, when the wage rate is at the equilibrium level, the quantity of labor supplied equals the quantity of labordemanded: There is neither a shortage of labor nor a surplus of labor.
But at a wage rate above the equilibrium wage, the quantity of labor supplied exceeds the quantity of labor demanded-there is a surplus of labor. So when a minimum wage is set above the equilibrium wage, there is a surplus of labor. The demand for labor determines the level of employment, and the surplus of labor is unemployed.

## FIGURE 6.3 Minimum Wage and Unemployment



The minimum wage rate is set at $\$ 7$ an hour. Any wage rate below $\$ 7$ an hour is illegal. At the minimum wage of $\$ 7$ an hour, 20 million hours are hired but 22 million hours are available. Unemployment-AB-of 2 million hours a year is created. With only 20 million hours demanded, someone is willing to supply the 20 millionth hour for $\$ 5$.

Figure 6.3 illustrates the effect of the minimum wage on unemployment. The demand for labor curve is $D$ and the supply of labor curve is $S$. The horizontal line shows the minimum wage set at $\$ 7$ an hour. A wage rate below this level is illegal, in the illegal region of the figure. At the minimum wage rate, 20 million hours of labor are demanded (point $A$ ) and 22 million hours of labor are supplied (point $B$ ), so 2 million hours of available labor are unemployed. With only 20
million hours demanded, someone is willing to supply that 20 millionth hour for $\$ 5$. Frustrated unemployed workers spend time and other resources searching for hard-to-find jobs.

## Is the Minimum Wage Fair?

The minimum wage is unfair on both views of fairness: It delivers an unfair result and imposes an unfair rule.
The result is unfair because only those people who have jobs and keep them benefit from the minimum wage. The unemployed end up worse off than they would be with no minimum wage. Some of those who search for jobs and find them end up worse off because of the increased cost of job search they incur. Also those who search and find jobs aren't always the least well off. When the wage rate doesn't allocate labor, other mechanisms determine who finds a job. One such mechanism is discrimination, which is yet another source of unfairness.
The minimum wage imposes an unfair rule because it blocks voluntary exchange. Firms are willing to hire more labor and people are willing to workmore, but they are not permitted by the minimum wage law to do so.

## Inefficiency of a Minimum Wage

In the labor market, the supply curve measures the marginal social cost of labor to workers. This cost is leisure forgone. The demand curve measures the marginal social benefit from labor. This benefit is the value of the goods and services produced. An unregulated labor market allocates the economy's scarce labor resources to the jobs in which they are valued most highly. The market is efficient.
The minimum wage frustrates the market mechanism and results in unemployment and increased job search. At the quantity of labor employed, the marginal social benefit of labor exceeds its marginal social cost and a deadweight loss shrinks the firms' surplus and the workers' surplus.
Figure 6.4 shows this inefficiency. The minimum wage ( $\$ 7$ an hour) is above the equilibrium wage ( $\$ 6$ an hour) and the quantity of labor demanded and employed (20 million hours) is less than the efficient quantity ( 21 million hours).
Because the quantity of labor employed is less than the efficient quantity, there is a deadweight loss, shown by the gray triangle. The firms' surplus shrinks to theblue triangle and the workers' surplus shrinks to thegreen triangle. The red rectangle shows the potential loss from increased job search, which is borne by workers. The full loss from the minimum wage is the sum of the deadweight loss and the increased cost of job search.

## FIGURE 6.4 The Inefficiency of a Minimum Wage



A minimum wage decreases employment. Firms' surplus and workers' surplus shrink and a deadweight loss arises. Job search increases and the red area shows the loss from this activity.

## Taxes

Everything you earn and almost everything you buy is taxed. Income taxes and Social Security taxes are deducted from your earnings and sales taxes areadded to the bill when you buy something. Employers also pay a Social Security tax for their workers, and producers of tobacco products, alcoholic drinks, and gasoline pay a tax every time they sell something. Who really pays these taxes? Because the income tax and Social Security tax are deducted from yourpay, and the sales tax is added to the prices that you pay, isn't it obvious that you pay these taxes? And isn't it equally obvious that your employer pays the employer's contribution to the Social Security tax and that tobacco producers pay the tax on cigarettes? You're going to discover that it isn't obvious who really pays a tax and that lawmakers don't make that decision. We begin with a definition of tax incidence.

Tax Incidence is the division of the burden of a tax between buyers and sellers. When the government imposes a tax on the sale of a good,* the price paid by buyers might rise by the full amount of the tax, by a lesser amount, or not at all. If the price paid by buyers rises by the full amount of the tax, then the burden of the tax falls entirely on buyers - the buyers pay the tax. If the price paid by buyers rises by a lesser amount than the tax, then the burden of the tax falls partly on buyers and partly on sellers. And if the price paid by buyers doesn't change at all, then the burden of the tax falls entirely on sellers. Tax incidence does not depend on the tax law. The law might impose a tax on sellers or on buyers, but the outcome is the same in either case. To see why, let's look at the tax on cigarettes in New York City.

## A Tax on Sellers

On July I, 2002, Mayor Bloomberg put a tax of \$1.50 a pack on cigarettes sold in New York City. To work out the effects of this tax on the sellers of cigarettes, we begin by examining the effects on demand and supply in the market for cigarettes.
In Fig. 6.5, the demand curve is $D$, and the supply curve is $S$. With no tax, the equilibrium price is $\$ 3$ perpack and 350 million packs a year are bought and sold.
A tax on sellers is like an increase in cost, so it decreases supply. To determine the position of the new supply curve, we add the tax to the minimum quantity sold. You can see that without the tax, sellers are willing to offer 350 million packs a yearfor $\$ 3$ a pack. So with a $\$ 1.50$ tax, they will offer350 million packs a year only if the price is $\$ 4.50$ a pack. The supply curve shifts to the curve labeled $S+$ tax on sellers.

Equilibrium occurs where the new supply curve intersects the demand curve at 325 million packs a year. The price paid by buyers rises by $\$ 1$ to $\$ 4$ a pack. And the price received by sellers falls by 50 ct to $\$ 2.50$ a pack. So buyers pay $\$ 1$ of the tax and sellers pay the other $50 \%$.

FGuse 6.5 A Tax on Sellers


With no tax, 350 million packs a year are bought and sold at $\$ 3$ a pack. A tax on sellers of $\$ 1.50$ a pack shifts the supply curve from $S$ to $S+$ tox on sellers. The equilibrium quantity decreases to 325 million packs a year, the price paid by buyers rises to $\$ 4$ a pack, and the price received by sellers falls to $\$ 2.50 \mathrm{a}$ pack. The tax raises the price paid by buyers by less than the tax and lowers the price received by sellers, so buyers and sellers share the burden of the tax.

## A Tax on Buyers

Suppose that instead of taxing sellers, New York City taxes cigarette buyers $\$ 1.50$ a pack. A tax on buyers lowers the amount they are willing to pay sellers, so it decreases demand and shifts the demand curve leftward. To determine the position of this new demand curve, we subtract the tax from the maximum price that buyers are willing to pay for each quantity bought. You can see, in Fig. 6.6, that without the tax, buyers are willing to buy 350 million packs a year for $\$ 3$ a pack. So with a $\$ 1.50$ tax, they are willing to buy 350 million packs a year only if the price including the tax is $\$ 3$ a pack, which means that they're willing to pay sellers only $\$ 1.50$ a pack. The demand curve shifts to become the red curve labeled $D$ tax on buyers.
Equilibrium occurs where the new demand curve intersects the supply curve at a quantity of 325 million packs a year. The price received by sellers is $\$ 2.50$ a pack, and the price paid by buyers is $\$ 4$.

Can We Share the Burden Equally? Suppose that Mayor Bloomberg wants the burden of the cigarettetax to fall equally on buyers and sellers and declaresthat a $75 \%$ tax be imposed on each. Is the burden of the tax then shared equally?
You can see that it is not. The tax is still $\$ 1.50$ a pack. You've seen that the tax has the same effect regardless of whether it is imposed on sellers or buyers. So imposing half the tax on sellers and half on buyers is like an average of the two cases you've just examined. (Draw the demand-supply graph and work out what happens in this case. The demand curve shifts downward by 75¢ and the supply curve shifts upward by $75 \%$. The new equilibrium quantity is still 325 million packs a year. Buyers pay $\$ 4$ a pack, of which 75 c is tax. Sellers receive $\$ 3.25$ from buyers, but pay a 75 c tax, so sellers net $\$ 2.50$ a pack.)
When a transaction is taxed, there are two prices: the price paid by buyers, which includes the tax; and the price received by sellers, which excludes the tax.

Buyers respond to the price that includes the tax and sellers respond to the price that excludes the tax. A tax is like a wedge between the price buyers pay and the price sellers receive. The size of the wedge determines the effects of the tax, not the side of the market on which the government imposes the tax.

TheSocialSecurity TaxThe Social Securitytax is an example of a tax that Congress imposes equally on both buyers and sellers. But the principles you've just learned apply to this tax too. The market for labor, not Congress, decides how the burden of the Social Security tax is divided between firms and workers.
In the New York City cigarette tax example, buyers bear twice the burden of the tax borne by sellers. In special cases, either buyers or sellers bear the entire burden. The division of the burden of a tax betweenbuyers and sellers depends on the elasticities of demand and supply, as you will now see.
figure 6.6 A Tax on Buyers


With no tax, 350 million packs a year are bought and sold at $\$ 3$ a pack. A tax on buyers of $\$ 1.50$ a pack shifts the demand curve from $D$ to $D$-tax on buyers. The equilibrium quantity decreases to 325 million packs a year, the price paid by buyers rises to $\$ 4$ a pack, and the price received by sellers falls to $\$ 2.50$ a pack. The tax raises the price paid by buyers by less than the tax and lowers the price received by sellers, so buyers and sellers share the burden of the tax.

Tax Incidence and Elasticity of Demand The division of the tax between buyers and sellers depends in part on the elasticity of demand. There are two extreme cases:

- Perfectly inelastic demand-buyers pay.
- Perfectly elastic demand-sellers pay.

Perfectly Inelastic Demand Figure 6.7 showsthe market for insulin, a vital daily medication for those with .diabetes. Demand is perfectly inelastic at100,000 doses a day, regardless of the price, as shownby the vertical demand curve $D$. That is, a diabetic would sacrifice all other goods and services ratherthan not consume the insulin dose that provides good health. The supply curve of insulin is S. With no tax, the price is $\$ 2$ a dose and the quantity is 100,000 doses aday.
If insulin is taxed at $20 \%$ a dose, we must add the tax to the minimum price at which drug companies are willing to sell insulin. The result is the new supply curve $\mathrm{S}+$ tax. The price rises to $\$ 2.20$ a dose, but the quantity does not change. Buyers pay the entire tax of 20¢ a dose.

FIGURE 6.7 Tax with Perfectly Inelastic Demand


In this market for insulin, demand is perfectly inelastic. With no tax, the price is $\$ 2$ a dose and the quantity is 100,000 doses a day. A tax of $20 \$$ a dose shifts the supply curve to $S+$ tax. The price rises to $\$ 2.20$ a dose, but the quantity bought does not change. Buyers pay the entire tax.

Perfectly Elastic Demand Figure 6.8 shows the market for pink marker pens. Demand is perfectly elasticat $\$ 1$ a pen, as shown by the horizontal demand curve $D$. If pink pens are less expensive than the other colors, everyone uses pink. If pink pens are more expensive than other colors, no one uses pink.
The supply curve is $S$. With no tax, the price of a pink pen is $\$ 1$ and the quantity is 4,000 pens a week. Suppose that the government imposes a tax of 10¢ a pen on pink marker pens but not on other colors. The new supply curve is $S+$ tax. The price remains at $\$ 1$ a pen, and the quantity decreases to 1,000 pink pens a week. The $10 ¢$ tax leaves the price paid by buyers unchanged but lowers the amount received by sellers by the full amount of the tax. Sellers pay the entire tax of $10 ¢$ a pink pen.
We've seen that when demand is perfectly inelastic, buyers pay the entire tax and when demand is perfectly elastic, sellers pay the entire tax. In theusual case, demand is neither perfectly inelastic norperfectly elastic and the tax is split between buyers and sellers. But the division depends on the elasticityof demand: The more inelastic the demand, thelarger is the amount of the tax paid by buyers.

## Tax Incidence and Elasticity of Supply

The division of the tax between buyers and sellers also depends, in part, on the elasticity of supply. Again, there are two extreme cases:

- Perfectly inelastic supply - sellers pay.
- Perfectly elastic supply - buyers pay.

FIGURE 6.8 Tax with Perfectly Elastic Demand


In this market for pink pens, demand is perfectly elastic. With no tax, the price of a pen is $\$ 1$ and the quantity is 4,000 pens a week. A tax of $10 \$$ a pink pen shifts the supply curve to $S+$ tax. The price remains at $\$ 1$ a pen, and the quantity of pink pens sold decreases to 1,000 a week.
Sellers pay the entire tax.

Perfectly Inelastic Supply Figure6.9(a) shows the market for water from a mineral spring that flows at a constant rate that can't be controlled. Supply is perfectly inelastic at 100,000 bottles a week, as shown by the supply curve $S$. The demand curve for the water from this spring is $D$. With no tax, the price is $50 \%$ a bottle and the quantity is 100,000 bottles. Suppose this spring water is taxed at $S ¢$ a bottle. The supply curve does not change because the springowners still produce 100,000 bottles a week, eventhough the price they receive falls. But buyers arewilling to buy the 100,000 bottles only if the price is 50 C a bottle, so the price remains at $50 \%$ a bottle. The tax reduces the price received by sellers to $45 ¢$ a bottle, and sellers pay the entire tax.

We've seen that when supply is perfectly inelastic, sellers pay the entire tax, and when supply is perfectly elastic, buyers pay the entire tax. In the usual case,supply is neither perfectly inelastic nor perfectly elastic and the tax is split between buyers and sellers. But how the tax is split depends on the elasticity of supply: The more elastic the supply, the larger is the amount of the tax paid by buyers.

FIGURE 6.9 Tax and the Elasticity of Supply

(a) Perfectly inelastic supply


## (b) Perfectly elastic supply

Part (a) shows the market for water from a mineral spring. Supply is perfectly inelastic. With no tax, the price is $50 \$$ a bottle. With a tax of $5 \$$ a bottle, the price remains at $50 \$$ a bottle. The number of bottles bought remains the same, but the price received by sellers decreases to $45 \ddagger$ a bottle. Sellers pay the entire tax.

Part (b) shows the market for sand. Supply is perfectly elastic. With no tax, the price is $10 \$$ a pound. A tax of $1 \$$ a pound increases the minimum supply-price to $11 \$$ a pound. The supply curve shifts to $S+$ tax. The price increases to 11 a pound. Buyers pay the entire tax.

## Taxes and Efficiency

A tax drives a wedge between the buying price and the selling price and results in inefficient underproduction. The price buyers pay is also the buyers' willingness to pay, which measures marginalsocialbenefit. The price sellers receive is also the sellers' minimum supply-price, which equals marginal social cost.
A tax makes marginal social benefit exceed marginal social cost, shrinks the producer surplus and consumer surplus, and creates a deadweight loss.
Figure 6.10 shows the inefficiency of a tax on MP3 players. The demand curve, $D$, shows marginalsocial benefit, and the supply curve, $S$, shows marginal social cost. Without a tax, the market producesthe efficient quantity (5,000 players a week).
With a tax, the sellers' minimum supply-price rises by the amount of the tax and the supply curve shifts to $\mathrm{S}+$ tax. This supply curve does not show marginal social cost. The tax component isn't a social cost of production. It is a transfer of resources to the government. At the new equilibrium quantity ( 4,000 players a week), both consumer surplus and producer surplus shrink. Part of each surplus goes to the government in tax revenue-the purple area; part becomes a deadweight loss.

FIGURE 6.10 Taxes and Efficiency


With no tax, 5,000 players $a$ week are produced. With a $\$ 20$ tax, the buyers' price rises to $\$ 210$, the sellers' price falls to $\$ 190$, and the quantity decreases to 4,000 players a week. Consumer surplus shrinks, and the producer surplus shrinks. Part of the loss of consumer surplus and producer surplus goes to the government as tax revenue and part becomes a deadweight loss.

Only in the extreme cases of perfectly inelastic demand and perfectly inelastic supply does a tax not change the quantity bought and sold so that no dead-weight loss arises.

## Taxes and Fairness

We've examined the incidence and the efficiency of taxes. Bur when political leaders debate tax issues, it is fairness, not incidence and efficiency, that gets the most attention. Democrats complain that Republican tax cuts are unfair because they give the benefits of lower taxes to the rich. Republicans counter that it isfair that the rich get most of the tax cuts because they pay most of the taxes. No easy answers are available to the questions about the fairness of taxes.
Economists have proposed two conflicting principles of fairness to apply to a tax system:

- The benefits principle
- The ability-to-pay principle

The Benefits Principle is the proposition that people should pay taxes equal to the benefits they receive from the services provided by government. This arrangement is fair because it means that those who benefit most pay the most taxes. It makes tax payments and the consumption of government-provided services similar to private consumption expenditures.
The benefits principle can justify high fuel taxes to pay for freeways, high taxes on alcoholic beverages and tobacco products to pay for public healthcare services, and high rates of income tax on high incomes to pay for the benefits from law and order and from living in a secure environment, from which the rich might benefit more than the poor.

The Ability-to-Pay Principle is the proposition that people should pay taxes according to how easily they can bear the burden ofthe tax. A rich person can more easily bear the bur-den than a poor person can, so the ability-to-pay principle can reinforce the benefits principle to justify high rates of income tax on high incomes.

## Production Quotas and Subsidies

An early or late frost, a hot dry summer, and a wet spring present just a few of the challenges that fill the lives of farmers with uncertainty and sometimes with economic hardship.
Fluctuations in the weather bring fluctuations in farm output and prices and sometimes leave farmers with low incomes. To help farmers avoid low prices and low incomes, governments intervene in the markets for farm products.
Price floors that work a bit like the minimum wage that you've already studied might be used. But as you've seen, this type of government action creates a surplus and is inefficient. These same conclusions apply to the effects of a price floor for farmproducts.
Governments often use two other methods of
intervention in the markets for farm products:

- Production quotas
- Subsidies


## Production Quotas

In the markets for sugarbeets, tobacco leaf, and cotton (among others), governments have, from time to time, imposed production quotas. A production quota is an upper limit to the quantity of a good that may be produced in a specified period. To discover the effects of a production quota, let's look at what a quota does to the market for sugarbeets. Suppose that the growers of sugarbeets want to limit total production to get a higher price. They persuade the government to introduce a productionquota on sugarbeets.
The effect of the production quota depends on whether it is set below or above the equilibrium quantity. If the government introduced a productionquota above the equilibrium quantity, nothing would change because sugarbeet growers would already be producing less than the quota. But a production quota set below the equilibrium quantity has big effects, which are

- A decrease in supply
- A rise in price
- A decrease in marginal cost
- Inefficient underproduction
- An incentive to cheat and overproduce Figure 6.11 illustrates these effects.

A Decrease in Supply A production quota on sugarbeets decreases the supply of sugarbeets. Each grower is assigned a production limit that is less than the amount that would be produced - and supplied - without the quota. The total of the growers' limits equals the quota, and any production in excess of the quota is illegal.
The quantity supplied becomes the amount permitted by the production quota, and this quantity is fixed. The supply of sugarbeets becomes perfectly inelastic at the quantity permitted under the quota.
In Fig. 6.11, with no quota, growers would produce 60 million tons of sugarbeets a year - the market equilibrium quantity. With a production quota set at 40 million tons a year, the grayshaded area shows the illegal region. As in the case of price ceilings and price floors, market forces and political forces are in conflict in this illegal region. The vertical line labeled "Quota" becomes the supply curve of sugarbeets at prices above $\$ 20$ a ton.

A Rise in Price The production quota raises the price of sugarbeets. When the government sets a production quota, it leaves market forces free to determine the price. Because the quota decreases the supply of sugar- beets, it raises the price. In Fig. 6.11, with no quota, the price is $\$ 30$ a ton. With a quota of 40 million tons, the price rises to $\$ 50$ a ton.

A Decrease in Marginal Cost The production quota lowers the marginal cost of growing sugarbeets. Marginal cost decreases because growers produce less and stop using the resources with the highest marginal cost. Sugarbeet growers slide down their supply (and marginal cost) curves. In Fig. 6.11, marginalcost decreases to $\$ 20$ a ton.

FIGURE 6.11 The Effects of a Production Quota


With no quota, growers produce 60 million tons a year and the price is $\$ 30 \mathrm{a}$ ton. A production quota of 40 mil lion tons a year restricts total production to that amount. The quantity produced decreases to 40 million tons a year, the price rises to $\$ 50$ a ton, and the farmers' marginal cost falls to $\$ 20$ a ton. Because marginal social cost (on the supply curve) is less than marginal social benefit (on the demand curve), a deadweight loss arises from the underproduction.

Inefficiency The production quota results in inefficient underproduction. Marginal social benefit at the quantity produced is equal to the market price, which has increased. Marginal social cost at the quantity produced has decreased and is less than the market price. So marginal social exceeds marginal social cost and a deadweight loss arises.

An Incentive to Cheat and Overproduce The production quota creates an incentive for growers to cheat and produce more than their individual production limit. With the quota, the price exceeds marginal cost, so the grower can get a larger profit by producing one more unit. Of course, if all growers produce more than their assigned limit, the production quota becomes ineffective, and the price falls to the equilibrium (no quota) price.
To make the production quota effective, growers must set up a monitoring system to ensure that no one cheats and overproduces. But it is costly to set upand operate a monitoring system and it is difficult todetect and punish producers who violate their quotas. Because of the difficulty of operating a quota, producers often lobby governments to establish a quota and provide the monitoring and punishment systemsthat make it work.

## Subsidies

In the United States, the producers of peanuts, sugarbeets, milk, wheat, and many other farm products receive subsidies. A subsidy is a payment made by the government to a producer. A large and controversial Farm Bill passed by Congress in 2008 renewed and extended a wide range of subsidies. The effects of a subsidy are similar to the effects of a tax but they go in the opposite directions. These effects are

- An increase in supply
- A fall in price and increase in quantity produced
- An increase in marginal cost
- Payments by government to farmers
- Inefficient overproduction

Figure 6.12 illustrates the effects of a subsidy to peanut farmers.
FIGURE 6.12 The Effects of a Subsidy


With no subsidy, farmers produce 40 million tons a year of $\$ 40 \mathrm{a}$ ton. A subsidy of $\$ 20 \mathrm{a}$ ton shifts the supply curve rightward to $S$ - subsidy. The equilibrium quantity increases to 60 million tons a year, the price falls to $\$ 30$ a ton, and the price plus the subsidy received by farmers rises to $\$ 50$ a ton. In the new equilibrium, marginal social cost (on the supply curve) exceeds marginal social benefit (on the demand curve) and the subsidy results in inefficient overproduction.
An Increase in Supply In Fig. 6.12, with no subsidy, the demand curve $D$ and the supply curve $S$ determine the price of peanuts at $\$ 40$ a ton and the quantity of peanuts at 40 million tons a year.
Suppose that the government introduces a subsidy of $\$ 20$ a ton to peanut farmers. A subsidy is like a negative tax. A tax is equivalent to an increase in cost, so a subsidy is equivalent to a decrease in cost. The subsidy brings an increase in supply.

To determine the position of the new supply curve, we subtract the subsidy from the farmers' minimum supply-price. In Fig. 6.12, with no subsidy, farmers are willing to offer 40 million tons a year at a price of $\$ 40$ a ton. With a subsidy of $\$ 20$ a ton, theywill offer 40 million tons a year if the price is as low as $\$ 20$ a ton. The supply curve shifts to the curvelabeled $S$-subsidy.

A Fall in Price and Increase in Quantity Produced The subsidy lowers the price of peanuts and increases the quantity produced. In Fig. 6.12, equilibrium occurs where the new supply curve intersects the demand curve at a price of $\$ 30$ a ton and a quantity of 60 million tons ayear.

An Increase in Marginal Cost The subsidy lowers the price paid by consumers but increases the marginal cost of producing peanuts. Marginal cost increases because farmers grow more peanuts, which means that they must begin to use someresources that are less ideal for growing peanuts. Peanut farmers slide up their supply (and marginalcost) curves. In Fig. 6.12, marginal cost increases to $\$ 50$ a ton.
Payments by Government to Farmers The government pays a subsidy to peanut farmers on each ton of peanuts produced. In this example, farmers increase production to 60 million tons a year and receive a subsidy of $\$ 20$ a ton. So peanut farmers receive payments from the government that total \$1,200 million a year.

Inefficient Overproduction The subsidy results in inefficient overproduction. At the quantity produced with the subsidy, marginal social benefit is equal to the market price, which has fallen. Marginal social cost has increased and it exceeds the market price.
Because marginal social cost exceeds marginal social benefit, the increased production brings inefficiency.
Subsidies spill over to the rest of the world. Because a subsidy lowers the domestic market price, subsidized farmers will offer some of their output for sale on the world market. The increase in supply on the world market lowers the price in the rest of the world. Faced with lower prices, farmers in other countries decrease production and receive smaller revenues.

## ECONOMICS IN ACTION

## Rich High-Cost Farmers the Winners

Farm subsidies are a major obstacle to achieving an efficient use of resources in the global markets for farm products and are a source of tension between the United States, Europe, and developing nations.
The United States and the European Union are the world's two largest and richest economies. They also pay their farmers the biggest subsidies, which create inefficient overproduction of food in these rich economies.
At the same time, U.S. and European subsidies make it more difficult for farmers in the developing nations of Africa, Asia, and Central and South America to compete in global food markets. Farmers in these countries can often produce at a lower opportunity cost than the U.S. and European farmers.

Two rich countries, Australia and New Zealand, have stopped subsidizing farmers. The result has been an improvement in the efficiency of farming in these countries. New

Zealand is so efficient at producing lamb and dairy products that it has been called the Saudi Arabia of milk (an analogy with Saudi Arabia's huge oil reserve and production.) International opposition to U.S. and European farm subsidies is strong. Opposition to farm subsidies inside the United States and Europe is growing, but it isn't as strong as the profarm lobby, so don't expect an early end to these subsidies.

## Markets for Illegal Goods

The markets for many goods and services are regulated and buying and selling some goods is illegal.The best-known examples of such goods are drugssuch as marijuana, cocaine, ecstasy, and heroin. Despite the fact that these drugs are illegal, trade in them is a multibillion-dollar business. This trade can be understood by using the same economic model and principles that explain trade in legal goods. To study the market for illegal goods, we're first going to examine the prices and quantities that would prevail if these goods were not illegal. Next, we'll see how prohibition works. Then we'll see how a tax might be used to limit the consumption of these goods.

## A Free Market for a Drug

Figure 6.13 shows the market for a drug. The demand curve, $D$, shows that, other things remaining the same, the lower the price of the drug, the larger is the quantity of the drug demanded. The supply curve, 5 , shows that, other things remaining the same, the lower the price of the drug, the smaller is the quantity supplied. If the drug were not illegal, the quantity bought and sold would be $Q c$ and the price would be Pc.

## A Market for an Illegal Drug

When a good is illegal, the cost of trading in the good increases. By how much the cost increases and who bears the cost depend on the penalties for violating the law and the degree to which the law isenforced. The larger the penalties and the better thepolicing, the higher are the costs. Penalties might be imposed on sellers, buyers, or both.

Penalties on Sellers Drug dealers in the United States face large penalties if their activities are detected. Penalties for convicted dealers range from jail terms of 10 years to life and fines that might be as large as $\$ 2$ million. These penalties are part of the cost of supplying illegal drugs, and they bring a decrease in supply-a leftward shift in the supply curve. To determine the new supply curve, we add the cost of breaking the law to the minimum price that drug dealers are willing to accept.

In Fig. 6.13, the cost of breaking the law by selling drugs (CBL) is added to the minimum price those dealers will accept and the supply curve shiftsleftward to $S+C B L$. If penalties were imposed only on sellers, the market equilibrium would move from point $E$ to point $F$.

Fhuse 6.13 A Market for an Illegal Good


The demand curve for drugs is $D$, and the supply curve is $S$. If drugs are not illegal, the quantity bought and sold is $Q_{C}$ at a price of $P_{C}$ - point $E$. If selling drugs is illegat, the cost of breaking the law by selling drugs (CBL) is added to the minimum supply-price and supply decreases to $S+$ CBL. The market moves to point $F$. If buying drugs is illegal, the cost of breaking the law is subtracted from the maximum price that buyers are willing to pay, and demand decreases to $D-C B L$. The market moves to point $G$. With both buying and selling illegal, the supply curve and the demand curve shift and the market moves to point $H$. The market price remains at $P_{C}$, but the market price plus the penaliy for buying rises - point $J$-and the market price minus the penalty for selling falls-point $K$.

Penalties on Buyers In the United States, it is illegal to possess drugs such as marijuana, cocaine, ecstasy, and heroin. Possession of marijuana can bring a prison term of 1 year, and possession ofheroin can bring a prison term of 2 years. Penalties fall on buyers and the cost of breaking the law must be subtracted from the value of the good to determine the maximum price buyers are willing to pay for the drugs. Demand decreases, and the demand curve shifts leftward. In Fig. 6.13, the demand curve shifts to $D$-CBL. If penalties were imposed only on buyers, the market equilibrium would move from point $E$ to point $G$.

Penalties on Both Sellers and Buyers If penalties are imposed on both sellers and buyers, both supplyand demand decrease and both the supply curve and the demand curve shift. In Fig. 6.13, the costs ofbreaking the law are the same for both buyers and sellers, so both curves shift leftward by the sameamount. The market equilibrium moves to point $H$. The market
price remains at the competitive market price $P_{c}$, but the quantity bought decreases to $Q_{P}$. Buyers pay $P_{C}$ plus the cost of breaking the law, which equals $P_{B}$. Sellers receive $P_{C}$ minus the cost of breaking the law, which equals $P_{s}$.
The larger the penalties and the greater the degree of law enforcement, the larger is the decrease in demand and/or supply. If the penalties are heavier on sellers, the supply curve shifts farther than the demand curve and the market price rises above $P_{C}$. If the penalties are heavier on buyers, the demand curve shifts farther than the supply curve and the market price falls below $P_{c}$. In the United States, the penalties on sellers are larger than those on buyers, so the quantity of drugs traded decreases and the market price increases compared with a free market.
With high enough penalties and effective law enforcement, it is possible to decrease demand and/orsupply to the point at which the quantity bought is zero. But in reality, such an outcome is unusual. It does not happen in the United States in the case of illegal drugs. The key reason is the high cost of law enforcement and insufficient resources for the police to achieve effective enforcement. Because of this situation, some people suggest that drugs (and other illegal goods) should be legalized and sold openly but also taxed at a high rate in the same way that legaldrugs such as alcohol are taxed. How would such an arrangement work?

## Legalizing and Taxing Drugs

From your study of the effects of taxes, it is easy to see that the quantity bought of a drug could be decreased if the drug were legalized and taxed. Imposing a sufficiently high tax could decrease the supply, raise the price, and achieve the same decrease in the quantity bought as does a prohibition on drugs. The government would collect large tax revenue.

Illegal Trading to Evade the Tax It is likelythat an extremely high tax rate would be needed to cut the quantity of drugs bought to the level prevailing with a prohibition. It is also likely that many drug dealers and consumers would try to cover up their activities to evade the tax. If they did act in this way, they would face the cost of breaking the law - the tax law. If the penalty for tax law violation is as severe and as effectively policed as drug-dealing laws, the analysis we've already conducted applies alsoto this case. The quantity of drugs bought would depend on the penalties for law breaking and on the way in which the penalties are assigned to buyers and sellers.

Taxes Versus Prohibition: Some Pros and Cons Which is more effective: prohibition or taxes? In favor of taxes and against prohibition is the fact that the taxrevenue can be used to make law enforcement moreeffective. It can also be used to run a more effective education campaign against illegal drug use. In favor of prohibition and against taxes is the fact that prohibition sends a signal that might influence preferences, decreasing the demand for illegal drugs. Also, somepeople intensely dislike the idea of the government profiting from trade in harmful substances.

## ECONOMIC ANALYSIS

- The state minimum wage rate exceeds the federal minimum wage rate in 23 states.
- The state average minimum wage in 2014 was $\$ 8.25$ an hour, $\$ 1$ more than the federal minimum of $\$ 7.25$ an hour, set in 2009.
- Other wage rates have risen and the cost of living has increased, and there was a widely held view that the minimum wage rate needed to rise.
- With disagreement leading to inaction in Washington, an increasing number of state governments began to raise their state minimum wage rates to levels that exceeded the federal minimum wage.
- Supporters of the increases believed that employment would not be adversely affected, while opponents argued that the higher minimum wage would bring job losses.
- The figures illustrate these two opinions about the effect of a higher minimum wage rate.
- Figure 1 illustrates a market for low-skilled labor in which the equilibrium wage rate exceeds the minimum wage.
- The demand for labor is $D_{0}$ and the supply of labor is $S_{0}$, so the equilibrium wage rate is $\$ 8.75$ an hour.
- The minimum wage rate is $\$ 8.25$ an hour, so no one earns the minimum wage.


Figure 1 Minimum Wage Below Equilibrium Wage

- Because the equilibrium wage rate exceeds the minimum wage rate, the quantity of labor demanded equals the quantity of labor supplied and the minimum wage rate has no effect on the market outcome.
- Figure 2 illustrates a market for low-skilled labor in which the equilibrium wage rate is lower than the minimum wage rate.
- The demand for labor is $D_{1}$ and the supply of labor is $S_{1}$, so the equilibrium wage rate is $\$ 7.75$ an hour.
- The minimum wage rate is $\$ 8.25$ an hour, and the quantity of labor employed equals the quantity demanded at the minimum wage rate.
- The quantity of labor demanded and employed is 3.6 million - the quantity estimated by the Bureau of Labor Statistics.
- At the equilibrium wage rate, the quantity of labor supplied is 3.7 million (an assumed quantity) and 0.1 million workers are unemployed.
- With the equilibrium wage rate less than the minimum wage rate, the quantity of labor demanded is less than the quantity of labor supplied and the minimum wage rate brings an increase in unemployment.


Figure 2 Minimum Wage Above Equilibrium Wage

## CHAPTER 7: Global Markets in Action

After studying this chapter, you will be able to:

- Explain how markets work with international trade
- Identify the gains from international trade and its winners and losers
- Explain the effects of international trade barriers
- Explain and evaluate arguments used to justify restricting international trade
iPhones, Wii games, and Nike shoes are just three of the items that you might buy that are not produced in the United States. Why don't we produce phones, games, and shoes in America? Isn't the globalization of production killing good American jobs?
You will find the answers in this chapter. And you will see why global trade is a win-win deal for buyersand sellers. You will also see why governments restrict trade, and in Economics in the News at the end of the chapter, why it is difficult for the United States to make a free-trade deal with Japan. But first, we study the gains from international trade.


## How Global Markets Work

Because we trade with people in other countries, the goods and services that we can buy and consume are not limited by what we can produce. The goods and services that we buy from other countries are our imports; and the goods and services that we sell to people in other countries are our exports.

## International Trade Today

Global trade today is enormous. In 2013, global exports and imports were $\$ 23$ trillion, which is one third of the value of global production. The United States is the world's largest international trader and accounts for 10 percent of world exports and 12 percent of world imports. Germany and China, which rank 2 and 3 behind the United States, lag by a large margin. In 2013, total U.S. exports were $\$ 2.3$ trillion, which is about 14 percent of the value of U.S. production. Total U.S. imports were $\$ 2.7$ trillion, which is about 17 percent of total expenditure in the United States. We trade both goods and services. In 2013, exports of services were about one third of total exports and imports of services were about one fifth of total imports.

## What Drives International Trade?

Comparative advantage is the fundamental force that drives international trade.
Comparative advantage (see Chapter $2, \mathrm{p} .78$ ) is a situation in which a person can perform an activity or produce a good or service at a lower opportunity cost than anyone else. This same idea applies to nations. We can define national comparative advantage as a situation in which a nation can perform an activity or produce a good or service at a lower opportunity cost than any other nation.
The opportunity cost of producing a T-shirt is lower in China than in the United States, so China has a comparative advantage in producing T-shirts. The opportunity cost of
producing an airplane is lower in the United States than in China, so the United States has a comparative advantage in producing airplanes.
You saw in Chapter 2 how Liz and Joe reap gains from trade by specializing in the production of the good at which they have a comparative advantage and then trading with each other. Both are better off.

## ECONOMICS IN ACTION

## We Trade Services for Oil

We import huge amounts of oil - about $\$ 300$ billion in 2012 . How do we pay for all this oil? The answer is by exporting business, professional, and technical services, airplanes, food and drinks, and chemicals. Wealso trade a large quantity of automobiles, but we both export and import them (mainly in trade with Canada).

U.S. Exports and Imports in 2012

This same principle applies to trade among nations. Because China has a comparative advantageat producing T -shirts and the United States has a comparative advantage at producing airplanes, the people of both countries can gain from specializationand trade. China can buy airplanes from the United States at a lower opportunity cost than that at which Chinese firms can produce them. And Americans canbuy T-shirts from China for a lower opportunity cost than that at which U.S. firms can produce them.
Also, through international trade, Chinese producerscan get higher prices for their T-shirts and Boeing can sell airplanes for a higher price. Both countries gain from international trade.
Let's now illustrate the gains from trade that we've just described by studying demand and supply in the global markets for T -shirts and airplanes.

FIGURE 7.1 A Market with Imports

(a) Equilibrium with no international trade

Part (a) shows the U.S. market for T-shirts with no international trade. The U.S. domestic demand curve DUS and U.S. domestic supply curve SUS determine the price of a T -shirt at $\$ 8$ and the quantity of T - shirts produced and bought in the United States at 40 million a year.

Part (b) shows the U.S. market for T-shirts with

(b) Equillbrium in a market with imports
international trade. World demand for and world supply of T shirts determine the world price of a T-shirt, which is \$5. The price in the U.S. market falls to $\$ 5$ a shirt. U.S. purchases of T-shirts increase to 60 million a year, and U.S. production of $T$-shirts decreases to 20 million a year. The United States imports 40 million T-shirts a year.

## Why the United States Imports T-Shirts

The United States imports T-shirts because the rest of the world has a comparative advantage in producing $T$-shirts. Figure 7.1 illustrates how this comparative advantage generates international trade and how trade affects the price of a $T$-shirt and the quantities produced and bought. The demand curve DuS and the supply curve Sus show the demand and supply in the U.S. domestic market only. The demand curve tells us the quantity of T-shirts that Americans are willing to buy at various prices. The supply curve tells us the quantity of T-shirts that U.S. garment makers are willing to sell at various prices-that is, the quantity supplied at each price when all T-shirts sold in the United States are produced in the United States.
Figure 7.I (a) shows what the U.S. T-shirt market would be like with no international trade. The price of a shirt would be $\$ 8$ and 40 million shirts a year would be produced by U.S. garment makers and bought by U.S. consumers. Figure 7.1(b) shows the market for Tshirts with international trade. Now the price of a T-shirt is determined in the world market, not the U.S. domes- tic market. The world price of a T-shirt is less than $\$ 8$, which means that the rest of the world has a comparative advantage in producing T-shirts. The world price line shows the world price at $\$ 5$ a shirt.
The U.S. demand curve, $D_{U S}$, tells us that at $\$ 5$ a shirt, Americans buy 60 million shirts a year.
The U.S. supply curve, $S_{U S}$, tells us that at $\$ 5$ a shirt, U.S. garment makers produce 20 million Tshirts a year. To buy 60 million T-shirts when only 20 million are produced in the United

States, we must import T-shirts from the rest of the world. The quantity of T-shirts imported is 40 million a year.

## Why the United States Exports Airplanes

Figure 7.2 illustrates international trade in airplanes. The demand curve $D_{U S}$ and the supply curve $S_{\text {us }}$ show the demand and supply in the U.S. domestic market only. The demand curve tells us the quantity of airplanes that U.S. airlines are willing to buy at various prices. The supply curve tells us the quantity of airplanes that U.S. aircraft makers are willing to sell at various prices.
Figure 7.2(a) shows what the U.S. airplane market would be like with no international trade. The price of an airplane would be $\$ 100$ million and 400 airplanes a year would be produced by U.S. aircraft makers and bought by U.S. airlines.

Figure 7.2(b) shows the U.S. airplane market with international trade. Now the price of an airplane is determined in the world market and the world priceof an airplane is higher than $\$ 100$ million, which means that the United States has a comparative advantage in producing airplanes. The world priceline shows the world price at $\$ 150$ million.
The U.S. demand curve, $D_{U S}$, tells us that at $\$ 150$ million an airplane, U.S. airlines buy 200 airplanes a year. The U.S. supply curve, Sus, tells us that at $\$ 150$ million an airplane, U.S. aircraft makers produce 700airplanes a year. The quantity produced in the UnitedStates (700 a year) minus the quantity purchased by U.S. airlines (200 a year) is the quantity of airplanes exported, which is 500 airplanes a year.

FIGURE 7.2 A Market with Exports

(a) Equilibrium without international trade

In part (a), the U.S. market with no international trade, the U.S. domestic demand curve $D_{U S}$ and the U.S. domestic supply curve $S_{U S}$ determine the price of an airplane at $\$ 100$ million and 400 airplanes are produced and bought each year. In part (b), the U.S. market with international trade,

(b) Equilibrium in a market with exports
world demand and world supply determine the world price, which is $\$ 150$ million per airplane. The price in the U.S. market rises. U.S. airplane production increases to 700 a year, and U.S. purchases of airplanes decrease to 200 a year. The United States exports 500 airplanes a year.

## Winners, Losers, and Net Gain from Trade

In Chapter 1 (see p. 44), we asked whether globalization is in the self-interest of the low-wage worker in Malaysia who sews your new running shoes and the displaced shoemaker in Atlanta? Is it in the social interest? We're now going to answer these questions. You will learn why producers complain about cheap foreign imports, but consumers of imports never complain.

## Gains and Losses from Imports

We measure the gains and losses from imports by examining their effect on consumer surplus, producersurplus, and total surplus. In the importing country the winners are those whose surplus increases and thelosers are those whose surplus decreases.
Figure 7.3(a) shows what consumer surplus and producer surplus would be with no international trade in T-shirts. U.S. domestic demand, $D_{U S}$, and domestic supply, $S_{U S}$, determine the price and quantity. Total surplus is the sum of consumer surplus and producer surplus. Figure 7.3(b) shows how these surpluses change when the U.S. market opens to imports. The U.S. price falls to the world price. The quantity bought increases to the quantity demanded at the world price and consumer surplus expands from $A$ to the larger area $A+B+D$. The quantity produced in the United States decreases to the quantity supplied at the world price and producer surplus shrinks to the smaller area $C$. Part of the gain in consumer surplus, the area $B$, is a loss of producer surplus - a redistribution of total surplus. But the other part of the increase in consumer surplus, the area $D$, is a net gain. This increase in total surplus results from the lower price and increased purchases and is the gain from imports.
FIGURE 7.3 Gains and Losses in a Market with Imports

(a) Consumer surplus and producer surplus with no international trade

In part (a), with no international trade, the green area shows the consumer surplus and the blue area shows the producer surplus. In part (b), with international trade, the price falls to the

(b) Gains and losses from imports
world price of $\$ 5$ a shirt. Consumer surplus expands from area $A$ to the area $A+B+D$. Producer surplus shrinks to area C. Area $B$ is a transfer of surplus from producers to consumers. Area $D$ is an increase in total surplus-the gain from imports.

## Gains and Losses from Exports

We measure the gains and losses from exports just like we measured those from imports, by their effect onconsumer surplus, producer surplus, and total surplus.
Figure 7.4(a) on page 194 shows the situation with no international trade. Domestic demand, Dus, and domesticsupply, Sus, determine the price and quantity, the consumer surplus, and the producer surplus.
Figure 7.4(b) shows how the consumer surplus and producer surplus change when the good is exported. The price rises to the world price. The quantity bought decreases to the quantity demanded at the world price and the consumer surplus shrinks to the green area $A$. The quantity produced increases to the quantity supplied at the world price and the producer surplus expands to the blue area $B+C+D$.
Part of the gain in producer surplus, the area $B$, is a loss in consumer surplus - a redistribution of the total surplus. But the other part of the increase in producer surplus, the area $D$, is a net gain. This increase in total surplus results from the higher price and increased production and is the gain from exports.

## Gains for All

You've seen that both imports and exports bring gains. Because one country's exports are other countries' imports, international trade brings gain for all countries. International trade is a win-win game.

## International Trade Restrictions

Governments use four sets of tools to influence inter- national trade and protect domestic industries from foreign competition. They are

- Tariffs
- Import quotas
- Other import barriers
- Export subsidies


## Tariffs

A tariff is a tax on a good that is imposed by the importing country when an imported good crosses its international boundary. For example, the government of India imposes a 100 percent tariff on wine imported from California. So when an Indian imports a $\$ 10$ bottle of Californian wine, he pays the Indian government a $\$ 10$ import duty.
Tariffs raise revenue for governments and serve the self-interest of people who earn their incomes in import-competing industries. But as you will see, restrictions on free international trade decrease the gains from trade and are not in the social interest.

The Effects of a Tariff To see the effects of a tariff, let's return to the example in which the United Statesimports T-shirts. With free trade, the T-shirts are imported and sold at the world price. Then, under pressure from U.S. garment makers, the U.S. government imposes a tariff on imported T-shirts. Buyers ofT-shirts must now pay the world price plus the tariff. Several consequences follow and Fig. 7.5 illustratesthem. Figure 7.5(a) shows the situation with free
international trade. The United States produces 20 million T-shirts a year and imports 40 million a year at the world price of $\$ 5$ a shirt. Figure $7.5(\mathrm{~b})$ shows what happens with a tariff set at $\$ 2$ per T-shirt.
FIGURE 7.5 The Effects of a Tariff

(a) Free trade

The world price of a T-shirt is $\$ 5$. With free trade in part [a], Americans buy 60 million T-shirts a year. U.S. garment makers produce 20 million T-shirts a year and the United States imports 40 million a year.

With a tariff of \$2 per T-shirt in part (b), the price in

(b) Market with tariff
the U.S. market rises to $\$ 7$ a T-shirt. U.S. production increases, U.S. purchases decrease, and the quantity imported decreases. The U.S. government collects a tariff revenue of $\$ 2$ on each T-shirt imported, which is shown by the purple rectangle.

The following changes occur in the market for T-shirts:

- The price of a T-shirt in the United States rises by $\$ 2$.
- The quantity of T-shirts bought in the United States decreases.
- The quantity of T-shirts produced in the United States increases.
- The quantity of T-shirts imported into the United States decreases.
- The U.S. government collects a tariff revenue.

Rise in Price of a T-Shirt To buy a T-shirt, Americans must pay the world price plus the tariff, so the price of a T-shirt rises by the $\$ 2$ tariff to $\$ 7$. Figure $7.5(\mathrm{~b})$ shows the new domestic price line, which lies $\$ 2$ above the world price line. The price rises by the full amount of the tariff. The buyer pays the entire tariff because supply from the rest of the world is perfectly elastic (see Chapter 6, pp. 174-175).

Decreases in Purchases The higher price of a T-shirt brings a decrease in the quantity demanded along thedemand curve. Figure 7.5(b) shows the decrease from 60 million T-shirts a year at $\$ 5$ a shirt to 45 million ayear at $\$ 7$ a shirt.

Increase in Domestic Production The higher price of a T-shirt stimulates domestic production, and U.S. garment makers increase the quantity supplied along the supply curve. Figure 7.5(b) shows the increase from 20 million T-shirts at $\$ 5$ a shirt to 35 million a yearat $\$ 7$ a shirt.

Decrease in Imports T-shirt imports decrease by 30 million, from 40 million to 10 million a year. Both the decrease in purchases and the increase in domestic production contribute to this decrease in imports.

Tariff Revenue The government's tariff revenue is $\$ 20$ million- $\$ 2$ per shirt on 10 million imported shirts-shown by the purple rectangle.

Winners, Losers, and the Social Loss from a Tariff A tariff on an imported good creates winners and losers and a social loss. When the U.S. governmentimposes a tariff on an imported good,

- U.S. consumers of the good lose.
- U.S. producers of the good gain.
- U.S. consumers lose more than U.S. producers gain.
- Society loses: a deadweight loss arises.
U.S. Consumers of the Good Lose because the price of a T-shirt in the United States rises, the quantity of T-shirts demanded decreases. The combination of a higher price and smaller quantity bought decreases consumer surplus - the loss to U.S. consumers that arises from a tariff.


## ECONOMICS IN ACTION

## U.S. Tari影 Almosf Gone

The Smoot-Hawley Act, which was passed in 1930, took U.S. tariffs to a peak average rate of 20 percent in 1933. (One third of imports was subject to a 60 percent tariff.) The General Agreement on Tariffs and Trade (GATT) was established in 1947. Since then tariffs have fallen in a series of negotiating rounds, the most significant of which are identified in the figure. Tariffs are now as low as they have ever been but import quotas and other trade barriers persist.


Tariffs: 1930-2013
Sources of dato: U.S. Buracu of the Census, Historical Statistics of the United Stales, Colonial Times to 1970, Bicentennial Edition, Port 1 (Washington, D.C., 1975); Series U-212: updatod from Stalisfical Abstract of the Uniled States: various editions.
U.S. Producers of the Good Gain because the price of an imported T-shirt rises by the amount of the tariff, U.S. T-shirt producers are now able to sell their T- shirts for the world price plus the tariff. At the higher price, the quantity ofT-shirts supplied by U.S. producers increases.

The combination of a higher price and larger quantity produced increases producer surplus the gain to U.S. producers from the tariff.
U.S. Consumers Lose More Than U.S. Producers Gain Consumer surplus decreases for four reasons: Some becomes producer surplus, some is lost in a highercost of production (domestic producers have highercosts than foreign producers), some is lost because imports decrease, and some goes to the governmentas tariff revenue. Figure 7.6 shows these sources of lost consumer surplus.
Figure 7.6(a) shows the consumer surplus and producer surplus with free international trade in T-shirts. Figure 7.6(b) shows the consumer surplus and producer surplus with a \$2 tariff on imported T-shirts.
By comparing Fig. 7.6(b) with Fig. 7.6(a), you can see how a tariff changes these surpluses. Consumer surplus shrinks for four reasons. First, the higher price transfers surplus from consumers to producers. The blue area $B$ represents this loss (and gain of producer surplus). Second, domestic production costs more than imports. The supply curve $S_{U S}$ shows the higher cost of production and the area C shows this loss of consumer surplus. Third, some of the consumer surplus is transferred to the government. The purple area $D$ shows this loss (and gain of government revenue). Fourth, some of the consumer surplus is lost because imports decrease. The gray area $E$ shows this loss.

SocietyLoses:A Deadweight LossArises Some of the loss of consumer surplus is transferred to producersand some is transferred to the government and spenton government programs that people value. But the increase in production cost and the loss from decreased imports is transferred to no one: It is a social loss - a deadweight loss. The areas labeled $C$ and $E$ represent this deadweight loss. Total surplus decreases by the area $C+E$.

## Import Quotas

We now look at the second tool for restricting trade: import quotas. An import quota is a restriction that limits the quantity of a good that may be imported in a given period. Most countries impose import quotas on a wide range of items. The United States imposes them on food products such as sugar and bananas and manufactured goods such as textiles and paper. Import quotas enable the government to satisfy the self-interest of the people who earn their incomes in the import-competing industries. But you will dis-cover that, like a tariff, an import quota decreases the gains from trade and is not in the social interest.

The Effects of an Import Quota The effects of an import quota are similar to those of a tariff The price rises, the quantity bought decreases, and the quantityproduced in the United States increases. Figure 7.7 illustrates the effects.
Figure 7.7(a) shows the situation with free inter- national trade. Figure 7.7(b) shows what happens with an import quota of 10 million T-shirts a year. The U.S. supply curve of Tshirts becomes thedomestic supply curve, $S_{u S}$, plus the quantity thatthe import quota permits. So the supply curve becomes $S_{U S}+q u o t a$. The price of a $T$-shirt rises to
\$7, the quantity of T-shirts bought in the UnitedStates decreases to 45 million a year, the quantity of T-shirts produced in the United States increases to 35 million a year and the quantity of T-shirtsimported into the United States decreases to thequota quantity of 10 million a year. All the effects of this quota are identical to the effects of a $\$ 2$ per shirt tariff, as you can check in Fig. 7.5(b).

FIGURE 7.6 The Winners and Losers from a Tariff

(a) Free trade

The world price of a T-shirt is $\$ 5$. In part ( $a$ ), with free trade, the United States imports 40 million T-shirts. Consumer surplus, producer surplus, and the gains from free trade are as large as possible.

In part (b), a tariff of $\$ 2$ per T-shirt raises the U.S. price

(b) Market with tariff
of a T -shirt to \$7. The quantity imported decreases. Consumer surplus shrinks by the areas $B, C, D$, and $E$. Producer surplus expands by area $B$. The government's tariff revenue is area $D$, and the tariff creates a deadweight loss equal to the area $C+E$.

Winners, Losers, and the Social Loss from an Import Quota An import quota creates winners and losers that are similar to those of a tariff but with an interesting difference.
When the government imposes an import quota,

- U.S. consumers of the good lose.
- U.S. producers of the good gain.
- Importers of the good gain.
- Society loses: a deadweight loss arises.

Figure 7.8 shows these gains and losses from a quota. By comparing Fig. 7.8(b) with a quota and Fig. 7.8(a) with free trade, you can see how an import quota of 10 million T-shirts a year changes the consumer and producer surpluses. Consumer surplus shrinks. This decrease is the loss to consumers from the import quota. The decrease in consumer surplus is made up of four parts. First, some of the consumer surplus is transferred to producers. The blue area Brepresents this loss of consumer surplus (and gain of producer surplus). Second, part of the consumer surplus is lost because the domestic cost of production is higher than the
world price. The area C represents this loss. Third, part of the consumer surplus is transferred to importers who buy T-shirts for $\$ 5$ (the world price) and sell them for $\$ 7$ (the U.S. domestic price). The two areas $D$ represent this loss of consumer surplus and profit for importers. Fourth, part of the consumer surplus is lost because imports decrease. The area Erepresents this loss.

FIGURE 7.7 The Effects of an Import Quota

(a) Free trade

With free international trade, in part (a), Americans buy 60 million T-shirts at the world price. The United States produces 20 million T-shirts and imports 40 million a year. With an import quota of 10 million T-shirts a year, in part (b).

(b) Market with import quota
the supply of T-shirts in the United States is shown by the curve $S_{U S}+$ quota. The price in the United States rises to $\$ 7$ a T-shirt. U.S. production increases, U.S. purchases decrease, and the quantity of T-shirts imported decreases.

The loss of consumer surplus from the higher cost of production and the decrease in imports is a social loss - a deadweight loss. The areas labeled $C$ and $E$ represent this deadweight loss. Total surplusdecreases by the area C+E.
You can now see the one difference between a quota and a tariff. A tariff brings in revenue for the government while a quota brings a profit for the importers. All the other effects are the same, provided the quota is set at the same quantity of imports that results from the tariff.

The world price of a T-shirt is $\$ 5$. In part (a), with freetrade, the United States produces 20 million T-shirts a ye ar and imports 40 million T-shirts. Consumer surplus, producer surplus, and the gain from free international trade are as large as possible.
In part (b), the import quota raises the price of a T-shirt to \$7 The quantity imported decreases. Consumer surplus shrinks by the areas $B, C, D$, and $E$. Producer surplus expands by area B . Importers' profit is the two area D , and the quota creates a deadweight loss equal to $C+E$.

FIGURE 7.8 The Winners and Losers from an Import Quota

(a) Free trade

(b) Market with import quota

## Other Import Barriers

Two sets of policies that influence imports are

- Health, safety, and regulation barriers
- Voluntary export restraints

Health, Safety, and Regulation Barriers Thousands of detailed health, safety, and other regulations restrict international trade. For example, U.S. food imports are examined by the Food and Drug Administration to determine whether the food is "pure, wholesome, safe to eat, and produced under sanitary conditions." The discovery of BSE (mad cow disease) in just one U.S. cow in 2003 was enough to close down international trade in U.S. beef The European Union bans imports of most genetically modified foods, such as U.S.produced soybeans. Although regulations of the type we've just described are not designed to limit international trade, they have thateffect.

Voluntary Export Restraints A voluntary export restraint is like a quota allocated to a foreign exporter of a good. This type of trade barrier isn'tcommon. It was initially used during the 1980s when Japan voluntarily limited its exports of carparts to the United States.

## Export Subsidies

A subsidy is a payment by the government to a producer. You studied the effects of a subsidy on the quantity produced and the price of a subsidized farm product in Chapter 6, pp. 178-179. An export subsidy is a payment by the government to the producer of an exported good. Export subsidies are illegal under a number of international agreements, including the North American Free Trade Agreement (NAFTA), and the rules of the World Trade Organization (WTO).

Although export subsidies are illegal, the subsidies that the U.S. and European Union governments pay to farmers end up increasing domestic production, some of which gets exported. These exports of subsidized farm products make it harder for producers in other countries, notably in Africa and Central and South America, to compete in global markets. Export subsidies bring gains to domestic producers, but they result in inefficient underproduction in the rest of the world and create a deadweight loss.

## ECONOMICS IN ACTION

## Self-Interest Beats the Social Interest

The World Trade Organization (WTO) is an international body established by the world's major trading nations for the purpose of supervising internationaltrade and lowering the barriers to trade. In 2001, at a meeting of trade ministers from all the WTO member-countries held in Doha, Qatar, an agreement was made to begin negotiations to lower tariff barriers and quotas that restrict international trade in farm products and services. These negotiations are called the Doha Development Agenda or the Doha Round. In the period since 2001, thousands of hours of conferences in Cancun in 2003, Geneva in 2004, and Hong Kong in 2005, and ongoing meetings at WTOheadquarters in Geneva, costing millions of taxpayers' dollars, have made disappointing progress.
Rich nations, led by the United States, the European Union, and Japan, want greater access to the markets of developing nations in exchange for allowing those nations greater access to the markets of the rich world, especially those for farm products.
Developing nations, led by Brazil, China, India, and South Africa, want access to the markets of farm products of the rich world, but they also want to protect their infant industries.
With two incompatible positions, these negotiations are stalled and show no signs of a breakthrough. Theself-interests of rich nations and developing nations are preventing the achievement of the social interest.

## The Case Against Protection

You've just seen that free trade promotes prosperityand protection is inefficient. Yet trade is restricted with tariffs, quotas, and other barriers. Why? Seven arguments for trade restrictions are that protecting domestic industries from foreign competition:

- Helps an infant industry grow.
- Counteracts dumping.
- Saves domestic jobs.
- Allows us to compete with cheap foreign labor.
- Penalizes lax environmental standards
- Prevents rich countries from exploiting developing countries.
- Reduces offshore outsourcing that sends good U.S. jobs to other countries.


## Helps an Infant Industry Grow

Comparative advantages change with on-the-job experience-learning-by-doing. When a new industry or a new product is born - an infant industry - it is not as productive as it will become with experience. It is argued that such an industry should be protected from
international competition until it can stand alone and compete. It is true that learning-bydoing can change comparative advantage, but this fact doesn't justify protecting an infant industry. Firms anticipate and benefit from learning-by-doing without protection from foreign competition. When Boeing started to build airplanes, productivity was at first low. But after a period of learning-by- doing, huge productivity gains followed. Boeing didn't need a tariff to achieve these productivity gains.

## Counteracts Dumping

Dumping occurs when a foreign firm sells its exports at a lower price than its cost of production. Dumping might be used by a firm that wants to gain a global monopoly. In this case, the foreign firm sells its output at a price below its cost to drive domestic firms out of business. When the domesticfirms have gone, the foreign firm takes advantage of its monopoly position and charges a higher price for its product. Dumping is illegal under the rules of the World Trade Organization and is usually regarded as a justification for temporary tariffs, which are called countervailing duties. But it is virtually impossible to detect dumping because it is hard to determine a firm's costs. As a result, the test for dumping is whether a firm's export price is below its domestic price. But this test is weak because it is rational for a firm to charge a low price in a market in which the quantity demanded is highly sensitive to price and a higher price in a market in which demand is less price-sensitive.

## Saves Domestic Jobs

First, free trade does destroy some jobs, but it also creates other jobs. It brings about a global rationalization of labor and allocates labor resources to their highest- valued activities. International trade in textiles has cost tens of thousands of U.S. jobs as U.S. textile mills and other factories closed. But tens of thousands ofjobs have been created in other countries as textile mills opened. And tens of thousands of U.S. workershave better-paying jobs than as textile workers because U.S. export industries have expanded and created new jobs. More jobs have been created than destroyed.
Although protection can save particular jobs, it does so at a high cost. For example, until 2005, U.S. textile jobs were protected by an international agreement called the Multifibre Arrangement. The U.S. International Trade Commission (ITC) has estimated that because of import quotas, 72,000 jobs existed in the textile industry that would otherwise have disappeared and that the annual clothing expenditure in the United States was \$15.9 billion ( $\$ 160$ per family) higher than it would have been with free trade. Equivalently, the ITC estimated that each textile job saved cost $\$ 221,000$ a year.
Imports don't only destroy jobs. They create jobs for retailers that sell imported goods and for firms that service those goods. Imports also create jobs by creating income in the rest of the world, some of which is spent on U.S.-made goods and services.

## Allows Us to Compete with Cheap Foreign Labor

With the removal of tariffs on trade between the United States and Mexico, people said we would hear a "giant sucking sound" as jobs rushed to Mexico.
That didn't happen. Why?

It didn't happen because low-wage labor is low-productivity labor. If a U.S. autoworker earns $\$ 30$ an hourand produces 15 units of output an hour, the average labor cost of a unit of output is $\$ 2$. If a Mexican auto-worker earns $\$ 3$ an hour and produces I unit of output an hour, the average labor cost of a unit of output is $\$ 3$. Other things remaining the same, the higher a worker's productivity, the higher is the worker's wage rate. High-wage workers have high productivity; low- wage workers have low productivity.
Itis comparative advantage, notwage differences, that drive international trade and that enable us tocompete with Mexico and Mexico to compete with us.

## Penalizes Lax environmental Standards

Another argument for protection is that it provides an incentive to poor countries to raise their environmental standards - free trade with the richer and "greener" countries is a reward for improved environmental standards.
This argument for protection is weak. First, a poor country cannot afford to be as concerned about itsenvironmental standard as a rich country can. Today, some of the worst pollution of air and water is foundin China, Mexico, and the former communist countries of Eastern Europe. But only a few decades ago, London and Los Angeles topped the pollution league chart. The best hope for cleaner air in Beijing and Mexico City is rapid income growth, which free trade promotes. As incomes in developing countries grow, they have the means to match their desires to improve their environment. Second, a poor country may have a comparative advantage at doing "dirty" work, which helps it to raise its income and at the same time enables the global economy to achieve higher environmental standards than would otherwise be possible.

## Prevents Rich Countries from Exploiting Developing Countries

Another argument for protection is that international trade must be restricted to prevent the people of the rich industrial world from exploiting the poorer people of the developing countries and forcing them to work for slave wages.
Child labor and near-slave labor are serious problems. But by trading with poor countries, we increase the demand for the goods that these countries produce and increase the demand for their labor. When the demand for labor in developing countries increases, the wage rate rises. So, rather than exploiting people in developing countries, trade can improve their opportunities and increase their incomes.

## Reduces Offshore Outsourcing that Sends Good U.S. Jobs to Other Countries

Offshore outsourcing-buying goods, components, or services from firms in other countriesbrings gainsfrom trade identical to those of any other type of trade. We could easily change the names of the items traded from T-shirts and airplanes (the examples in the previous sections of this chapter) to banking services and call-center services (or any other pair of services). A U.S. bank might export banking servicesto Indian firms, and Indians might provide call-center services to U.S. firms. This type of trade would benefit both Americans and Indians, provided the United States has a comparative advantage in banking services and India has a comparative advantage in call- center services.

Despite the gain from specialization and trade that offshore outsourcing brings, many people believe thatit also brings costs that eat up the gains. Why?
A major reason is that it seems to send good U.S. jobs to other countries. It is true that some manufacturing and service jobs are going overseas. But others are expanding at home. The United States imports call-center services, but it exports education, healthcare, legal, financial, and a host of other types of services. The number of jobs in these sectors is expanding and will continue to expand.
The exact number of jobs that have moved to lower-cost offshore locations is not known, and estimates vary. But even the highest estimate is small compared to the normal rate of job creation and laborturnover.
Gains from trade do not bring gains for every single person. Americans, on average, gain from offshore outsourcing, but some people lose. The losers are thosewho have invested in the human capital to do a specific job that has now gone offshore.
Unemployment benefits provide short-term temporary relief for these displaced workers. But the long- term solution requires retraining and the acquisition of new skills.
Beyond bringing short-term relief through unemployment benefits, government has a larger role toplay. By providing education and training, it canenable the labor force of the twenty-first century to engage in the ongoing learning and sometimes rapid retooling that jobs we can't foresee today will demand.
Schools, colleges, and universities will expand and become better at doing their job of producing a morehighly educated and flexible labor force.

## Avoiding Trade Wars

We have reviewed the arguments commonly heard in favor of protection and the counterarguments against it. But one counterargument to protection that is general and quite overwhelming is that protection invites retaliation and can trigger a trade war.
A trade war is a contest in which when one country raises its import tariffs, other countries retaliate with increases of their own, which trigger yet further increases from the first country.
A trade war occurred during the Great Depression of the 1930s when the United States introduced the Smoot-Hawley tariff. Country after country retaliated with its own tariff, and in a short period, world trade had almost disappeared. The costs to all countries were large and led to a renewed international resolve to avoid such selfdefeating moves in the future. The costs also led to attempts to liberalize trade following World War II.

## Why is International Trade Restricted?

Why, despite all the arguments against protection, istrade restricted? There are two key reasons:

- Tariff revenue
- Rent seeking

Tariff Revenue Government revenue is costly to col- lect. In developed countries such as the UnitedStates, a well-organized tax collection system is in place that can generate billions of dollars of incometax and sales tax revenues.
But governments in developing countries have a difficult time collecting taxes from their citizens.
Much economic activity takes place in an informal economy with few financial records. The one area in which economic transactions are well recorded isinternational trade. So tariffs on international trade are a convenient source of revenue in these countries.

Rent Seeking Rent seeking is the major reason why international trade is restricted. Rent seeking is lobbying for special treatment by the government to create economic profit or to divert consumer surplus or producer surplus away from others. Free trade increases consumption possibilities on average, but not everyone shares in the gain and some people even lose. Free trade brings benefits to some and imposes costs on others, with total benefits exceeding total costs. The uneven distribution of costs and benefits is the principal obstacle to achieving more liberal international trade.
The number of winners from free trade is large, but because the gains are spread thinly over a large number of people, the gain per person is small. The winners could organize and become a political forcelobbying for free trade. But political activity is costly. It uses time and other scarce resources and the gainsper person are too small to make the cost of politicalactivity worth bearing.
In contrast, the number of losers from free trade is small, but the loss per person is large. Because the loss per person is large, the people who lose are willing to incur considerable expense to lobby against free trade. Both the winners and losers weigh benefits and costs. Those who gain from free trade weigh the benefits it brings against the cost of achieving it. Those who lose from free trade and gain from protection weigh the benefit of protection against the cost of maintaining it. The protectionists undertake a larger quantity of political lobbying than the free traders.

## Compensating Losers

If, in total, the gains from free international trade exceed the losses, why don't those who gain compensate those who lose so that everyone is in favor of free trade?
Some compensation does take place. When Congress approved the North American Free Trade Agreement (NAFTA) with Canada and Mexico, it set up a $\$ 56$ million fund to support and retrain workers who lost their jobs as a result of the new trade agreement. During NAFTA's first six months, only 5,000 workers applied for benefits under this scheme. The losers from international trade are also compensated indirectly through the normal unemployment compensation arrangements. But only limited attempts are made to compensate those who lose.
The main reason full compensation is not attempted is that the costs of identifying all the losers and estimating the value of their losses would be enormous. Also, it would never be clear whether a person who has fallen on hard times is suffering because of free trade or for other reasons that might be largely under her or his control. Furthermore, some people who look like losers at one point in time might, in fact, end up gaining.

## ECONOMIC ANALYSIS

- Twelve Pacific-rim nations are attempting to reach a trade deal that lowers barriers to trade.
- The United States and Japan, the two largest of the economies, are seeking agreement ahead of the broader negotiations, but obstacles stand in the way of a deal.
- The core of the problem is Japan's wish to protect its farmers, and especially its rice farmers.
- Figure 1 shows how Japan is protecting its rice farmers but damaging its consumers' interest.
- The demand curve is $D_{j}$ and the supply curve is $S_{j}$. With a total ban on rice imports (an import quota of zero), the price of rice is $\$ 4,000$ per ton and 10 million tons are produced and consumed per year.
- If Japan opened up its rice market to free international trade, the price of rice would fall. In Fig. 1, the price falls to the world price (assumed) of $\$ 2,800$ per ton, shown by the line $P W$.
- With free trade, Japan can buy rice for $\$ 2,800$ per ton and the price in Japan falls to that level. The quantity of rice demanded increases to 16 million tons, the quantity supplied decreases to 6 million tons, and 10 million tons are imported. (Assumed quantities.)
- The Japanese rice producers' surplus shrinks and the Japanese consumer surplus expands by the amount of the light green area. Consumer surplus also increases and total surplus increases by the darker green area.
- The farm lobby in Japan is strong and the government is unwilling to risk losing votes by permitting free trade in rice. But it has moved in that direction in its deal with Australia, and it is expected to move further in a deal with the United States.
- Figure 2 shows why the United States is interested in this deal. The demand curve for rice is $D_{\text {US }}$ and the supply curve is $S_{U S}$. If there were no international trade in rice, the price in the United States would be $\$ 1,600$ per ton and 4 million tons would be produced and consumed each year.
- With free trade, the United States can sell rice at the world price of $\$ 2,800$ per ton and the price in the United States rises to that level. The quantity of rice demanded decreases to 2 million tons, the quantity supplied increases to 8 million tons, and 6 million tons are exported. (Assumed quantities.)


Figure 1 The Market for Rice in Japan


Figure 2 The Market for Rice in the United States

- The U.S. consumer surplus shrinks and the U.S. rice producers' surplus expands by the amount of the light blue area. Producer surplus also increases and total surplus increases by the darker blue area.


## PART 3: HOUSEHOLDS' CHOICES

## CHAPTER 8: Utility and Demand

After studying this chapter, you will be able to:

- Explain the limits to consumption and describe preferences using the concept of utility
- Explain the marginal utility theory of consumer choice
- Use marginal utility theory to predict the effects of changes in prices and incomes and to explain the paradox of value
- Describe some new ways of explaining consumer choices

You enjoy sugary drinks and sometimes, perhaps, drink more than is good for your health. What determines our choices about the quantity of sugary drinks we consume? You know that diamonds are expensive and water is cheap. Doesn't that seem odd? Why do we place a higher value on useless diamonds than on essential-to-life water?
The theory of consumer choice that you're going to study in this chapter answers questions like the ones we've just posed. Economics in the News at the end of the chapter applies what you learn to a debate about whether sugary drinks should be banned or taxed to discourage their consumption.

## Consumption Choices

The choices that you make as a buyer of goods and services - your consumption choices - are influenced by many factors. We can summarize them under two broad headings:

- Consumption possibilities
- Preferences


## Consumption Possibilities

Your consumption possibilities are all the things that you can afford to buy. You can afford many different combinations of goods and services, but they are all limited by your income and by the prices that you must pay. For example, you might decide to spend a big part of your income on a gym membership and personal trainer and little on movies and music, or you might spend lots on movies and music and use the free gym at school. The easiest way to describe consumption possibilities is to consider a model consumer who buys onlytwo items. That's what we'll now do. We'll study the consumption possibilities of Lisa, who buys only movies and soda.
Consumption possibilities are limited by income and by the prices of movies and soda. When Lisa spends all her income, she reaches the limits to her consumption possibilities. We describe this limit with a budget line, which marks the boundary between those
combinations of goods and services that a household can afford to buy and those that it cannot afford.
Figure 8.1on page 216 illustrates Lisa's consumption possibilities of movies and soda and her budget line. Lisa hasan income of $\$ 40$ a month, the price of a movie is $\$ 8$, and the price of soda is $\$ 4$ a case. Rows $A$ through $F$ in the table show six possible ways of allocating $\$ 40$ to these two goods. For example, in row $A$ Lisa buys 10 cases of soda and sees no movies; in row $F$ she sees 5 movies and buys no soda; and in row Cshe sees 2 movies and buys 6 cases of soda.
Points $A$ through $F$ in the graph illustrate the possibilities presented in the table, and the line passing through these points is Lisa's budget line.
The budget line constrains choices: It marks the boundary between what is affordable and unaffordable. Lisa can afford all the points on the budget line and inside it. Points outside the line are unaffordable.

Changes in Consumption Possibilities Consumption possibilities change when income or prices change. A rise in income shifts the budget line outward butleaves its slope unchanged. A change in a pricechanges the slope of the line. Our goal is to predict the effects of such changes on consumption choices. To do so, we must determine the choice a consumer makes. The budget line shows what is possible; preferences determine which possibility is chosen. We'll now describe a consumer's preferences.

## Preferences

Lisa's income and the prices that she faces limit her consumption choices, but she still has lots of choice. The choice that she makes depends on her preferences - a description of her likes and dislikes.
You saw one way that economists use to describe preferences in Chapter 2 (p. 74), the concept of marginal benefit and the marginal benefit curve. But you also saw in Chapter 5 (p. 146) that a marginal benefitcurve is also a demand curve. The goal of a theory of consumer choice is to derive the demand curve from a deeper account of how consumers make their buy-ing plans. That is, we want to explain what determines demand and marginal benefit. To achieve this goal, we need a deeper way of describing preferences. One approach to this problem uses the idea of utility, and defines utility as the ben-efit or satisfaction that a person gets from the consumption of goods and services. We distinguish two utility concepts:

- Total utility
- Marginal utility

Total Utility The total benefit that a person gets from the consumption of all the different goods and services is called total utility. Total utility depends on the level of consumptionmore consumption generallygives more total utility.
Toillustrate the concept of total utility, think about Lisa's choices. We tell Lisa that we want to measure her utility from movies and soda. We can use any scale that we wish to measure her total utility and we give her two starting points: (1) We will call the total utility from no
movies and no soda zero utility; and (2) we will call the total utility she getsfrom seeing 1 movie a month 50 units.
We then ask Lisa to tell us, using the same scale, how much she would like 2 movies, and more, up to 10 movies a month. We also ask her to tell us, on the same scale, how much she would like 1 case of soda a month, 2 cases, and more, up to 10 cases a month.
In Table 8.1, the columns headed "Total utility" show Lisa's answers. Looking at those numbers, you can say a lot about how much Lisa likes soda and movies. She says that 1 case of soda gives her 75 units of utility- 50 percent more than the utility that she gets from seeing 1 movie. You can also see that her total utility from soda climbs more slowly than her total utility from movies. This difference turns on the second utility concept: marginal utility.

## TABLE 8.1 Lisa's Utility from Movies and Soda

| Movies |  |  | Soda |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Quantity (per month) | Total utility | Marginal utility | Cases (per month) | Total utility | Marginal utility |
| 0 | 0 | 50 | 0 | 0 | 75 |
| 1 | 50 | 40 | 1 | 75 | 48 |
| 2 | 90 | 32 | 2 | 123 | 36 |
| 3 | 122 | 28 | 3 | 159 | 24 |
| 4 | 150 | 26 | 4 | 183 | 22 |
| 5 | 176 | 24 | 5 | 205 |  |
| 6 | 200 |  | 6 | 225 | 13 |
| 7 | 222 | 20 | 7 | 238 | 10 |
| 8 | 242 | 17 | 8 | 248 | 7 |
| 9 | 259 |  | 9 | 255 | 5 |
| 10 | 275 |  | 10 | 260 |  |

Marginal Utility We define marginal utility as the change in total utility that results from a one-unit increase in the quantity of a good consumed.
In Table 8.1, the columns headed "Marginal utility" show Lisa's marginal utility from movies and soda. You can see that if Lisa increases the soda she buys from 1 to 2 cases a month, her total utility from soda increases from 75 units to 123 units. ForLisa, the marginal utility from the second case each month is 48 units (123-75).
The marginal utility numbers appear midway between the quantities of soda because it is the change in the quantity she buys from 1 to 2 cases that produces the marginal utility of 48 units.
Marginal utility ispositive, but it diminishes as the quantity of a good consumed increases.

Positive Marginal Utility All the things that people enjoy and want more of have a positive marginal utility. Some objects and activities can generate negative marginal utility-and lower total utility. Two examples are hard labor and polluted air. But all the goods and services that people value and that we are thinking about here have positive marginal utility: Total utility increases as the quantity consumed increases.

Diminishing Marginal Utility As Lisa sees more movies, her total utility from movies increases but her marginal utility from movies decreases. Similarly, as she consumes more soda, her total utility from soda increases but her marginal utility from soda decreases.

FIGURE 8.2 Total Utility and Marginal Utility

(a) Total utility

The figure graphs Lisa's total utility and marginal utility from soda based on the numbers for the first 5 cases of soda a month in Table 8.1. Part (a) shows her total utilityincreasing total utility. The bars along the total utility curve show the extra total utility from each additional case of soda-marginal utility. Part (b) shows Lisa's diminishing marginal utility from soda.

(b) Marginal utility

The tendency for marginal utility to decrease as the consumption of a good increases is so general and universal that we give it the status of a principle - theprinciple of diminishing marginalutility.
You can see Lisa's diminishing marginal utility by calculating a few numbers. Her marginal utility from soda decreases from 75 units from the first case to 48 units from the second case and to 36 units from thethird. Her marginal utility from movies decreases from 50 units for the first movie to 40 units for thesecond and 32 units for the third. Lisa's marginal utility diminishes as she buys more of each good.

YourDiminishing Marginal Utility You've beenstudying all day and into the evening, and you've been too busy finishing an assignment to shop for soda. A friend drops by with a can of soda. The utility you get from that soda is the marginal utility from your first soda of the day - from one can. On another day you've been on a soda binge. You've been working on an assignment, but you've guzzled 10 cans of soda while doing so, and are now totally wired. You are happy enough to have one more
can, but the thrill that you get from it is not very large. It is the marginal utility from the eleventh can in a day.

Graphing Lisa's Utility Schedules Figure 8.2(a) illustrates Lisa's total utility from soda. The more soda Lisa consumes in a month, the more total utility she gets. Her total utility curve slopes upward.
Figure 8.2(b) illustrates Lisa's marginal utility from soda. It is a graph of the marginal utility numbers in Table 8.1. This graph shows Lisa's diminishing marginal utility from soda. Her marginal utility curve slopes downward as she consumes more soda. We've described Lisa's consumption possibilities and preferences. Your next task is to see how Lisachooses what to consume.

## Utility-Maximizing Choice

Consumers want to get the most utility possible from their limited resources. They make the choice that maximizes utility. To discover this choice, we combine the constraint imposed by the budget and the consumer's preferences and find the point on the budget line that gives the consumer the maximum attainable utility. Let's find Lisa's utility-maximizing choice.

## A Spreadsheet Solution

Lisa's most direct way of finding the quantities of movies and soda that maximize her utility is to makea table in a spreadsheet with the information and calculations shown in Table 8.2. Let's see what that table tells us.

Find the Just-Affordable Combinations Table 8.2 shows the combinations of movies and soda that Lisacan afford and that exhaust her $\$ 40$ income. For example, in row $A$, Lisa buys only soda and at $\$ 4$ acase she can buy 10 cases. In row $B$, Lisa sees 1 movie and buys 8 cases of soda. She spends $\$ 8$ on the movie. At $\$ 4$ a case, she spends $\$ 32$ on soda and can buy 8 cases. The combination in row $B$ just exhausts her $\$ 40$. The combinations shown in the table are the same as those plotted on her budget line in Fig. 8.1.
We noted that the budget line shows that Lisa can also afford any combination inside the budget line. The quantities in those combinations would be smaller than the ones shown in Table 8.2 and they do not exhaust her $\$ 40$. But smaller quantities don't maximize her utility. Why? The marginal utilities of movies and soda are positive, so the more of each that Lisa buys, the more total utility she gets.

Find the Total Utility for Each Just-Affordable Combination Table 8.2 shows the total utility that Lisa gets from the just-affordable quantities of movies and soda. The second and third columns show the numbers for movies and the fourth and fifth columns show those for soda. The center column adds the total utility from movies to the total utility from soda. This number, the total utility from movies and soda, is what Lisa wants to maximize.
In row $A$ of the table, Lisa sees no movies and buys 10 cases of soda. She gets no utility from movies and 260 units of utility from soda. Her total utility from movies and soda (the center column) is 260 units.

## TABLE 8.2 Lisa's Utility-Maximizing Choice

|  | Movies \$8 |  | Total utility from movies and soda | Soda \$4 |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | Quantity (per month) | Total utility |  | Total utility | $\begin{gathered} \text { Cases } \\ \text { (per month) } \end{gathered}$ |
| A | 0 | 0 | 260 | 260 | 10 |
| B | 1 | 50 | 298 | 248 | 8 |
| $C$ | 2 | 90 | 315 | 225 | 6 |
| D | 3 | 122 | 305 | 183 | 4 |
| E | 4 | 150 | 273 | 123 | 2 |
| F | 5 | 176 | 176 | 0 | 0 |

In row C of the table, Lisa sees 2 movies and buys 6 cases of soda. She gets 90 units of utility from movies and 225 units of utility from soda. Her total utility from movies and soda is 315 units. This combination of movies and soda maximizes Lisa'stotal utility. That is, given the prices of movies andsoda, Lisa's best choice when she has $\$ 40$ to spend is to see 2 movies and buy 6 cases of soda.
If Lisa sees 1 movie, she can buy 8 cases of soda, but she gets only 298 units of total utility 17 units less than the maximum attainable. If she sees 3 movies, she can buy only 4 cases ofsoda. Shegets 305 units of total utility - 10 units less than the maximum attainable.

Consumer Equilibrium We've just described Lisa's consumer equilibrium. A consumer equilibrium is a situation in which a consumer has allocated all of his or her available income in the way that maximizes his or her total utility, given the prices of goods and services. Lisa's consumer equilibrium is 2 movies and 6 cases of soda.
To find Lisa's consumer equilibrium, we did something that an economist might do but that a consumer is not likely to do: We measured her total utility from all the affordable combinations of movies and soda and then, by inspection of the numbers, selected the combination that gives the highest total utility. There is a more natural way offinding a consumer's equilibrium-a way that uses the idea that choices are made at the margin, as you first met in Chapter 1. Let's look at this approach.

## Choosing at the Margin

When you go shopping you don't do utility calculations. But you do decide how to allocate your budget, and you do so in a way that you think is best for you. If you could make yourself better off by spending a few more dollars on an extra unit of one item and the same number of dollars less on something else, you would make that change. So, when you've allocated your budget in the best possible way, you can't make yourself better off by spending more on one item and less on others.

Marginal Utility per Dollar Economists interpret your best possible choice by using the idea of marginal utility per dollar. Marginal utility is the increasein total utility that results from consuming one moreunit of a good. Marginal utility per dollar is the marginal utility from a good that results from spending one more dollar on it.
The distinction between these two marginal concepts is clearest for a good that is infinitely divisible, such as gasoline. You can buy gasoline by the smallest fraction of a gallon and literally choose to spend one more or one less dollar at the pump. The increase in total utility that results from spending one more dollar at the pump is the marginal utility per dollar from gasoline. When you buy a movie ticket or a case of soda, you must spend your dollars in bigger lumps.
To buy our marginal movie ticket or case of soda, you must spend the price of one unit and your total utility increases by the marginal utility from that item.
So to calculate the marginal utility per dollar for movies (or soda), we must divide marginal utility from the good by its price.
Call the marginal utility from movies MUM and the price of a movie $P M$ Then the marginal utility perdollar from movies is

$$
M U_{M} / P_{M}
$$

Call the marginal utility from soda MUs and the price of a case of soda Ps. Then the marginal utilityper dollar from soda is

$$
M U_{S} / P_{S}
$$

By comparing the marginal utility per dollar from all the goods that a person buys, we can determine whether the budget has been allocated in the way that maximizes total utility. Let's see how we use the marginal utility per dollar to define a utility-maximizing rule.

Utility-Maximizing Rule A consumer's total utility is maximized by following the rule:

- Spend all the available income.
- Equalize the marginal utility per dollar for all goods.

Spend all the Available Income Because more consumption brings more utility, only those choices that exhaust income can maximize utility. For Lisa, combinations of movies and soda that leave her with money to spend don't give her as much total utility asthose that exhaust her $\$ 40$ per month income.

Equalize the Marginal Utilityper Dollar The basic idea behind this rule is to move dollars from good $B$ togood $A$ if doing so increases the utility from good $A$ by more than it decreases the utility from good $B$. Such a utility-increasing move is possible if the marginal utility per dollar from good $A$ exceeds that from good $B$.
But buying more of good $A$ decreases its marginal utility. And buying less of good $B$ increases its marginal utility. So by moving dollars from good $B$ to good $A$, total utility rises, and the gap between the marginal utilities per dollar gets smaller.
As long as the gap exists-as long as the marginal utility per dollar from good $A$ exceeds that from good $B$-total utility can be increased by spending more on $A$ and less on $B$. But when
enough dollars have been moved from $B$ to $A$ to make the two marginal utilities per dollar equal, total utility cannot be increased further. Total utility is maximized.

Lisa's Marginal Calculation Let's apply the basic idea to Lisa. To calculate Lisa's marginal utility per dollar, we divide her marginal utility numbers for each quantity of each good by the price of the good. The table in Fig. 8.3 shows these calculations for Lisa, and the graph illustrates the situation on Lisa's budget line. The rows of the table are three of her affordable combinations of movies and soda.

Too much Soda and Too Few Movies Inrow B, Lisa sees 1 movie a month and consumes 8 cases of soda a month. Her marginal utility from seeing 1 movie a month is 50 units. Because the price of a movie is $\$ 8$, Lisa's marginal utility per dollar from movies is 50 units divided by $\$ 8$, or 6.25 units of utility per dollar. Lisa's marginal utility from soda when she consumes 8 cases of soda a month is 10 units. Because the price of soda is \$4 a case, Lisa's marginal utility per dollar from soda is 10 units divided by $\$ 4$, or 2.50 units of utility per dollar.
When Lisa sees 1 movie and consumes 8 cases of soda a month, her marginal utility per dollar from soda is less than her marginal utility per dollar from movies. That is,

$$
M U_{S} / P_{S}<M U_{M} I P_{M}
$$

If Lisa spent an extra dollar on movies and a dollar less on soda, her total utility would increase. She would get 6.25 units from the extra dollar spent on movies and lose 2.50 units from the dollar less spent on soda. Her total utility would increase by 3.75 units ( $6.25-$ 2.50).

Too Little Soda and Too Many Movies In row D, Lisasees 3 movies a month and consumes 4 cases of soda. Hermarginal utility from seeing the third movie a month is 32 units. At a price of $\$ 8$ a movie, Lisa's marginal utility per dollar from movies is 32 units divided by $\$ 8$, or 4 units of utility per dollar. Lisa's marginal utility from soda when she buys 4 cases a month is 24 units. At a price of $\$ 4$ a case, Lisa's marginal utility per dollar from soda is 24 units divided by $\$ 4$, or 6 units of utility per dollar.
When Lisa sees 3 movies and consumes 4 cases of soda a month, her marginal utility per dollar from soda exceeds her marginal utility per dollar from movies.
That is,

$$
M U_{S} / P_{S}>M U_{M} I P_{M}
$$

If Lisa spent an extra dollar on soda and a dollar less on movies, her total utility would increase. She would get 6 units from the extra dollar spent on soda and she would lose 4 units from the dollar less spent on movies. Her total utility would increase by 2 units (6-4).

Utility-Maximizing Movies and Soda InFig. 8.3, if Lisa
moves from row $B$ to row $C$, she increases the movies she sees from 1 to 2 a month and
decreases the sodashe consumes from 8 to 6 cases a month. Her marginal utility per dollar from movies falls to 5 and her marginal utility per dollar from soda rises to 5 .
Similarly, if Lisa moves from row $D$ to row $C$, she decreases the movies she sees from 3 to 2 a month and increases the soda she consumes from 4 to 6 cases a month. Her marginal utility per dollar from movies rises to 5 and her marginal utility per dollar from soda falls to 5 .
When Lisa sees 2 movies and consumes 6 cases of da a month, her marginal utility per dollar from da equals her marginal utility per dollar from movies. That is,

$$
M U_{S} / P_{S}=M U_{M} / P_{M}
$$

Lisa can't move from this allocation of her budget without making herself worse off.

## The Power of Marginal Analysis

The method we've just used to find Lisa's utility-maximizing choice of movies and soda is an example of the power of marginal analysis. Lisa doesn't need a computer and a spreadsheet program to maximize utility. She can achieve this goal by comparing the marginal gain from having more of one good with the marginal loss from having less of another good.
The rule that she follows is simple: If the marginal utility per dollar from movies exceeds the marginal utility per dollar from soda, see more movies and buy less soda; if the marginal utility per dollar from sodaexceeds the marginal utility per dollar from movies, buy more soda and see fewer movies.
More generally, if the marginal gain from an action exceeds the marginal loss, take the action. You will meet this principle time and again in your studyof economics, and you will find yourself using it when you make your own economic choices, especially when you must make big decisions.

## Revealing Preferences

When we introduced the idea of utility, we arbitrarily chose 50 units as Lisa's total utility from 1 movie, and we pretended that we asked Lisa to tell us how many units of utility she got from different quantities of soda and movies.
You're now about to discover that we don't need to ask Lisa to tell us her preferences. We can figure them out for ourselves by observing what she buys at various prices. Also, the units in which we measure Lisa's preferences don't matter. Any arbitrary units will work. In this respect, utility is like temperature. Predictions about the freezing point of water don't depend on the temperature scale; and predictions about a household's consumption choice don't depend onthe units of utility.

FIGURE 8.3 Equalizing Marginal Utilities per Dollar


|  |  | Movies (\$8 each) |  |  | Soda (\$4 per ce |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Quantity | Marginal utility | Marginal utility per dollar | Cases | Marginal utility | Marginal utility per dollar |
| B | 1 | 50 | 6.25 | 8 | 10 | 2.50 |
| C | 2 | 40 | 5.00 | 6 | 20 | 5.00 |
| D | 3 | 32 | 4.00 | 4 | 24. | 6.00 |

The graph shows Lisa's budget line and identifies three points on it. The rows of the table describe these points.

At point $B$ (row $B$ ), with 1 movie and 8 cases of soda, Lisa's marginal utility per dollar from soda is less than that from movies: Buy less soda and see more movies.

At point $D$ (row $D$ ), with 3 movies and 4 cases of soda, Lisa's marginal utility per dollar from soda is greater than that from movies: Buy more soda and see fewer movies.

At point $C$ (row $C$ ), with 2 movies and 6 cases of soda, Lisa's marginal utility per dollar from soda is equal to that from movies: Lisa's utility is maximized.

Lisa's Preferences In maximizing total utility by making the marginal utility per dollar equal for all goods, the units in which utility is measured do not matter.
You've seen that when Lisa maximizes her total utility, her marginal utility per dollar from soda, $M U_{5}!P_{5}$, equals her marginal utility per dollar from movies, $M U M / P M$ That is,

$$
M U_{S} / P_{S}=M U_{M} / P_{M}
$$

Multiply both sides of this equation by the price of soda, $P s$, to obtain

$$
M U_{S}=M U_{M} \times\left(P_{S} / P_{M}\right)
$$

This equation says that the marginal utility from soda, $M U_{5}$, is equal to the marginal utility from movies, MUM, multiplied by the ratio of the price of soda, $P_{5}$, to the price of a movie, $P M$.
The ratio Psi PM is the relative price of soda in terms of movies: It is the number of movies that must be forgone to get 1 case of soda. It is also the opportunity cost of soda. (See Chapter 2, p. 71 and Chapter 3, p. 94.)
For Lisa, when $P M=\$ 8$ and $P_{5}=\$ 4$ we observe that in a month she goes to the movies twice and buys 6 cases of soda. So we know that her MUs from 6 cases of soda equals her MUM from 2 movies multi- plied by $\$ 4 / \$ 8$ or 0.5 . That is, for Lisa, the marginalutility from 6 cases of soda equals one-half of the marginal utility from 2 movies.
If we observe the choices that Lisa makes at more prices, we can find more rows in her utility schedule. By her choices, Lisa reveals her preferences.

Units of Utility Don't Matter Lisa's marginal utility from 6 cases of soda is one-half of her marginal utility from 2 movies. So if the marginal utility from the second movie is 40 units, then the marginal utility from the sixth case of soda is 20 units. But if we call the marginal utility from the second movie 50 units, then the marginal utility from the sixth case of soda is 25 units. The units of utility are arbitrary.

## Predictions of Marginal Utility Theory

We're now going to use marginal utility theory to make some predictions. You will see that marginal utility theory predicts the law of demand. The theory also predicts that a fall in the price of a substitute of a good decreases the demand for the good and that for a normal good, a rise in income increases demand.
All these effects, which in Chapter 3 we simply assumed, are predictions of marginal utility theory.
To derive these predictions, we will study the effects of three events:

- A fall in the price of a movie
- A rise in the price of soda
- A rise in income


## A Fall in the Price of a Movie

With the price of a movie at $\$ 8$ and the price of soda at $\$ 4$, Lisa is maximizing utility by seeing 2 movies and buying 6 cases of soda each month. Then, with no change in her $\$ 40$ income and no change in the price of soda, the price of a movie falls from $\$ 8$ to $\$ 4$. How does Lisa change her buying plans?

Finding the New Quantities of Movies and Soda You can find the effect of a fall in the price of a movie on the quantities of movies and soda that Lisa buys in a three-step calculation.

1. Determine the just-affordable combinations of movies and soda at the new prices.
2. Calculate the new marginal utilities per dollar from the good whose price has changed.
3. Determine the quantities of movies and soda that make their marginal utilities per dollar equal.

Affordable Combinations The lower price of a movie means that Lisa can afford more movies or moresoda. Table 8.3 shows her new affordable combinations. In row $A$, if she continues to see 2 movies a month, she can now afford 8 cases of soda and in row $B$, if she continues to buy 6 cases of soda, she can now afford 4 movies. Lisa can afford any of the combinations shown in the rows of Table 8.3.
The next step is to find her new marginal utilities per dollar from movies.
New Marginal Utilities per Dollar from Movies A person's preferences don't change just because a price has changed. With no change in her preferences, Lisa's marginal utilities in Table 8.3 are the same as those inTable 8.1. But because the price of a movie has changed, the marginal utility per dollar from movies changes. In fact, with a halving of the price of a movie from $\$ 8$ to $\$ 4$, the marginal utility per dollar from movies has doubled.
The numbers in Table 8.3 show Lisa's new marginal utility per dollar from movies for each quantity of movies. The table also shows Lisa's marginal utility per dollar from soda for each quantity.

Equalizing the Marginal Utilities per Dollar You can see that if Lisa continues to see 2 movies a month and buy 6 cases of soda, her marginal utility per dollar from movies (row $A$ ) is 10 units and her marginal utility per dollar from soda (row $B$ ) is 5 units. Lisa isbuying too much soda and too few movies. If she spends a dollar more on movies and a dollar less on soda, her total utility increases by 5 units ( $10-5$ ). If Lisa continues to buy 6 cases of soda and increases the number of movies to 4 (row $B$ ), her marginal utility per dollar from movies falls to 7 units, but her marginal utility per dollar from soda is 5 units. Lisa is still buying too much soda and seeing too few movies. If she spends a dollar more on movies and a dollar less on soda, her total utility increases by 2 units ( $7-5$ ).
But if Lisa sees 6 movies and buys 4 cases of soda a month (row C), her marginal utility per dollar from movies ( 6 units) equals her marginal utility per dollar from soda and she is maximizing utility. If Lisa moves from this allocation of her budget in either direction, her total utility decreases.
table 8.3 How a Change in the Price of Movies Affects Lisa's Choices

|  | Movies <br> (\$4 each) |  |  | Soda (S4 per case) |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Quantity | Marginal utility | Marginal utility per dollar | Cases | Marginal utility | Marginal utility per dollar dolla |
|  | 0 | 0 |  | 10 | 5 | 1.25 |
|  | 1 | 50 | 12.50 | 9 | 7 | 1.75 |
| A | 2 | 40 | 10.00 | 8 | 10 | 2.50 |
|  | 3 | 32 | 8.00 | 7 | 13 | 3.25 |
| B | 4 | 28 | 7.00 | 6 | 20 | 5.00 |
|  | 5 | 26 | 6.50 | 5 | 22 | 5.50 |
| C | 6 | 24 | 6.00 | 4 | 24 | 6.00 |
|  | 7 | 22 | 5.50 | 3 | 36 | 9.00 |
|  | 8 | 20 | 5.00 | 2 | 48 | 12.00 |
|  | 9 | 17 | 4.25 | 1 | 75 | 18.75 |
|  | 10 | 16 | 4.00 | 0 | 0 |  |

Lisa's increased purchases of movies results from a substitution effect-she substitutes the now lower-priced movies for soda - and an income effect - shecan afford more movies.

A Change in the Quantity Demanded Lisa's increase in the quantity of movies that she sees is a change inthe quantity demanded. It is the change in the quantity of movies that she plans to see each month whenthe price of a movie changes and all other influences on buying plans remain the same. We illustrate a change in the quantity demanded by a movement along a demand curve.
Figure 8.4(a) shows Lisa's demand curve for movies. When the price of a movie is $\$ 8$, Lisa sees 2 movies a month. When the price of a movie falls to $\$ 4$, she sees 6 movies a month. Lisa moves downward along her demand curve for movies.
The demand curve traces the quantities that maximize utility at each price, with all other influences remaining the same. You can also see that utility- maximizing choices generate a downward-sloping demand curve. Utility maximization with diminishing marginal utility implies the law of demand.

## A Change in Demand

The decrease in the quantity is the change in the quantity of soda that she plans to buy at a given price of soda when the price of a movie changes. It is a change in her demand for soda. We illustrate a change in demand by a shift of a demand curve. Figure 8.4(b) shows Lisa's demand curve for soda.
The price of soda is fixed at $\$ 4$ a case. When theprice of a movie is $\$ 8$, Lisa buys 6 cases of soda ondemand curve $D_{0}$. When the price of a movie falls to $\$ 4$, Lisa buys 4 cases of soda on demand curve $D_{1}$.
The fall in the price of a movie decreases Lisa's demand for soda. Her demand curve for soda shifts leftward. For Lisa, soda and movies are substitutes.

## A Rise in the Price of Soda

Now suppose that with the price of a movie at $\$ 4$, the price of soda rises from $\$ 4$ to $\$ 8$ a case. How does this price change influence Lisa's buying plans?
We find the answer by repeating the three-step calculation with the new price of soda. Table 8.4 shows Lisa's new affordable combinations. In row $A$, if she continues to buy 4 cases of soda a month she can afford to see only 2 movies; and in row $B$, if she continues to see 6 movies a month, she can afford only 2 cases of soda.
Table 8.4 show Lisa's marginal utility per dollar from soda for each quantity of soda when the price is $\$ 8$ a case. The table also shows Lisa's marginal utility per dollar from movies for each quantity.
If Lisa continues to buy 4 cases of soda (row $A$ ), her marginal utility per dollar from soda is 3. But she must cut the movies she sees to 2 , which increases her marginal utility per dollar from movies to 10 . Lisa is buying too much soda and too few movies. If she spends a dollar less on soda and a dollar more on movies, her utility increases by 7 units ( $10-3$ ).
But if Lisa sees 6 movies a month and cuts her soda to 2 cases (row $B$ ), her marginal utility per dollar from movies ( 6 units) equals her marginal utility perdollar from soda. She is maximizing utility.
Lisa's decreased purchases of soda results from an income effect - she can afford fewer cases and she buys fewer cases. But she continues to buy the same quantity ofmovies.

Lisa's Demand for Soda Now that we've calculated the effect of a change in the price of soda on Lisa'sbuying plans when income and the price of movies remain the same, we have found two points on her demand curve for soda: When the price of soda is $\$ 4$ a case, Lisa buys 4 cases a month; and when the price of soda is $\$ 8$ a case, she buys 2 cases a month.
Figure 8.5 shows these points on Lisa's demand curve for soda. It also shows the change in the quantity of soda demanded when the price of soda rises and all other influences on Lisa's buying plans remain the same.
In this example, Lisa continues to buy the same quantity of movies, but this outcome does not always occur. It is a consequence of Lisa's preferences. With different marginal utilities, she might have decreased or increased the quantity of movies that she sees when the price of soda changes.

TABLE 8.4 How a Change in the Price of Soda Affects Lisa's Choices

|  | Movies <br> (\$4 each) |  |  | Soda (\$8 per case) |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Quantity | Marginal utility | Marginal utility per dollar | Cases | Marginal utility | Marginal utility per dollar dollar |
|  | 0 | 0 |  | 5 | 22 | 2.75 |
| A | 2 | 40 | 10.00 | 4 | 24 | 3.00 |
|  | 4 | 28 | 7.00 | 3 | 36 | 4.50 |
| B | 6 | 24 | 6.00 | 2 | 48 | 6.00 |
|  | 8 | 20 | 5.00 | 1 | 75 | 9.38 |
|  | 10 | 16 | 4.00 | 0 | 0 |  |

You've seen that marginal utility theory predicts the law of demand-the way in which the quantity demanded of a good changes when its price changes. Next, we'll see how marginal utility theory predicts the effect of a change in income on demand.

## A Rise in Income

Suppose that Lisa's income increases from $\$ 40$ to $\$ 56$ a month and that the price of a movie is $\$ 4$ and the price of soda is $\$ 4$ a case. With these prices and with an income of $\$ 40$ a month, Lisa sees 6 movies and buys 4 cases of soda a month (Table 8.3). How does the increase in Lisa's income from $\$ 40$ to $\$ 56$ change her buying plans? Table 8.5 shows the calculations needed to answer this question. If Lisa continues to see 6 movies a month, she can now afford to buy 8 cases of soda (row A); if she continues to buy 4 cases of soda, she can now afford to see 10 movies (row C).
Suppose that Lisa's income increases from $\$ 40$ to $\$ 56$ a month and that the price of a movie is $\$ 4$ and the price of soda is $\$ 4$ a case. With these prices and with an income of $\$ 40$ a month, Lisa sees 6 movies and buys 4 cases of soda a month (Table 8.3). How does the increase in Lisa's income from \$40 to \$56 change her buying plans? Table 8.5 shows the calculations needed to answer this question. If Lisa continues to see 6 movies a month, she can now afford to buy 8 cases of soda (row A); if she continues to buy 4 cases of soda, she can now afford to see 10 movies (row C).

FIGURE 8.5 A Rise in the Price of Soda


When the price of soda rises and the price of a movie and Lisa's income remain the same, the quantity of soda demanded by Lisa decreases. Lisa moves along her demand curve for soda.

In row $A$, Lisa's marginal utility per dollar from movies is greater than her marginal utility per dollarfrom soda. She is buying too much soda and too fewmovies. In row C, Lisa's marginal utility per dollarfrom movies is less than her marginal utility per dollar from soda. She is buying too little soda and too many movies. But in row $B$, when Lisa sees 8 movies a month and buys 6 cases of soda, her marginal utility per dollar from movies equals that from soda. Sheis maximizing utility.
Figure 8.6 on page 226 shows the effects of the rise in Lisa's income on her demand curves for movies and soda. The price of each good is $\$ 4$. When Lisa's income rises to $\$ 56$ a month, she sees 2 more movies and buys two more cases of soda. Her demand curves for both movie and soda increases. With a larger income, the consumer always buys more of a normal good. For Lisa, movies and soda are normal goods.

## TABLE 8.5 Lisa's Choices with an Income of $\$ 56$ a Month

|  | Movies (\$4 each) |  |  | Soda <br> (\$4 per case) |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Quantity | Marginal utility | Marginal utility per dollar | Cases | Marginal utility | Marginal utility per dollar |
|  | 4 | 28 | 7.00 | 10 | 5 | 1.25 |
|  | 5 | 26 | 6.50 | 9 | 7 | 1.75 |
| A | 6 | 24 | 6.00 | 8 | 10 | 2.50 |
|  | 7 | 22 | 5.50 | 7 | 13 | 3.25 |
| B | 8 | 20 | 5.00 | 6 | 20 | 5.00 |
|  | 9 | 17 | 4.25 | 5 | 22 | 5.50 |
| C | 10 | 16 | 4.00 | 4 | 24 | 6.00 |

## The Paradox of Value

The price of water is low and the price of a diamond is high, but water is essential to life while diamonds are used mostly for decoration. How can valuable water be so cheap while a relatively useless diamond is so expensive? This so-called paradox ofvalue has puzzled philosophers for centuries. Not until the theory ?f marginal utility had been developed could anyone give a satisfactory answer.

The Paradox Resolved The paradox is resolved by distinguishing between total utility and marginal utility. The total utility that we get from water is enormous. But remember, the more we consume of something, the smaller is its marginal utility.
We use so much water that its marginal utility- the benefit we get from one more glass of water or another 30 seconds in the shower-diminishes to a small value.
Diamonds, on the other hand, have a small total utility relative to water, but because we buy few diamonds, they have a high marginal utility.
When a household has maximized its total utility, it has allocated its income in the way that makes the marginal utility per dollar equal for all goods. That is, the marginal utility from a good divided by the price of the good is equal for all goods.
This equality of marginal utilities per dollar holds true for diamonds and water: Diamonds have a high price and a high marginal utility. Water has a low price and a low marginal utility. When the high marginal utility from diamonds is divided by the high price of a diamond, the result is a number that equals the low marginal utility from water divided by the low price of water. The marginal utility per dollar is the same for diamonds and water.

Value and Consumer Surplus Another way to think about the paradox of value and illustrate how it is resolved uses consumersurplus. Figure 8.7 explains the paradox of value by using this idea. The supply of water in part (a) is perfectly elastic at price $P_{w}$, so the quantity of water consumed is $Q_{w}$ and the large area shows the consumer surplus from water. The supply of diamonds in part (b) is perfectly inelastic at the quantity $Q_{D}$, so the price of a diamond is $P_{D}$ and the all green area shows the consumer surplus from diamonds. Water is cheap, but brings a large consumer surplus; diamonds are expensive, but bring a small consumer surplus.

FIGURE 8.7 The Paradox of Value

(a) Water

(b) Diamonds

Part (a) shows the demand for and supply of water. Supply is perfectly elastic at the price $P_{w}$, At this price, the quantity of water consumed is $\mathrm{Q}_{\mathrm{w}}$ and the large triangle shows consumer
surplus. Part (b) shows the demand for and supply of diamonds. Supply is perfectly inelastic at the quantity QO. At this quantity, the price of a diamond is Poand the small triangle shows consumer surplus. Water is valuable - has a large consumer surplus - but cheap. Diamonds are less valuable than water - have a smallerconsumer surplus - but are expensive.

## Temperature: An Analogy

Utility is similar to temperature - both are abstract concepts. You can't observe temperature. You can observe water turning to steam if it is hot enough or turning to ice if it is cold enough. You can also construct an instrument - a thermometer - that can help you predict when such changes will occur. We call the scale on the thermometer temperature and we call the units of temperature degrees. But like the units of utility, these degree units are arbitrary. We can use Celsius units or Fahrenheit units or some other units.
The concept of utility helps us to make predictions about consumption choices in much the same way that the concept of temperature helps us to make predictions about physical phenomena.
Admittedly, marginal utility theory does not enable us to predict how buying plans change with the same precision that a thermometer enables us to predict when water will turn to ice or steam. But the theory provides important insights into buying plans and has some powerful implications. It helps us to understand why people buy more of a good or service when its price falls and why people buy more of most goods when their incomes increase. It also resolves the paradox of value.
We're going to end this chapter by looking at some new ways of studying individual economic choices and consumer behavior.

## ECONOMICS IN ACTION

## Maximizing Utility from Recoded Music

In 2012, Americans spent $\$ 7$ billion on recorded music, down from more than $\$ 14$ billion in 2000. But the combined quantity of discs and downloadsbought increased from 1 billion in 2000 to 1.6 billion in 2012 and the average price of a unit of recorded music fell from \$14.00 to \$3.90.
The average price fell because the mix of formats bought changed dramatically. In 2001, we bought 90 million CDs; in 2012, we bought only 211 million CDs and downloaded 1.4 billion music files.

Figure 1 shows the longer history of the changing formats of recorded music.
The music that we buy isn't just one good - it is several goods. Singles and albums are different goods; downloads and discs are different goods; and downloads to a computer and downloads to a cellphone are different goods. There are five major categories and the table shows the quantities of each that webought in 2012 (excluding DVDs and cassettes).

|  | Singles | Albums |
| :--- | :---: | :---: |
| Format | (millions in 2012) |  |
| Disc | 211 | 241 |
| Download | 1,400 | 105 |
| Mobile | 116 | - |

Most people buy all their music in digital form, but many still buy physical CDs and some people buy both downloads and CDs. We get utility from the singles and albums that we buy, and the more songs and albums we have, the more utility we get. But our marginal utility from songs and albums decreases as the quantity that we own increases.
We also get utility from convenience. A song that we can buy with a mouse click and play with the spin of awheel is more convenient both to buy and to use than a song on a CD. The convenience of songs downloaded over the Internet means that, song for song, we getmore utility from a song downloaded than we get from a song on a physical CD. But most albums are still played at home on a CD player. So for most people, a physical CD is a more convenient medium for delivering an album. Album for album, people on average get more utility from a CD than from a download.


Figure 1 Changing Formats of Recorded Music
'When we decide how many singles and albums to download and how many to buy on CD, we compare the marginal utility per dollar from each type of music in each format. We make the marginal utility per dollar from each type of music in each format equal, as the equations below show.
The market for single downloads has created an enormous consumer surplus. Figure 2 shows the demand curve for singles. One point on the demand curve is the 2001 price and quantity- 100 million singles were bought at an average price of $\$ 5.00$. Another point on the demand curve is that for 2012-1,400 million singles downloaded at $\$ 1.20$ each.

If the demand curve has not shifted and is linear (assumed here), we can calculate the increase in consumer surplus generated by the fall in the price and the increase in the quantity demanded. The green area of the figureshows this increase in consumer surplus. Consumer surplus increases by ( $\$ 5.00-\$ 1.20$ ) = \$3.80 on the first 100 million and by $\$ 3.80 \times 1,300 / 2$ on the additional 1,300 singles. So the increase in consumer surplus is $\$ 2.85$ billion.


Figure 2 The Demand for Singles

$$
\begin{aligned}
& \frac{M U_{\text {single downloads }}}{P_{\text {single downloads }}}=\frac{M U_{\text {album downloads }}}{P_{\text {album downloods }}}=\frac{M U_{\text {physical singles }}}{P_{\text {physical singles }}}=\frac{M U_{\text {physical albums }}}{P_{\text {physical albums }}}=\frac{M U_{\text {mobile }}}{P_{\text {mobile }}} \\
& \frac{M U_{\text {single downloads }}}{\$ 1.20}=\frac{M U_{\text {album downloads }}}{\$ 10.99}=\frac{M U_{\text {physical singles }}}{\$ 4.76}=\frac{M U_{\text {physical albums }}}{\$ 9.99}=\frac{M U_{\text {mobile }}}{\$ 2.39}
\end{aligned}
$$

## New Ways of Explaining Consumer Choices

When William Stanley Jevons developed marginal utility theory in the 1860s, he would have loved to look inside people's brains and "see" their utility. But he believed that the human brain was the ultimate black box that could never be observed directly. For Jevons, and for most economists today, the purpose of marginal utility theory is to explain our actions, not what goes on inside our brains.
Economics has developed over the past 150 years with little help from and paying little attention toadvances being made in psychology. Botheconomics and psychology seek to explain human behavior, but they have developed different ways of attacking the challenge.
A few researchers have paid attention to the potential payoff from exploring economic problems by using the tools of psychology. These researchers, some economists and some psychologists, think that marginal utility theory is based on a view of how peoplemake
choices that attributes too much to reason and rationality. They propose an alternative approach based on the methods of psychology.
Other researchers, some economists and some neuroscientists, are using new tools to look inside the human brain and open up Jevons' "black box."
This section provides a very brief introduction to these new and exciting areas of economics. We'llexplore the two related research agendas:

- Behavioral economics
- Neuroeconomics

Behavioral Economics studies the ways in which limits on the human brain's ability to compute and implement rational decisions influences economic behavior - both the decisions that people make and the consequences of those decisions for the way markets work.
Behavioral economics starts with observed behavior. It looks for anomalies-choices that do not seem to be rational. It then tries to account for the anomalies by using ideas developed by psychologists that emphasize features of the human brain that limit rational choice. In behavioral economics, instead of being rational utility maximizers, people are assumed to have threeimpediments that prevent rational choice: bounded rationality, bounded willpower, and bounded self-interest.

Bounded Rationality Bounded rationality is rationality that is limited by the computing power of the human brain. We can't always work out the rationalchoice.
For Lisa, choosing between movies and soda, it seems unlikely that she would have much trouble figuring out what to buy. But toss Lisa some uncertainty and the task becomes harder. She's read the reviews of Ironman 2 on Fandango, but does she really want to see that movie? How much marginal utilitywill it give her? Faced with uncertainty, people might use rules of thumb, listen to the views of others, and make decisions based on gut instinct rather than on rational calculation.

Bounded Willpower Bounded willpower is the less-than-perfect willpower that prevents us from making a decision that we know, at the time of implementing the decision, we will later regret.
Lisa might be feeling particularly thirsty when she passes a soda vending machine. Under Lisa's rational utility-maximizing plan, she buys her soda at the discount store, where she gets it for the lowest possibleprice. Lisa has already bought her soda for this month, but it is at home. Spending $\$ 1$ on a can now means giving up a movie later this month. Lisa's rational choice is to ignore the temporary thirst and stick to her plan. But she might not possess the willpower to do so - sometimes she will and sometimes she won't.

Bounded Self-Interest Bounded self-interest is the limited self-interest that results in sometimes sup- pressing our own interests to help others.
A hurricane hits the Florida coast and Lisa, feeling sorry for the victims, donates $\$ 10$ to a fund-raiser. She now has only $\$ 30$ to spend on movies and soda this month. The
quantities that she buys are not, according to her utility schedule, the ones that maximize her utility.
The main applications of behavioral economics are in two areas: finance, where uncertainty is a key factor in decision making, and savings, where the future is a key factor. But one behavior observed by behavioral economists is more general and might affectyour choices. It is called the endowment effect.

The Endowment Effect The endowment effect is the tendency for people to value something more highlysimply because they own it. If you have allocatedyour income to maximize utility, then the price you would be willing to accept to give up something that you own (for example, your coffee mug) should bethe same as the price you are willing to pay for anidentical one. In experiments, students seem to display the endowment effect: The price they are willing to pay for a coffee mug that is identical to the one theyown is less than the price they would be willing to accept to give up the coffee mug that they own. Behavioral economists say that this behavior contradicts marginal utility theory.

Neuroeconomics is the study of the activity of the human brain when a person makes an economic decision. The discipline uses the observational tools and ideas of neuroscience to obtain a better understanding of economic decisions.
Neuroeconomics is an experimental discipline. In an experiment, a person makes an economic decisionand the electrical or chemical activity of the person'sbrain is observed and recorded using the same type of equipment that neurosurgeons use to diagnose brain disorders.
But observations also show that some economic decisions generate activity in the region of the brain (called the hippocampus) where we store memories of anxiety and fear. Decisions that are influenced by activity in this part of the brain might not be rationalbut instead be driven by fear or panic.
Neuroeconomists are also able to observe the amount of a brain hormone (called dopamine), the quantity of which increases in response to pleasurableevents and decreases in response to disappointingevents. These observations might one day enable neuroeconomists to actually measure utility and shine a bright light inside what was once believed to be the ultimate black box.

## Controversy

The new ways of studying consumer choice that we've briefly described here are being used morewidely to study business decisions and decisions infinancial markets, and this type of research is surelygoing to become more popular.
But behavioral economics and neuroeconomics generate controversy. Most economists hold the view of Jevons that the goal of economics is to explain the decisions that we observe people making and not to explain what goes on inside people's heads.
Most economists would prefer to probe apparent anomalies more deeply and figure out why they are not anomalies after all. Economists also point to the power of marginal utility theory and its ability to explain consumer choice and demand as well as resolve the paradox of value.

## ECONOMIC ANALYSIS

- Concerned that people are choosing to consume more sugary soda than is healthy, New York mayor Michael Bloomberg wants to ban large servings and the city of Richmond, CA, is considering a 1 -penny-per-ounce tax.
- Consumers choose to buy the quantity of sugary drinks that maximizes utility.
- To do so, they make the marginal utility per dollar for all other goods and services equal to the marginal utility per dollar for sugary drinks. That is:

$$
\frac{M U_{O}}{P_{O}}=\frac{M U_{S}}{P_{S}} .
$$

- Because of the way in which drinks are sold, there isn't a single price. The table shows some prices in St. Louis in June 2012.


## Prices of Sugary Drinks

| Size (ounces) | At the movies | Ata $\mathbf{7}$-Eleven |
| :---: | :---: | :---: |
| 20 | 4.00 | 1.09 |
| 32 | 4.50 | 1.29 |
| 44 | 5.00 | 1.49 |
| 52 | 5.50 | 1.69 |
| 104 | [free refill] |  |

- These prices fell us that people who buy their drinks in big cups pay a lower price per ounce.
- Because a person who buys a 52 -ounce cup gets a free refill, the price of the marginal ounce for that person is zero. With a price of zero, the buyer drinks the quantity at which marginal utility is also zero.
- The suggestions for decreasing the consumption of sugary drinks are ways of raising the price.
- A tax raises the price because the tax is added to the price received by the seller. A ban on large cups raises the price because the price per ounce is higher for small cups than for large cups. Both have a similar outcome.
- Faced with a higher price of sugary drinks, a consumer maximizes utility by consuming a smaller quantity of sugary drinks.
- The reason is that consumer equilibrium becomes

$$
\frac{M U_{O}}{P_{O}}=\frac{M U_{S}}{\left(P_{S}+\operatorname{tax}\right)^{\prime}}
$$

- When a tax is imposed, the price doesn't rise by the entire amount of the tax, but generally $P_{S}+\operatorname{tax}$ is greater than the price before that tax was imposed.
- Because $P_{S}+$ tax is greater than $P_{S}, M U_{S}$ must rise to restore the equality of the marginal utilities per dollar. But to increase $M U_{S}$, the quantity of sugary drinks consumed must decrease.
- Figure 1 illustrates and makes clear why consumption of sugary drinks decreases.
- Suppose that with no tax, the budget line is $B L_{0}$ and to make the marginal utilities per dollar equal, the consumer buys 60 ounces of drinks and 100 units of other goods and services per day.
- A tax raises the price to $P_{S}+\operatorname{tax}$ and the budget line becomes steeper as $B L_{1}$.
- If the consumer continues to drink 60 ounces per day, the quantity of other items bought must fall to 50 units a day. $M U_{O}$ rises and

$$
\frac{M U_{O}}{P_{O}}>\frac{M U_{S}}{\left(P_{S}+\operatorname{tax}\right)^{\prime}}
$$

- To restore maximum utility, the consumer buys a smaller quantity of sugary drinks, which increases $M U_{\mathrm{s}}$, and a greater quantity of other goods, which decreases $M U_{O}$. A movement up along the budget line $B L_{1}$ shows these changes in quantities consumed.
- The consumer substitutes other goods and services for sugary drinks until

$$
\frac{M U_{O}}{P_{O}}=\frac{M U_{S}}{\left(P_{S}+\operatorname{tax}\right)^{2}}
$$

- At this point, the consumer is again maximizing utility.


Figure 1 The Effect of a Tax

## CHAPTER 9: Possibilities, Preferences, and Choices

After studying this chapter, you will be able to:

- Describe a household's budget line and show how it changes when prices or income change
- Use indifference curves to map preferences and explain the principle of diminishing marginal rate of substitution
- Predict the effects of changes in prices and income on consumption choices

The iPad has revolutionized the way we read magazines and books and check our grades. Yet the magazine racks and bookstore shelves are still stuffed with traditional printed paper. Similarly, low-priced on-demand movies and DVD rentals have made it easierto watch a movie at home. Yet we're also going to movie theaters in ever-greater numbers. In this chapter, we're going to study a model that explains the choices we make and applies it to choices about using new and old technologies. At the end of the chapter in Economics in the News, we use the model to explain why e-books are taking off and replacing printed books.

## Consumption Possibilities

Consumption choices are limited by income and by prices. A household has a given amount of income to spend and cannot influence the prices of the goods and services it buys. A household's budget line describes the limits to its consumption choices. Let's look at Lisa's budget line.

## Budget Line

Lisa has an income of $\$ 40$ a month to spend. She buys two goods: movies and soda. The price of a movie is $\$ 8$, and the price of soda is $\$ 4$ a case.
Figure 9.1 shows alternative combinations of movies and soda that Lisa can afford. In row $A$, she sees no movies and buys 10 cases of soda. In row $F$, she sees 5 movies and buys no soda. Both of these combinations of movies and soda exhaust the \$40 available. Check that the combination of movies and soda in each of the other rows also exhausts Lisa's $\$ 40$ of income. The numbers in the table and the points $A$ through $F$ in the graph describe Lisa's consumption possibilities.

Divisible and Indivisible Goods Some good - called divisible goods - can be bought in any quantity desired. Examples are gasoline and electricity. We can best understand household choice if we suppose that all goods and services are divisible. For example, Lisa can see half a movie a month on average by seeing one movie every two months. When we think of goods as being divisible, the consumption possibilities are not only the points Athrough F shown in Fig. 9.1, but also all the inter-mediate points that form the line running from $A$ to $F$. This line is Lisa's budget line.

## Ficuke 9.1 The Budget Line



| Consumption <br> possibility | Movies <br> (per month) | Soda <br> (cases per month) |
| :---: | :---: | :---: |
| A | 0 | 10 |
| B | 1 | 8 |
| C | 2 | 6 |
| D | 3 | 4 |
| E | 4 | 2 |
| F | 5 | 0 |

Lisa's budget line shows the boundary between what she can and cannot afford. The rows of the table list Lisa's affordable combinations of movies and soda when her income is $\$ 40$, the price of soda is $\$ 4$ a case, and the price of a movie is $\$ 8$. For example, row $A$ tells us that Lisa spends all of her $\$ 40$ income when she buys 10 cases of soda and sees no movies. The figure graphs Lisa's budget line. Points $A$ through $F$ in the graph represent the rows of the table. For divisible goods, the budget line is the continuous line AF. To calculate the equation for Lisa's budget line, start with expenditure equal to income:

$$
\$ 4 Q_{S}+\$ 8 Q_{M}=\$ 40
$$

Divide by $\$ 4$ to obtain

$$
Q_{S}+2 Q_{M}=10 .
$$

Subtract $2 Q_{M}$ from both sides to obtain

$$
Q_{5}=10-2 Q_{M}
$$

Affordable and Unaffordable Quantities Lisa's budget line is a constraint on her choices. It marks the boundary between what is affordable and what is unaffordable. She can afford any point on the line and inside it. She cannot afford any point outside theline. The constraint on her consumption depends on the prices and her income, and the constraint changes when the price of a good or her income changes. To see how, we use a budget equation.

## Budget Equation

We can describe the budget line by using a budget equation. The budget equation starts with the fact that
Expenditure = Income.

Expenditure is equal to the sum of the price of eachgood multiplied by the quantity bought. For Lisa, Expenditure $=($ Price of soda $X$ Quantity of soda $)+($ Price of a movie $X$ Quantity of movies $)$.

Call the price of soda $P_{s}$, the quantity of soda $\mathrm{Q}_{s}$, the price of a movie $P M$, the quantity of movies $\mathrm{Q}_{\mathrm{m}}$, and income $Y$. We can now write Lisa's budget equation as

$$
P_{S} Q_{S}+P_{M} Q_{M}=Y
$$

Or, using the prices Lisa faces, $\$ 4$ a case of soda and $\$ 8$ a movie, and Lisa's income, $\$ 40$, we get

$$
\$ 4 Q_{S}+\$ 8 Q_{M}=\$ 40
$$

Lisa can choose any quantities of soda ( $Q_{s}$ ) and movies ( $Q_{m}$ ) that satisfy this equation. To find the relationship between these quantities, divide both sides of the equation by the price of soda $\left(P_{S}\right)$ to get

$$
Q_{s}+\frac{P_{M}}{P_{S}} \times Q_{M}=\frac{Y}{P_{s}}
$$

Now subtract the term $\left(P M / P_{S}\right) \times O M$ from both
sides of this equation to get

$$
Q_{S}=\frac{Y}{P_{S}}-\frac{P_{M}}{P_{S}} \times Q_{M}
$$

For Lisa, income $(Y)$ is $\$ 40$, the price of a movie $\left(P_{M}\right)$ is $\$ 8$, and the price of soda $\left(P_{S}\right)$ is $\$ 4$ a case. So Lisa must choose the quantities of movies and soda to satisfy the equation

$$
Q_{s}=\frac{\$ 40}{\$ 4}-\frac{\$ 8}{\$ 4} \times Q_{M}
$$

Or

$$
Q_{S}=10-2 Q_{M}
$$

To interpret the equation, look at the budget line in Fig. 9.1 and check that the equation delivers that budget line. First, set $O M$ equal to zero. The budget equation tells us that $Q_{s}$, the quantity of soda, is $Y / P_{S}$, which is 10 cases. This combination of $Q_{M}$ and $Q_{S}$ is the one shown in
row $A$ of the table in Fig. 9.1. Next set equal to 5. $Q_{S}$ now equals zero (row $F$ of the table). Check that you can derive the other rows.
The budget equation contains two variables chosen by the household ( $Q_{M}$ and $Q_{S}$ ) and two variables that the household takes as given $\left(Y / P_{S}\right.$ and $\left.P_{M} / P_{S}\right)$. Let's look more closely at these variables.

Real Income A household's realincome is its income expressed as a quantity of goods that the household can afford to buy. Expressed in terms of soda, Lisa's real income is YIPs. This quantity is the maximum quantity of soda that she can buy. It is equal to her money income divided by the price of soda. Lisa's money income is $\$ 40$ and the price of soda is $\$ 4$ a case, so her real income in terms of soda is 10 cases, which is shown in Fig. 9.1 as the point at which the budget line intersects the $y$-axis.

Relative Price A relative price is the price of one good divided by the price of another good. In Lisa's budgetequation, the variable $P_{M} / P s$ is the relative price of amovie in terms of soda. For Lisa, $P M$ is $\$ 8$ a movieand Ps is $\$ 4$ a case, so $P_{M} / P s$ is equal to 2 cases ofsoda per movie. That is, to see 1 movie, Lisa must give up 2 cases of soda.
You've just calculated Lisa's opportunity cost of seeing a movie. Recall that the opportunity cost of anaction is the best alternative forgone. For Lisa to see 1 more movie a month, she must forgo 2 cases of soda. You've also calculated Lisa's opportunity cost of soda. For Lisa to buy 2 more cases of soda a month, she must forgo seeing 1 movie. So her opportunity cost of 2 cases of soda is 1 movie. The relative price of a movie in terms of soda is the magnitude of the slope of Lisa's budget line. To calculate the slope of the budget line, recall the formula for slope (see the Chapter 1 Appendix): Slope equals the change in the variable measured on the $y$-axis divided by the change in the variable measured on the $x$-axis as we move along the line. In Lisa's case (Fig. 9.1), the variable measured on the $y$-axis is the quantity of soda and the variable measured on the $x$-axis is the quantity of movies. Along Lisa's budget line, as soda decreases from 10 to 0 cases, movies increase from 0 to 5 . So the magnitude of the slope of the budget line is 10 cases divided by 5 movies, or 2 cases of soda per movie. The magnitude of this slope is exactly the same as the relative price we've just calculated. It is also the opportunity cost of a movie.

A Change in Prices When prices change, so does the budget line. The lower the price of the good measured on the x-axis, other things remaining the same, the flatter is the budget line. For example, if the price of a movie falls from $\$ 8$ to $\$ 4$, real income in terms of soda does not change but the relative price of a movie falls. The budget line rotates outward and becomes flatter, as Fig. 9.2(a) on page 242 illustrates The higher the price of the good measured on the $x$-axis, other things remaining the same, the steeper is the budget line. For example, if the price of a movie rises from $\$ 8$ to $\$ 16$, the relative price of a movie increases. The budget line rotates inward and becomes steeper, as Fig. 9.2(a) illustrates.

A Change in Income A change in money income changes real income but does not change the relative price. The budget line shifts, but its slope does not change. An
increase in money income increases real income and shifts the budget line rightward. A decrease in money income decreases real income and shifts the budget line leftward. Figure 9.2(b) shows the effect of a change in money income on Lisa's budget line. The initial budget line when Lisa's income is $\$ 40$ is the same as in Fig. 9.1. The new budget line shows how much Lisa can buy if her income falls to $\$ 20$ a month. The two budget lines have the same slope because the relative price is the same. The new budget line is closer to the origin because Lisa's real income has decreased.

## Preferences and Indifference Curves

You are going to discover a very cool idea: that of drawing a map of a person's preferences. A preference map is based on the intuitively appealing idea that people can sort all the possible combinations of goods into three groups: preferred, not preferred, and indifferent. To make this idea more concrete, let's ask Lisa to tell us how she ranks various combinations of movies andsoda.
Figure 9.3 shows part of Lisa's answer. She tells us that she currently sees 2 movies and buys 6 cases of soda a month at point C . She then lists all the combinations of movies and soda that she says are just as acceptable to her as her current situation. When we plot these combinations of movies and soda, we get the green curve in Fig. 9.3(a). This curve is the key element in a preference map and is called an indifference curve.
An indifference curve is a line that shows combinations of goods among which a consumer is indifferent. The indifference curve in Fig. 9.3(a) tells us that Lisa is just as happy to see 2 movies and buy 6 cases of soda a month at point $C$ as she is to have thecombination of movies and soda at point $G$ or at anyother point along the curve.
Lisa also says that she prefers all the combinations of movies and soda above the indifference curve in Fig. 9.3(a)-the yellow area-to those on the indifference curve. And she prefers any combination on the indifference curve to any combination in the gray area below the indifference curve.
The indifference curve in Fig. 9.3(a) is just one of a whole family of such curves. This indifference curve appears again in Fig. 9.3(b), labeled $I_{1}$. The curves labeled $I_{0}$ and $I_{2}$ are two other indifference curves. Lisa prefers any point on indifference curve $I_{2}$ to any point on indifference curve $I_{1}$, and she prefers any point on $I_{1}$ to any point on $I_{0}$. We refer to $I_{2}$ as being a higher indifference curve than $I_{1}$ and $I_{1}$ as being higher than $I_{0}$. A preference map is a series of indifference curves that resemble the contour lines on a map. By looking at the shape of the contour lines on a map, we can draw conclusions about the terrain. Similarly, by looking at the shape of the indifference curves, we can draw conclusions about a person's preferences. Let's learn how to "read" a preference map.

## FIGURE 9.3 A Preference Map


(a) An indifference curve

(b) Lisa's preference map

Part (a) shows one of Lisa's indifference curves. She is indifferent between point $C$ (with 2 movies and 6 cases of soda) and all other points on the indifference curve, such as G. She prefers points above the indifference curve to points on it, and she prefers points on the indifference curve to points below it. Part (b) shows three of the indifference curves $I_{0}, I_{1}$, and $I_{2}$ - in Lisa's preference map. She prefers point $J$ to point $C$ or $G$, and she prefers all the points on $\mathrm{I}_{2}$ to those on $\mathrm{I}_{1}$.

## Marginal Rate of Substitution

The marginal rate of substitution (MRS) is the rate at which a person will give up goody (the good measured on the $y$-axis) to get an additional unit of good $x$ (the good measured on the $x$-axis) while remaining indifferent (remaining on the same indifference curve). The magnitude of the slope of an indifferencecurve measures the marginal rate of substitution.

- If the indifference curve is steep, the marginal rate of substitution is high. The person is willing to give up a large quantity of good $y$ to get an additional unit of good $x$ while remaining indifferent.
- If the indifference curve is flat, the marginal rate of substitution is low. The person is willing to give up a small amount of good $y$ to get an additional unit of good $x$ while remaining indifferent.
Figure 9.4 shows you how to calculate the marginal rate of substitution.
At point $C$ on indifference curve $I_{1}$, Lisa buys 6 cases of soda and sees 2 movies. Her marginal rate of substitution is the magnitude of the slope of the indifference curve at point C. To measure this magnitude, place a straight line against, or tangent to, the indifference curve at point $C$. Along that line, as the quantity of soda decreases by 10 cases, the number of movies increases by 5 or 2 cases per movie. At point $C$, Lisa is willing to give up soda for movies at the rate of 2 cases per movie - a marginal rate of substitution of 2 . At point $G$ on indifference curve $I_{1}$, Lisa buys 1.5 cases of soda and sees 6 movies. Her marginal rate of substitution is measured by the slope of the indifference curve at point $G$. That slope is the same as the slope of the tangent to the indifference curve at point G. Now, as the quantity of soda decreases by 4.5 cases, the number of movies increases by 9 - or $1 / 2$ case per movie. At point G , Lisa is willing to give up soda for movies at the rate of $1 / 2$ case per movie - a marginal rate of substitution of $1 / 2$.As Lisa sees more movies and buys less soda, her marginal rate of substitution diminishes. Diminishing marginal rate of substitution is the key assumption about preferences. A diminishing marginal rate of substitution is a general tendency for a person to be willing to give up less of good $y$ to get one more unit of good $x$, while at the same time remaining indifferent as the quantity of $x$ increases. InLisa's case, she isless willing to give up soda to see one more movie asthe number of movies she sees increases.

The magnitude of the slope of an indifference curve is called the marginal rate of substitution (MRS). The line at point $C$ tells us that Lisa is willing to give up 10 cases of soda to see 5 movies. Her marginal rate of substitution at point $C$ is 10 divided by 5 , which equals 2 . The line at point $G$ tells us that Lisa is willing to give up 4.5 cases ofsoda to see 9 movies. Her marginal rate of substitution at point $G$ is 4.5 divided by 9 , which equals $1 / 2$.

Your Diminishing Marginal Rate of Substitution Think about your own diminishing marginal rateof substitution. Imagine that in a week, you drink 10 cases of soda and see no movies. Most likely, you are willing to give up a lot of soda so that you can see just 1 movie. But now imagine that in a week, you buy 1 case of soda and see 6 movies.
Most likely, you will now not be willing to give up much soda to see a seventh movie. As a general rule, the greater the number of movies you see, the smaller is the quantity of soda you are willing to give up to see one additional movie.

The shape of a person's indifference curves incorporates the principle of the diminishing marginal rate of substitution because the curves are bowed toward the origin. The tightness of the bend of an indifference curve tells us how willing a person is to substitute one good for another while remaining indifferent. Let's look at some examples that make this point clear.

FIGURE 9.4 The Marginal Rate of Substitution


## Degree of Substitutability

Most of us would not regard movies and soda as being close substitutes, but they are substitutes. Nomatter how much you love soda, some increase in the number of movies you see will compensate you forbeing deprived of a can of soda. Similarly, no matterhow much you love going to the movies, some number of cans of soda will compensate you for being deprived of seeing one movie. A person's indifferencecurves for movies and soda might look something like those for most ordinary goods and services shown in Fig. 9.5(a).

Close Substitutes Somegoods substitute soeasily for each other that most of us do not even notice which we are consuming. The different brands of marker pens and pencils are examples. Most people don't care which brand of these items they use or where they buy them. A marker pen from the campus bookstore is just as good as one from the local grocery store. You would be willing to forgo a pen fromthe campus store if you could get one more pen from the local grocery store. When two goods are perfectsubstitutes, their indifference curves are straight linesthat slope downward, as Fig. 9.5(b) illustrates. The marginal rate of substitution is constant.

Complements Some goods do not substitute for each other at all. Instead, they are complements. The complements in Fig. 9.5(c) are left and right running shoes. Indifference curves of perfect complements are L-shaped. One left running shoe and one right running shoe are as good as one left shoe and two right shoes. Having two of each is preferred to having one of each, but having two of one and one of the other is no better than having one of each.
The extreme cases of perfect substitutes and perfect complements shown here don't often happen in reality, but they do illustrate that the shape of the indifference curve shows the degree ofsubstitutability between two goods. The closer the two goods are to perfect substitutes, the closer the marginal rate of substitution is to being constant (a straight line), rather than diminishing (a curved line). Indifference curves for poor substitutes are tightly curved and lie between the shapes of those shown in Figs. 9.5(a) and 9.5(c). As you can see in the cartoon, according to the waiter's preferences, Coke and Alsatian white wine are perfect substitutes and each is a complement of pork. We hope the customers agree with him.

FIGURE 9.5 The Degree of Substitutability


The shape of the indifference curves reveals the degree of substitutability between two goods. Part (a) shows the indifference curves for two ordinary goods: movies and soda. To drink less soda and remain indifferent, one must see more movies. The number of movies that compensates for a reduction in soda increases as less soda is consumed. Part (b) shows the indifference curves for two perfect substitutes. For
the consumer to remain indifferent, one fewer marker pen from the local grocery store must be replaced by one extra marker pen from the campus bookstore. Part (c) shows two perfect complements-goods that cannot be substituted for each other at all. Having two left running shoes with one right running shoe is no better than having one of each. But having two of each is preferred to having one of each.

## Predicting Consumer Choices

We are now going to predict the quantities of movies and soda that Lisa chooses to buy. We're also going to see how these quantities change when a price changes or when Lisa's income changes. Finally, we're going to see how the substitution effect and the income effect, two ideas that you met in Chapter 3 (see p. 95), guarantee that for a normal good, the demand curve slopes downward.
figure 9.6 The Best Affordable Choice


Lisa's best affordable choice is at point $C$, the point on her budget line and on her highest attainable indifference curve. At point $C_{\text {, }}$ Lisa's marginal rate of substitution between movies and soda (the magnitude of the slope of the indifference curve $h_{1}$ ) equals the relative price of movies and soda (the slope of the budget line).

## Best Affordable Choice

When Lisa makes her best affordable choice ofmovies and soda, she spends all her income and is on her highest attainable indifference curve. Figure 9.6illustrates this choice: The budget line is from Fig. 9.1 and the indifference curves are from Fig. 9.3(b). Lisa's best affordable choice is 2 movies and 6 cases of soda at point C-the best affordable point

On the Budget Line The best affordable point is on the budget line. For every point inside the budget line, such as point $I$, there are points on the budget line that Lisa prefers. For example, she prefers all the points on the budget line between $F$ and $H$ to point $J$, so she chooses a point on the budget line.

On the Highest Attainable Indifference Curve Every point on the budget line lies on an indifference curve. For example, points $F$ and $H$ lie on the indifferencecurve $I_{0}$. By moving along her budget line from either For $H$ toward $C$, Lisa reaches points on ever higher indifference curves that she prefers to points $F$ or $H$. When Lisa gets to point $C$, she is on the highest attainable indifference curve.

Marginal Rate of Substitution Equals Relative Price At point C, Lisa's marginal rate of substitution between movies and soda (the magnitude of the slope of the indifference curve) is equal to the relative price of movies and soda (the magnitude of the slope of the budget line). Lisa's willingness to pay for a movie equals her opportunity cost of a movie. Let's now see how Lisa's choices change when a price changes.

## A Change in Price

The effect of a change in the price of a good on thequantity of the good consumed is called the priceeffect. We will use Fig. 9.7(a) to work out the price effect of a fall in the price of a movie. We start with the price of a movie at $\$ 8$, the price of soda at $\$ 4$ a case, and Lisa's income at $\$ 40$ a month. In this situation, she buys 6 cases of soda and sees 2 movies a - month at point C. Now suppose that the price of a movie falls to $\$ 4$. With a lower price of a movie, the budget line rotates outward and becomes flatter. The new budget line is the darker orange one in Fig. 9.7(a). For a refresher on how a price change affects the budget line, check back to Fig. 9.2(a). Lisa's best affordable point is now point], where she sees 6 movies and drinks 4 cases of soda. Lisa drinks less soda and watches more movies now that movies are cheaper. She cuts her soda purchases from 6 to 4 cases and increases the number of movies she sees from 2 to 6 a month. When the price of a movie falls and the price of soda and her income remain constant, Lisa substitutes movies for soda.

## ECONOMICS IN ACTION

## Best Affordable Choice of Movies and DVDs

Between 2005 and 2014, box-office receipts rose by more than ticket prices, which means that moviegoing has increased. Why has moviegoing increased? One answer is that the consumer's experience has improved. Movies in 3-D such as Godzilla play better on the big screen than at home. Also, movie theaters are able to charge a higher price for 3-D films and other big hits, which further boosts receipts. But there is another answer, and at first thought an unlikely one: Events in the marketfor DVD rentals have impacted going to the movies. To see why, let's look at the recent history of the DVDrentals market. Back in 2005, Blockbuster was the main player and the price of a DVD rental was around $\$ 4$ a night. Redbox was a fledgling. It had started a year earlier with just 140 kiosks in selected McDonald's restaurants. But Redbox expanded rapidly and in 2014 had outlets across the nation renting DVDs at a price of $\$ 1.20$ anight. Blockbuster was history. The easy access to DVDs at \$1.20 a night trans- formed the markets for movie watching and the figure shows why. A student has a budget of $\$ 50$ a month to allocate to movies. To keep the story dear, we'll suppose that it cost $\$ 10$ to go to a movie in both 2005 and 2014. The price of a DVD rental in 2005 was $\$ 4$, so the student's budget line is the one that runs from 5 movies on the $y$-axis to 12.5 DVD rentals on the $x$-axis. The student's best affordable point is 7 rentals and 2 movies a month. In 2014, the price of a rental falls to $\$ 1.20$ a night but the price of a movie ticket remains at $\$ 10$. So the budget line rotates outward. The student's best affordable point is now at 17 rentals and 3 movies a month. (This student loves movies!) Many other things changed between 2005 and 2014

FIGURE 9.7 Price Effect and Demand Curve

(b) Demand curve

Initially, Lisa's best affordable point is $C$ in part (a). If the price of a movie falls from $\$ 8$ to $\$ 4$, Lisa's best affordable point is $J$. The move from $C$ to $J$ is the price effect.

At a price of $\$ 8$ a movie, Lisa sees 2 movies a month, at point $A$ in part (b). At a price of $\$ 4$ a movie, she sees 6 movies a month, at point $B$. Lisa's demand curve for movies traces out her best affordable quantity of movies as the price of a movie varies.
that influenced the markets for movies and DVD rentals, but the fall in the price of a DVD rental was the biggest influence.


## Best Affordable Movies and DVD Rentals

The Demand Curve In Chapter 3, we asserted that the demand curve slopes downward. We can now derive a demand curve from a consumer's budget line and indifference curves. By doing so, we can see that the law of demand and the down-ward-sloping demand curve are consequences of a consumer's choosing her or his best affordable combination of goods. To derive Lisa's demand curve for movies, lower the price of a movie and find her best affordable point at different prices. We've just done this for two movie prices in Fig. 9.7(a). Figure 9.7(b) high- lights these two prices and two points that lie on Lisa's demand curve for movies. When the price of a movie is $\$ 8$, Lisa sees 2 movies a month at point $A$. When the price falls to $\$ 4$, she increases the number of movies she sees to 6 a month at point $B$. The demand curve is made up of these two points plus all the other points that tell us Lisa's best affordable quantity of movies at each movie price, with the price of soda and Lisa's income remaining the same. As you can see, Lisa's demand curve for movies slopes downward-the lower the price of a movie, the more movies she sees. This is the law of demand. Next, let's see how Lisa changes her purchases of movies and soda when her income changes.

## A Change in Income

The effect of a change in income on buying plans is called the income effect. Let's work out the income effect by examining how buying plans change when income changes and prices remain constant. Figure 9.8 shows the income effect when Lisa's income falls. With an income of $\$ 40$, the price of a movie at $\$ 4$, and the price of soda at $\$ 4$ a case, Lisa's best afford- able point is -she buys 6

FIGURE 9.8 Income Effect and Change in Demand

(b) Demand curve for movies

A change in income shifts the budget line, changes the best affordable point, and changes demand.

In part (a), when lisa's income decreases from $\$ 40$ to $\$ 28$, she sees fewer movies and buys less soda.

In part (b), when Lisa's income is $\$ 40$, her demand curve for movies is $D_{0}$. When Lisa's income falls to $\$ 28$, her demand curve for movies shifts leftward to $D_{1}$. For Lisa, going to the movies is a normal good. Her demand for movies decreases because she now sees fewer movies at each price.
movies and 4 cases of soda. If her income falls to $\$ 28$, her best affordable point is K - she sees 4 movies and buys 3 cases of soda. When Lisa's income falls, she buys less of both goods. Movies and soda are normal goods.

The Demand Curve and the Income Effect A change in income leads to a shift in the demand curve, as shown in Fig. 9.8(b). With an income of $\$ 40$, Lisa'sdemand curve for movies is $D_{0}$, the same as in Fig.9.7(b). But when her income falls to $\$ 28$, she plans to see fewer movies at each price, so her demand curve shifts leftward to $D_{1}$. For a normal good, a fall in its price always increases the quantity bought. We can prove this assertion bydividing the price effect into two parts:

- Substitution Effect
- Income Effect

Figure 9.9(a) shows the price effect and Figs. 9.9(b) and 9.9(c) show the two parts into which we separate the price effect.

Substitution Effect The substitution effect is the effect of a change in price on the quantity bought when theconsumer (hypothetically) remains indifferentbetween the original situation and the new one. To work out Lisa's substitution effect when the price of a movie falls, we must lower her income by enough tokeep her on the same indifference curve as before.
Figure 9.9(a) shows the price effect of a fall in the price of a movie from $\$ 8$ to $\$ 4$. The number of movies increases from 2 to 6 a month. When the price falls, suppose (hypothetically) that we cut Lisa's income to $\$ 28$. What's special about $\$ 28$ ? It is the income that is just enough, at the new price of a movie, to keep Lisa's best affordable point on the same indifference curve (11) as her original point C. Lisa's budget line is now the medium line in Fig. 9.9(b). With the lower price of a movie and a smaller income, Lisa's best affordable point is $K$. The move from $C$ to $K$ along the same indifference curve $I_{1}$ is the substitution effect of the price change. The substitution effect of the fall in the price of a movie is an increase in the quantity of movies from 2 to 4 . The direction of the substitution effect never varies: When the relative price of a good falls, the consumer substitutes more of that good for the other good.

Income Effect To calculate the substitution effect, we gave Lisa a $\$ 12$ pay cut. To calculate the income effect, we give Lisa back her \$12. The $\$ 12$ increase in income shifts Lisa's budget line outward, as shown in Fig. 9.9(c). The slope of the budget line does not change because both prices remain the same. This change in Lisa's budget line is similar to the one illustrated in Fig. 9.8. As Lisa's budget line shifts outward, her consumption possibilities expand and her best affordable point becomes on indifference curve $\mathrm{I}_{1}$. The move from $K$ to $J$ is the income effect of the price change. As Lisa's income increases, she sees more movies. For Lisa, a movie is a normal good. For a normal good, the income effect reinforces the substitution effect. Because the two effects work in the same direction, we can be sure that
the demand curve slopes downward. But some goods are inferior goods. What can we say about the demand for an inferior good?

Inferior Goods Recall that an inferior good is a good for which demand decreases when income increases. For an inferior good, the income effect is negative, which means that a lower price does not inevitablylead to an increase in the quantity demanded. The substitution effect of a fall in the price increases the quantity demanded, but the negative income effect works in the opposite direction and offsets the substitution effect to some degree. The key question is to what degree.
If the negative income effect equals the positive substitution effect, a fall in price leaves the quantity bought the same. When a fall in price leaves the quantity demanded unchanged, the demand curve is vertical and demand is perfectly inelastic.

Eabure 9.9 Substitution Effect and Income Effect

(a) Price effect

When the price of a movie falls from $\$ 8$ to $\$ 4$, Lisa moves from point $C$ to point $J$ in part (a). The price effect is an increase in the number of movies from 2 to 6 a month. This price effect is separated into a substitution effect in part (b) and an income effect in part (c).

(b) Substitution effect

To isolate the substitution effect, we confront Lisa with the new price but keep her on her original indifference curve, $\Lambda_{1}$. The substitution effect is the move from $C$ to $K$ along indifference curve $I_{1}$-an increase from 2 to 4 movies a month.

If the negative income effect issmaller than the positive substitution effect, a fall in price increases the quantity bought and the demand curve still slopes downward like that for a normal good. But the demand for an inferior good might be less elasticthan that for a normal good.
If the negative income effect exceeds the positive substitution effect, a fall in the price decreases the quantity bought and the demand curve slopes upward. This case does not appear to occur in the real world.
You can apply the indifference curve model that you've studied in this chapter to explain the changes in the way we buy recorded music, see movies, and make all our other consumption choices. We allocate our budgets to make our best affordable choices.

Changes in prices and incomes change our best affordable choices and change consumption patterns.
Economics in the News on pp. 252-253 applies the theory of household choice to explain how peoplechose whether to buy their books in electronic or paper format and why e-books boomed in 2011.

(c) Income effect

To isolate the income effect, we confront Lisa with the new price of movies but increase her income so that she can move from the original indifference curve, $l_{1}$, to the new one, $l_{2}$. The income effect is the move from $K$ to $J$-an increase from 4 to 6 movies a month.

## ECONOMIC ANALYSIS

- Sales of e-books are growing rapidly because of the choices that millions of consumers are making. One of these consumers is Andy.
- Andy loves reading, but he also enjoys music. His budget for books and music is limited. So he must choose among the many alternative combinations of books and albums that he can afford.
- Figure 1 shows Andy's indifference curves for books (of all types) and albums.
- Andy's annual budget for albums and books is $\$ 600$. The price of an album is $\$ 10$, the price of a print book is $\$ 20$, and the price of an e-book is $\$ 10$.
- Figure 1 shows two budget lines: one if Andy buys print books and albums and another if he buys e-books and albums.
- In Fig. 1, the price of an e-book reader is $\$ 200$. Andy must spend this amount on a reader if he is to buy e-books, which leaves him with $\$ 400$ for albums and e-books. If he buys 15 e-books, he can afford 25 albums $[(15 \times \$ 10)+(25 \times \$ 10)=\$ 400]$.
- If Andy buys print books and albums, he can afford 15 print books and 30 albums $[(15 \times \$ 20)+(30 \times \$ 10)=$ \$600].
- This combination is Andy's best affordable choice-15 print books and 30 albums shown at point A. Andy doesn't buy e-books.
- Now the price of an e-book reader falls, and today Andy can buy a reader that previously cost $\$ 200$ for $\$ 100$.
- Figure 2 shows what happens to Andy's budget line and his choices.
- If Andy buys print books and albums, nothing changes. He can still afford 15 print books and 30 albums $[(15 \times \$ 20)+(30 \times \$ 10)=\$ 600]$.
- But if he buys e-books, his situation has changed. After spending $\$ 100$ on an e-book reader, Andy is left with $\$ 500$ for albums and e-books. If he buys 15 e-books he can now afford 35 albums $[(15 \times \$ 10)+(35 \times \$ 10)=$ \$500].
- Andy can now afford more albums if he buys the same number of books that he bought when the reader cost $\$ 200$. But that's not Andy's best affordable combination of albums and books.
- The price of an e-book is lower than the price of a print book, so for Andy the relative price of a book has fallen and he can benefit by substituting books for albums.


Figure 1 When the Price of a Reader Is $\mathbf{\$ 2 0 0}$


Figure 2 When the Price of a Reader Is $\$ 100$

- Andy moves along his budget line to the point at which his marginal rate of substitution of books for albums equals the relative price.
- This point occurs at $B$ where Andy buys 25 e-books and 25 albums $[(25 \times \$ 10)+(25 \times \$ 10)=\$ 500]$.
- The surge in e-book sales is the consequence of Andy and other rational consumers responding to the incentive of a change in relative prices and, in particular, a fall in the price of an e-book reader.


## PART THREE: FIRMS AND MARKETS

## CHAPTER 10: Organizing production

After studying this chapter, you will be able to:

- Explain the economic problem that all firms face
- Distinguish between technological efficiency and economic efficiency
- Define and explain the principal-agent problem
- Distinguish among different types of markets
- Explain why markets coordinate some economic activities and why firms coordinate others

In 1990, a British scientist named Tim Berners-Lee invented the World Wide Web, a remarkable ideathat paved the way for the creation of thousands of profitable firms among which are Facebook, Twitter, Google, Amazon, and eBay.
What are the decisions that firms must make? That's the question you study in this chapter. In Economics in the News at the end of the chapter, we'll look at some decisions made by Facebook and Google in the Internet advertising market

## The Firm and Its Economic Problem

The 20 million firms in the United States differ in size and in the scope of what they do, but they all perform the same basic economic functions. Each firm is an institution that hires factors of production and organizes those factors to produce and sell goods and services. Our goal is to predict firms' behavior. To do so, we need to know a firm's goal and the constraints it faces. We start with the goal.

## The Firm's Goal

A firm's goal is to maximize profit. A firm that does not seek to maximize profit is either eliminated or taken over by a firm that does seek that goal. What is the profit that a firm seeks to maximize? To answer this question, we'll look at Campus Sweaters, Inc., a small producer of knittedsweaters owned and operated by Cindy.

## Accounting Profit

In 2012, Campus Sweaters received $\$ 400,000$ for the sweaters it sold and paid out $\$ 80,000$ for wool, $\$ 20,000$ for utilities, $\$ 120,000$ for wages, $\$ 5,000$ for the lease of a computer, and $\$ 5,000$ in interest on a bank loan. These expenses total $\$ 230,000$, so the firm had a cash surplus of $\$ 170,000$.
To measure the profit of Campus Sweaters, Cindy's accountant subtracted $\$ 20,000$ for the depreciation of buildings and knitting machines from the $\$ 170,000$ cash surplus.
Depreciation is the fall in the value of a firm's capital. To calculate depreciation, accountants use Internal Revenue Service rules based on standards established by the Financial Accounting Standards Board. Using these rules, Cindy's accountant calculated that Campus Sweaters made a profit of \$150,000 in 2012.

## Economic Accounting

Accountants measure a firm's profit to ensure that the firm pays the correct amount of income tax and to show its investors how their funds are being used.
Economists measure a firm's profit to enable them to predict the firm's decisions, and the goal of these decisions is to maximize economicprofit. Economic profit is equal to total revenue minus total cost, with total cost measured as the opportunity costof production.

## A Firm's Opportunity Cost of Production

The opportunity cost of any action is the highest-valued alternative forgone. The opportunity cost of production is the value of the best alternative use of theresources that a firm uses in production. A firm's opportunity cost of production is the value of real alternatives forgone. We express opportunity cost in money units so that we can compare andadd up the value of the alternatives forgone. A firm's opportunity cost of production is the sum of the cost of using resources

- Bought in the market
- Owned by the firm
- Supplied by the firm's owner

Resources Bought in the Market A firm incurs an opportunity cost when it buys resources in the market. The amount spent on these resources is an opportunity cost of production because the firm could have bought different resources to produce some other good or service. For Campus Sweaters, the resources bought in the market are wool, utilities, labor, a leased computer, and a bank loan. The $\$ 230,000$ spent on these items in 2012 could have been spent on something else, so it is an opportunity cost of producing sweaters.

Resources Owned by the Firm A firm incurs anopportunity cost when it uses its own capital. The cost of using capital owned by the firm is an opportunity cost of production because the firm could sellthe capital that it owns and rent capital from another firm. When a firm uses its own capital, it implicitly rents it from itself. In this case, the firm's opportunity cost of using the capital it owns is called the implicit rental rate of capital. The implicit rental rate of capital has two components: economic depreciation and forgone interest.

Economic Depreciation Accountants measure depreciation, the fall in the value of a firm's capital, using formulas that are unrelated to the change in the market value of capital. Economic depreciation is the fall in the market value of a firm's capital over a given period. It equals the market price of the capital at the beginning of the period minus the market price of the capital at the end of the period.
Suppose that Campus Sweaters could have sold its buildings and knitting machines on January 1, 2012, for $\$ 400,000$ and that it can sell the same capital on December 31, 2012, for $\$ 375,000$. The firm's economic depreciation during 2012 is $\$ 25,000$ ( $\$ 400,000-\$ 375,000$ ). This forgone $\$ 25,000$ is an opportunity cost of production.

Forgone Interest The funds used to buy capital could have been used for some other purpose, and in their next best use, they would have earned interest. This forgone interest
is an opportunity cost of production. Suppose that Campus Sweaters used $\$ 300,000$ of its own funds to buy capital. If the firm invested its $\$ 300,000$ in bonds instead of a knitting factory (and rented the capital it needs to produce sweaters), it would have earned $\$ 15,000$ a year in interest. This forgone interest is an opportunity cost of production. Resources Supplied by the Firm's Owner A firm's owner might supply both entrepreneurship and labor.

Entrepreneurship The factor of production that organizes a firm and makes its decisions might be supplied by the firm's owner or by a hired entrepreneur. The returnto entrepreneurship is profit, and the profit that an entrepreneurearnsonaverage is called normal profit. Normal profit is the cost of entrepreneurship and is an opportunity cost of production. IfCindy supplies entrepreneurial services herself, and if the normal profit she can earn on these services is $\$ 45,000$ a year, this amount is an opportunity cost of production at Campus Sweaters.

Owner's Labor Services In addition to supplying entrepreneurship, the owner of a firm might supply labor but not take a wage. The opportunity cost of theowner's labor is the wage income forgone by not taking the best alternative job. If Cindy supplies labor to Campus Sweaters, and if the wage she can earn on this labor at another firm is $\$ 55,000$ a year, this amount of wages forgone is an opportunity cost of production at Campus Sweaters.

## Economic Accounting: A Summary

Table 10.1 summarizes the economic accounting. Campus Sweaters' total revenue is $\$ 400,000$; its opportunity cost of production is $\$ 370,000$; and its economic profit is $\$ 30,000$. Cindy's personal income is the $\$ 30,000$ of economic profit plus the $\$ 100,000$ that she earns bysupplying resources to CampusSweaters.

## The Firm's Decisions

To achieve the objective of maximum economicprofit, a firm must make five decisions:

1. What to produce and in what quantities
2. How to produce
3. How to organize and compensate its managers and workers
4. How to market and price its products
5. What to produce itself and buy from others

In all these decisions, a firm's actions are limited by the constraints that it faces. Your next task is to learn about these constraints.

## The Firm's Constraints

Three features of a firm's environment limit the maxi- mum economic profit it can make. They are

- Technologyconstraints
- Information constraints
- Market constraints


## TABLE 10.1 Economic Accounting

| Item |  | Amount |
| :---: | :---: | :---: |
| Total Revenue |  | \$400,000 |
| Cost of Resources Bought in Market |  |  |
| Wool | \$80,000 |  |
| Utilities | 20,000 |  |
| Wages | 120,000 |  |
| Computer lease | 5,000 |  |
| Bank interest | 5,000 | \$230,000 |
| Cost of Resources Owned by Firm |  |  |
| Economic depreciation | \$25,000 |  |
| Forgone interest | 15,000 | \$40,000 |
| Cost of Resources Supplied by Owner |  |  |
| Cindy's normal profit | \$45,000 |  |
| Cindy's forgone wages | 55,000 | \$100,000 |
| Opportunity Cost of Production |  | \$370,000 |
| Economic Profit |  | \$30,000 |

MyEconLab Animation

Technology Constraints Economists definetechnology broadly. A technology is any method of producing a good or service. Technology includes the detailed designs of machines and the layout of the workplace.lt includes the organization of the firm. For example, the shopping mall is one technology for producing retail services. It is a different technology from the catalog store, which in turn is different from the downtown store.
It might seem surprising that a firm's profit is limited by technology because it seems that technological advances are constantly increasing profit opportunities. Almost every day, we learn about some newtechnological advance that amazes us. With computers that speak and recognize our own speech and carsthat can find the address we need in a city we've never visited, we can accomplish more than ever.
Technology advances over time. But at each point in time, to produce more output and gain more revenue, a firm must hire more resources and incur greater costs. The increase in profit that a firm can achieve is limited by the technology available. For example, by using its current plant and workforce, Ford can produce some maximum number of cars per day To produce more cars per day, Ford must hire more resources, which increases its costs and limits the increase in profit that it can make by selling the additional cars.

Information Constraints We never possess all the information we would like to have to make decisions. We lack information about both the future and the present. For example, suppose you plan to buy a new computer. When should you buy it? The answer depends on how the price is going to change in the future. Where should you buy it? The answer dependson the prices at hundreds of different computer stores.
To get the best deal, you must compare the quality and prices in every store. But the opportunity cost of this comparison exceeds the cost of the computer!
A firm is constrained by limited information about the quality and efforts of its workforce, the current and future buying plans of its customers, and the plans of its competitors. Workers might make too little effort, customers might switch to competing sup- pliers, and a competitor might enter the market and take some of the firm's business.
To address these problems, firms create incentives to boost workers' efforts even when no one is monitoring them; conduct market research to lower uncertainty about customers' buying plans; and "spy" on each other to anticipate competitive challenges. But these efforts don't eliminate incomplete information and uncertainty, which limit the economic profit that a firm can make.

Market Constraints The quantity each firm can sell and the price it can obtain are constrained by its customers' willingness to pay and by the prices and marketing efforts of other firms. Similarly, the resources that a firm can buy and the prices it must pay for them are limited by the willingness of people to work for and invest in the firm. Firms spend billions of dollars a year marketing and selling their products.
Some of the most creative minds strive to find the right message that will produce a knockout television advertisement. Market constraints and the expenditures firms make to overcome them limit the profit afirm can make.

## Technological and Economic Efficiency

Microsoft employs a large workforce, and most Microsoft workers possess a large amount of human capital. But the firm uses a small amount of physical capital. In contrast, a coal-mining company employsa huge amount of mining equipment (physical capital) and almost no labor. Why? The answer lies in theconcept of efficiency. There are two concepts of production efficiency: technological efficiency and economic efficiency. Technological efficiency occurs when the firm produces a given output by using the least amount of inputs. Economic efficiency occurs when the firm produces a given output at the least cost. Let's explore the two concepts of efficiency by studying anexample.

Suppose that there are four alternative techniques for making TVs:
A. Robot production. One person monitors the entire computer-driven process.
B. Production line. Workers specialize in a small part of the job as the emerging TV passes them on a production line.
C. Hand-tool production. A single worker uses afew hand tools to make a TV.
D. Bench production. Workers specialize in asmall part of the job but walk from bench to bench to perform their tasks.

Table 10.2 sets out the amounts of labor and capital required by each of these four methods to make 10 TVs a day. Which of these alternative methods are technologically efficient?

## TABLE 10.2 Four Ways of Making 10 TVs a Day

|  |  | Quantities of inputs |  |
| :--- | :--- | ---: | ---: |
|  | Method | Labor | Capital |
| A | Robot production | 1 | 1,000 |
| B | Production line | 10 | 10 |
| C | Hand-tool production | 1,000 | 1 |
| D | Bench production | 100 | 10 |

## Technological Efficiency

Recall that technological efficiency occurs when the firm produces a given output by using the least amount of inputs. Look at the numbers in the tableand notice that method $A$ uses the most capital and the least labor. Method $C$ uses the most labor and the least capital. Method $B$ and method $D$ lie between the two extremes. They use less capital and more labor than method $A$ and less labor but more capital than method C. Compare methods B and $D$. Method $D$ requires 100 workers and 10 units of capital to produce 10 TVs. Method $B$ can produce those same 10 TVs by using 10 workers and the same 10 units of capital. Because method $D$ uses the same amount of capital and more labor than method $B$, method $D$ is not technologically efficient. Are any of the other methods not technologically efficient? The answer is no. Each of the other methods is technologically efficient. Method $A$ uses more capital but less labor than method $B$, and method $C$ uses more labor but less capital than method $B$. Which of the methods are economically efficient?

## Economic Efficiency

Recall that economic efficiency occurs when the firm produces a given output at the least cost. Method $D$, which is technologically inefficient, is also economically inefficient. It uses the same amount of capital as method $B$ but 10 times as much labor, so it costs more. A technologically inefficient method is never economically efficient. One of the three technologically efficient methods is economically efficient. The other two are economically inefficient. But which method is economically efficient depends on factor prices.
In Table 10.3(a), the wage rate is $\$ 75$ per day and the rental rate of capital is $\$ 250$ per day. By studying Table 10.3(a), you can see that method $B$ has the lowest cost and is the economically efficient method.

In Table 10.3(b), the wage rate is $\$ 150$ a day and the rental rate of capital is $\$ 1$ a day. Looking at Table10.3(b), you can see that method $A$ has the lowest cost and is the economically efficient method. In thiscase, capital is so cheap relative to labor that the method that uses

(b) Wage rate $\$ 150$ per day; Capital rental rate $\$ 1$ per day

| Method | Labor | Inputs <br> Capital | Labor cost <br> $(\$ 150$ per day) | Capital cost <br> ( $\$ 1$ per day) |  | Total cost |  |
| :---: | ---: | ---: | :---: | ---: | ---: | ---: | ---: |
| A | $\mathbf{1}$ | 1,000 | $\$ 150$ | $\$$ | $\$ 1,000$ | $=$ | $\$ 1,150$ |
| B | 10 | 10 | 1,500 | + | 10 | $=$ | 1,510 |
| C | 1,000 | 1 | 150,000 | + | 1 | $=$ | 150,001 |

(c) Wage rate $\$ 1$ per day; Capital rental rate $\$ 1,000$ per day

| Method | Labor | Inputs <br> Capital | Labor cost <br> ( $\$ 1$ per day $)$ |  | Capital cost <br> $(\$ 1,000$ per day) |  | Total cost |
| :---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| A | 1 | 1,000 | $\$ 1$ | $\$ 1,000,000$ | $=$ | $\$ 1,000,001$ |  |
| B | 10 | 10 | 10 | + | 10,000 | $=$ | 10,010 |
| C | 1,000 | 1 | 1,000 | + | 1,000 | $=$ | 2,000 |

the most capital is the economically efficient method. In Table 10.3(c), the wage rate is $\$ 1 \mathrm{a}$ day and the rental rate of capital is $\$ 1,000$ a day. You can see thatmethod Chas the lowest cost and is the economicallyefficient method. In this case, labor is so cheap relative to capital that the method that uses the most labor is the economically efficient method. Economic efficiency depends on the relative costs of resources. The economically efficient method is the one that uses a smaller amount of the more expensive resource and a larger amount of the less expensive resource. A firm that is not economically efficient does not maximize profit. Natural selection favors efficient firms and inefficient firms disappear. Inefficient firmsgo out of business or are taken over by firms that pro- duce at lower costs.

## Information and Organization

Each firm organizes the production of goods and services using a combination of two systems:

- Command systems
- Incentive systems


## Command Systems

A command system is a method of organizing production that uses a managerial hierarchy. Commands pass downward through the hierarchy, and information passes upward.
The military uses the purest form of command system. A commander-in-chief makes the big decisions about strategic goals. Beneath this highest level, generals organize their military resources. Beneath the generals, successively lower ranks organize smaller and smaller units but pay attention to ever-increasingdegrees of detail. At the bottom of the hierarchy are the people who operate weapons systems.
Command systems in firms are not as rigid as those in the military, but they share some similar features. Achief executive officer (CEO) sits at the top of a firm'scommand system. Senior executives specialize in man- aging production, marketing, finance, and personnel. Beneath these senior managers are the people who supervise the day-to-day operations of the business and beneath them, the people who operate the firm's machines and who make and sell the firm's products. Managers try to be well informed, but they almost always have incomplete information about what is happening in the divisions for which they are responsible. For this reason, firms also use incentive systems.

## Incentive Systems

An incentive system is a method of organizing production that uses a market-like mechanism inside the firm. Instead of issuing commands, senior managers create compensation schemes to induce workers to perform in ways that maximize the firm's profit.
Incentive systems operate at all levels in a firm. The compensation plan of a CEO might include ashare in the firm's profit, factory floor workers some- times receive compensation based on the quantity they produce, and salespeople, who spend most of their working time alone, are induced to work hard by being paid a small salary and a large performancerelated bonus.

## The Principal-Agent Problem

The principal-agent problem is the problem of devising compensation rules that induce an agent to act in thebest interest of a principal. For example, the stockholders of Texaco are principals, and the firm's managers are agents. The stockholders (the principals) must induce the managers (agents) to act in the stockholders' best interest. Similarly, Mark Zuckerberg (a principal) must induce the designers who are working on the next generation Facebook (agents) to work efficiently.
Agents, whether they are managers or workers, pursue their own goals and often impose costs on a principal. For example, the goal of stockholders of Citicorp (principals) is to maximize the firm's profit-its true profit, not some fictitious paper profit. But the firm's profit depends on the actions ofits managers (agents), and they have their own goals. Perhaps a bank manager takes a customer to a ball game on the pretense that she is building customer loyalty, when in fact she is simply enjoying on-the-job leisure. This same manager is also a principal, and her tellers are agents. The manager wants the tellers to work hard and attract new customers so that she canmeet her operating targets. But the workers slack offand take on-the-job leisure.

## Coping with the Principle-Agent Problem

A principal must create incentives that induce each agent to work in the interests of the principal. Three ways of coping with the principal-agent problem are

- Ownership
- Incentive pay
- Long-term contracts

Ownership By assigning ownership (or part- ownership) of a business to managers or workers, it is sometimes possible to induce a job performance that increases a firm's profits. Part-ownership is quite common for senior managers but less common for workers. When United Airlines was running into problems a few years ago, it made most of its employees owners of the company.

Incentive Pay - pay related to performance - is very common. Incentives are based on a variety of performance criteria such as profits, production, or salestargets. Promoting an employee for good performance is another example of the use of incentive pay.

Long-term Contracts tie the long-term fortunes of managers and workers (agents) to the success of the principal(s) - the owner(s) of the firm. For example, a multiyear employment contract for a CEO encourages that person to take a long term view and devise strategies that achieve maxi- mum profit over a sustained period. These three ways of coping with the principal- agent problem give rise to different types of business organization.

## Types of business organization

The three main types of business organization are

- Proprietorship
- Partnership
- Corporation

Proprietorship Aproprietorship is a firm with a single owner-a proprietor-who has unlimited liability.
Unlimited liability is the legal responsibility for all the debts of a firm up to an amount equal to the entire personal wealth of the owner. Farmers, computer programmers, and artists are often proprietorships.
The proprietor makes management decisions, receives the firm's profits, and is responsible forits losses. Profits from a proprietorship are taxed at the same rate as other sources of the proprietor's personal income.

Partnership Apartnership is a firm with two or more owners who have unlimited liability. Partners must agree on an appropriate management structure and on how to divide the firm's profits among them- selves. The profits of a partnership are taxed as the personal income of the owners, but each partner is legally liable for all the debts of the partnership
(limited only by the wealth of that individual partner). Liability for the full debts of the partnership is calledjoint unlimited liability. Most law firms are partnerships.

Corporation A corporation is a firm owned by one or more limited liability stockholders. Limited liability means that the owners have legal liability only for the value of their initial investment. This limitation of liability means that if the corporation becomes bankrupt, its owners do not use their personal wealth to pay the corporation's debts.
Corporations' profits are taxed independently of stockholders' incomes. Stockholders pay a capital gains tax on the profit they earn when they sell a stock for a higher price than they paid for it.
Corporate stocks generate capital gains when a corporation retains some of its profit and reinvests it in profitable activities. So retained earnings are taxed twice because the capital gains they generate are taxed. Dividend payments are also taxed but at a lower rate than other sources ofincome.

Prosand Cons of Different Types of Firms The different types of business organization arisefrom firms trying to cope with the principal-agent problem. Each type has advantages in particular situations and because of its special advantages, each type continues to exist. Each type of business organizationalso has disadvantages.
Table 10.4 summarizes these and other pros and cons of the different types of firms.

## Markets and the Competitive Environment

The markets in which firms operate vary a great deal. Some are highly competitive, and profits in these markets are hard to come by. Some appear to be almost free from competition, and firms in these markets earn large profits. Some markets are dominated by fierce advertising campaigns in which each firm seeks to persuade buyers that it has the best products. And some markets display the character of a strategic game.
Economists identify four market types:

1. Perfect competition
2. Monopolistic competition
3. Oligopoly
4. Monopoly

Perfect Competition arises when there are many firms, each selling an identical product, many buyers, and no restrictions on the entry of new firms into the industry. The many firms and buyers are all well informed about the prices of the products of each firm in the industry. The worldwide markets for wheat, corn, rice, and other grain crops are examples of perfect competition.

| Type of Firm | Pros | Cons |
| :---: | :---: | :---: |
| Proprietorship | - Easy to set up <br> - Simple decision making <br> - Profits taxed only once as owner's income | - Bad decisions not checked; no need for consensus <br> - Owner's entire wealth at risk <br> - Firm dies with owner <br> - Cost of capital and labor is high relative to that of a corporation |
| Parmership | - Easy to set up <br> - Diversified decision making <br> - Can survive withdrawal of partner <br> - Profits taxed only once as owners' incomes | - Achieving consensus may be slow and expensive <br> - Owners' entire wealth at risk <br> - Withdrawal of partner may create capital shortage <br> - Cost of capital and labor is high relative to that of a corporation |
| Corporation | - Owners have limited liability <br> - Large-scale, low-cost capital available <br> - Professional management not restricted by ability of owners <br> - Perpetual life <br> - Long-term labor contracts cut labor costs | - Complex management structure can make decisions slow and expensive <br> - Retained profits taxed twice: as company profit and as stockholders' capital gains |

Monopolistic competition is a marketstructure in which a large number of firms compete by making similar but slightly different products. Making a product slightly different from the product of a competing firm is called product differentiation. Product differentiation gives a firm in monopolistic competition an element of market power. The firm is the sole producer of the particular version of the good in question. For example, in the market for pizzas, hundreds of firms make their own version of the perfect pizza. Each of these firms is the sole producer of a particular brand. Differentiated products are not necessarily different products. Whatmatters is that consumers perceive them to be different. For example, different brands of potato chips and ketchup might be almost identical but be perceived by consumers to be different.

Oligopoly is a market structure in which a small number of firms compete. Computer software, air- plane manufacture, and international air transportationare examples of oligopolistic industries. Oligopolies might produce almost identical products, such as thecolas produced by Coke and Pepsi. Or they might produce differentiated products such as Boeing and Airbus aircraft.

Monopoly arises when there is one firm, which produces a good or service that has no close substitutes and in which the firm is protected by a barrier preventing the entry of new firms. In some places, the phone, gas, electricity, cable television, and water suppliers are local monopolies-monopolies restricted to a given location. Microsoft Corporation, the software developer that created Windows, the operating system for the personalcomputer, is an
example of a global monopoly. Perfect competition is the most extreme form of competition. Monopoly is the most extreme absence of competition. The other two market types fall between these extremes.
Many factors must be taken into account to deter- mine which market structure describes a particular real-world market. One of these factors is the extent to which a small number of firms dominates the market. To measure this feature of markets, economists use indexes called measures of concentration. Let's look at these measures.

## Measures of Concentration

Economists use two measures of concentration:

- The four-firm concentration ratio
- The Herfindahl-Hirschman Index

The Four-Firm Concentration Ratio The four-firm con- centration ratio is the percentage of the value of sales accounted for by the four largest firms in an industry. The range of the concentration ratio is from almost zero for perfect competition to 100 percent for monopoly. This ratio is the main measure used to assess marketstructure.
Table 10.5 shows two calculations of the four-firm concentration ratio: one for tire makers and one for printers. In this example, 14 firms produce tires. The largest four have 80 percent of the sales, so the four- firm concentration ratio is 80 percent. In the printing industry, with 1,004 firms, the largest four firms have only 0.5 percent of the sales, so the four-firm concentration ratio is 0.5 percent.
A low concentration ratio indicates a high degree of competition, and a high concentration ratio indicates an absence of competition. A monopoly has a concentration ratio of 100 percent-the largest (andonly) firm has 100 percent of the sales. A four-firm concentration ratio that exceeds 60 percent is regarded as an indication of a market that is highly concentrated and dominated by a few firms in an oligopoly. A ratio of less than 60 percent is regarded as an indication of a competitive market.

The Herfindahl-Hirschman Index - also called the HHI - is the square of the percentage market share of each firm summed over the largest 50 firms (or summed over all the firms if there are fewer than 50) in a market. For example, if there are four firms in a market and the market shares of the firms are 50 percent, 25 percent, 15 percent, and 10 percent, the Herfindahl-Hirschman Index is:

$$
\mathrm{HHI}=50^{2}+25^{2}+15^{2}+10^{2}=3,450
$$

TABLE 10.5 Calculating the Four-Firm Concentration Ratio

| Tire makers |  | Printers |  |
| :--- | :---: | :--- | :---: |
| Firm | Sales <br> (millions of dollars) | Firm | Sales <br> (millions of dollars) |
| Top, Inc. | 200 | Fran's | 2.5 |
| ABC, Inc. | 250 | Ned's | 2.0 |
| Big, Inc. | 150 | Tom's | 1.8 |
| XYZ, Inc. | $\underline{100}$ | Jill's | $\underline{1.7}$ |
| Largest 4 firms | $\underline{175}$ | Largest 4 firms | 8.0 |
| Other 10 firms | $\underline{\underline{875}}$ | Other 1,000 firms | $\underline{1,592.0}$ |
| Industry | Industry | $\underline{\underline{1,600.0}}$ |  |

Four-firm concentration ratios:
Tire makers: $\frac{700}{875} \times 100=80$ percent $\quad$ Printers: $\frac{8}{1,600} \times 100=0.5$ percent

## ECONOMY IN ACTION

## Concentration in the U.S. Economy

The U.S. Department of Commerce calculates and publishes data showing concentration ratios and the HHI for each industry in the United States. The bars in the figure show the four-firm concentration ratio and the number at the end of each bar is the HHI .
Cigarette and battery manufacturing are two of the most concentrated industries. A very small number of firms dominate the markets for these products andtheir competitors are small firms that have small market shares. Glass containers, trucks, breakfast cereals, and major appliances are highly concentrated industries. They are oligopolies.
Cookies and crackers, motor vehicles, and soft drinks are moderately concentrated industries. They are examples of monopolistic competition.
Snack food, ice cream, milk, cheese, commercial printing, canvas bags manufacturing, quick printers, and bakers have low concentration measures and are highly competitive industries.
Concentration measures are useful indicators of the degree of competition in a market, but they must be supplemented by other information to determine the structure of the market. Newspapers and automobiles are examples of how the concentration measures give a misleading reading of the degree of competition. Most newspapers are local. They serve a single city or even a smaller area. So despite the low concentration measure, newspapers are concentrated in their own local areas. Automobiles are traded internationally and foreign cars are freely imported into the United States. Despite the moderately high U.S. concentration measure, the automobile industry is competitive.


## Concentration Measures in the United States

In perfect competition, the HHI is small. For example, if each of the largest 50 firms in an industry has a market share of 0.1 percent, then the HHI is
$0.1^{2} \times 50=0.5$. Ina monopoly, the HHI is 10,000 .
The firm has 100 percent of the market: $100^{2}=10,000$.
The HHI became a popular measure of the degree of competition during the 1980s, when the Justice Department used it to classify markets. A market in which the HHI is less than 1,500 is regarded as beingcompetitive. A market in which the HHI lies between1,500 and 2,500 is regarded as being moderately competitive. But a market in which the HHI exceeds 2,500 is regarded as being uncompetitive.

Table 10.6 summarizes the characteristics of the types of market structure, along with the measures of concentration and some examples of each type.

TABLE 10.6 Market Structure

| Characteristics | Perfect competition | Monopolistic competition | Oligopoly | Monopoly |
| :---: | :---: | :---: | :---: | :---: |
| Number of firms in industry | Many | Many | Few | One |
| Product | Identical | Differentiated | Either identical or differentiated | No close substitutes |
| Barriers to entry | None | None | Moderate | High |
| Firm's control over price | None | Some | Considerable | Considerable or regulated |
| Concentration ratio | 0 | Low | High | 100 |
| HHI (approx. ranges) | Close to 0 | Less than 2,500 | More than 2,500 | 10,000 |
| Examples | Wheat, corn | Food, clothing | Computer chips | Local water supply |

## Limitations of a Concentration Measure

The three main limitations of using only concentration measures as determinants of market structure are their failure to take proper account of

- The geographical scope of the market
- Barriers to entry and firm turnover
- The correspondence between a market and an industry

Geographical Scope of the Market Concentration measures take a national view of the market. Many goods are sold in a national market, but some are sold in a regional market and some in a global one. The concentration measures for newspapers are low, indicating competition, but in most cities the newspaper industry is highly concentrated. The concentrationmeasures for automobiles are high, indicating little competition, but the three biggest U.S. car makers compete with foreign car makers in a highly competitive global market.

Barriersto Entryand Firm Turnover Some markets are highly concentrated but entry is easy and the turnover of firms is large. For example, small townshave few restaurants, but no restrictions hinder a new restaurant from opening and many attempt to do so. Also, a market with only a few firms might be competitive because of potential entry. The few firms in a market face competition from the many potential firms that will enter the market if economic profit opportunities arise.

Market and Industry Correspondence To calculate concentration ratios, the Department of Commerce classifies each firm as being in a particularindustry. But markets do not always correspond closely to industries for three reasons.

First, markets are often narrower than industries. For example, the pharmaceutical industry, which has a low concentration ratio, operates in many separate markets for individual products - for example, measles vaccine and AIDS - fighting drugs. These drugs do not compete with each other, so this industry, which looks competitive, includes firms that are monopolies (or near monopolies) in markets for individual drugs. Second, most firms make several products. For example, Westinghouse makes electrical equipment and, among other things, gas-fired incinerators and plywood. So this one firm operates in at least three separate markets, but the Department of Commerce classifies Westinghouse as being in the electrical goods and equipment industry. The fact that Westinghouse competes with other producers of plywood does not show up in the concentration numbers for the ply- wood market.
Third, firms switch from one market to another depending on profit opportunities. For example, Motorola, which today produces cellular telephones and other communications products, has diversified from being a TV and computer chip maker. Motorola no longer produces TVs. Publishers of newspapers, magazines, and textbooks are today rapidly diversifying into Internet and multimedia products. These switches among markets show that there is much scope for entering and exiting a market, and so measures of concentration have limited usefulness.
Despite their limitations, concentration measures do provide a basis for determining the degree of competition in a market when they are combined with information about the geographical scope of the market, barriers to entry, and the extent to which large, multiproduct firms straddle a variety of markets.

## Produce or Outsource? Firms and Markets

To produce a good or service, even a simple one such as a shirt, factors of production must be hired and their activities coordinated. To produce a good as complicated as an iPhone, an enormous range of specialist factors of production must be coordinated.
Factors of production can be coordinated either by firms or markets. We'll describe these two ways oforganizing production and then see why firms play acrucial role in achieving an efficient use of resources.

## Firm Coordination

Firms hire labor, capital, and land, and by using amixture of command systems and incentive systems (see p. 267) organize and coordinate their activities toproduce goods and services. Firm coordination occurs when you take your car to the garage for service such as an oil change and brake check. The garage owner hires a mechanic and tools and coordinates all the activities that get your carserviced. Firms also coordinate the production of cornflakes, golf clubs, and a host of other items.

## Market Coordination

Markets coordinate production by adjusting prices and making the decisions of buyers and sellers of factors of production and components consistent.
Market coordination occurs to produce a rock con- cert. A promoter books a stadium, rents some stage equipment, hires some audio and video recording engineers and technicians, and
engages some rockgroups, a superstar, a publicity agent, and a ticket agent. The promoter sells tickets to thousands of rock fans, audio rights to a recording company, and video and broadcasting rights to a television network. All these transactions take place in markets that coordinate the buying and selling of this huge variety of factors of production. Outsourcing, buying parts or products from other firms, is another example of market coordination.

## Why Firms?

What determines whether a firm or a market coordinates a particular set of activities? How does a firm decide whether to buy an item from another firm ormanufacture it itself? The answer is cost. Taking account of the opportunity cost of time as well as thecosts of the other inputs, a firm uses the method thatcosts least. In other words, it uses the economically efficient method.
If a task can be performed at a lower cost by markets than by a firm, markets will do the job, and any attempt to set up a firm to replace such market activity will be doomed to failure. Firms coordinate economic activity when a task can be performed more efficiently by a firm than bymarkets. In such a situation, it is profitable to set up a firm. Firms are often more efficient than markets as coordinators of economic activity because they can achieve

- Lower transactions costs
- Economies of scale
- Economies of scope
- Economies of team production

Lower Transactions Costs Firms eliminate transactions costs. Transactions costs are the costs that arise from finding someone with whom to do business, of reaching an agreement about the price and other aspects ofthe exchange, and of ensuring that the terms of the agreement are fulfilled. Market transactions require buyers and sellers to get together and to negotiate the terms and conditions of their trading. Sometimes, lawyers have to be hired to draw up contracts. A brokencontract leads to still more expense. A firm can lower such transactions costs by reducing the number of individual transactions undertaken. Imagine getting your car fixed using market coordination. You hire a mechanic to diagnose the problems and make a list of the parts and tools needed to fix them. You buy the parts from several dealers, rentthe tools from ABC Rentals, hire an auto mechanic, return the tools, and pay your bills. You can avoid all these transactions and the time they cost you by let-ting your local garage fix the car.

Economies of Scale When the cost of producing a unit of a good falls as its output rate increases, economies of scale exist. An automaker experiences economies of scale because as the scale of productionincreases, the firm can use cost-saving equipment and highly specialized labor. An automaker that produces only a few cars a year must use hand-tool methods that are costly. Economies of scale arise from specialization and the division of labor that can be reaped more effectively by firm coordination rather than market coordination.

Economies of Scope A firm experiences economies of scope when it uses specialized (and often expensive) resources to produce a range of goods and services. For example, Toshiba uses its designers and specialized equipment to make the hard drive for the iPod. But itmakes many different types of hard drives and other related products. As a result, Toshiba produces the iPod hard drive at a lower cost than a firm making only the iPod hard drive could achieve.

Economies of Team Production A production process in which the individuals in a group specialize inmutually supportive tasks is team production. Sports provide the best examples of team activity. In base- ball, some team members specialize in pitching and others in fielding. In basketball, some team membersspecialize in defense and some in offense. The production of goods and services offers many examples of team activity. For example, production lines in a TV manufacturing plant work most efficiently when individual activity is organized in teams, each worker specializing in a few tasks. You can also think of an entire firm as being a team. The team has buyers ofraw materials and other inputs, production workers, and salespeople. Each individual member of the team specializes, but the value of the output of the team and the profit that it earns depend on the coordinated activities of all the team's members. Because firms can economize on transactions costs, reap economies of scale and economies of scope, and organize efficient team production, it is firms rather than markets that coordinate most of our economic activity.

## ECONOMICS IN ACTION

## Apple Doesn't Make the iPhone

Apple designed the iPhone and markets it, but Apple doesn't manufacture it. Why? Apple wants to produce the iPhone at the lowest possible cost. Apple achieves its goal by assigning the production task to more than 30 firms, some of which are listed in thetable opposite. These 30 firms produce the components in Asia, Europe, and North America and then the components are assembled in its sleek, iconic case by Foxconn and Quanta in Taiwan. Most electronic products - TVs, DVD players, iPods and iPads, and personal computers are produced in a similar way to the iPhone with a combination of firm and market coordination. Hundreds oflittle-known firms compete fiercely to get their components into well-known consumer products.

## ECONOMIC ANALYSIS

- Like all firms, Facebook and Google aim to maximize profit.
- Facebook provides social networking services and Google provides search services, a variely of other services, and with Google+ is offering a social networking service.
- Facebook and Google face constraints imposed by the market and technology.
- People who use social networks demand their services, and at the latest count Facebook and another 200 -odd firms supply social networking services.
- People looking for information demand Internet search services, and Google and more than 100 other firms supply Internet search services.
- The equilibrium price of social networking services and of Internet search services is zero, and the equilibrium quantity of each is the quantity demanded at a zero price.
- Social network and Internet search providers enjoy economies of scope: They produce advertising services as well as their other service.
- Unlike social networking and search, Internet advertising is a big revenue and profit generator.
- Because the providers of social networking and search know a lot about their users, they can offer advertisers access to potential customers and charge a high price for this precision.
- Google has been enormously successful at delivering advertising based on a user's search activity, and its revenue has grown from $\$ 1$ billion in 2003 to $\$ 56$ billion in 2013 (see Fig. 1). Google's profit in 2013 was $\$ 13$ billion (see Fig. 2).
- Facebook is still learning how to tap its advertising potential and the news article describes its innovation in 2012: real-time bidding for advertising based on a user's browsing.
- Facebook's revenue is beginning to grow, but by 2013 it had reached only $\$ 8$ billion (see Fig. 1).
- Providing a social networking service or search service doesn't guarantee success in generating advertising revenue and profit.
- Yahoo is an example of a firm that hasn't performed as well as its owners would wish.
- As Google and Facebook have seen explosive growth in users and revenues, Yahoo has struggled.


Figure 1 Total Revenue Comparison


## Figure 2 Profit Comparison

- Figure 1 shows that Yahoo's revenue peaked in 2008 and has been falling while Google's has soared and Facebook's has grown to overtake that of Yahoo.
- Figure 2 shows that Yahoo's profit has been flat while Google's has soared. Facebook's profit remains modest.
- The data shown in Figs. 1 and 2 suggest that so far, Internet search is a more effective tool for generating revenue and profit than social networking. Perhaps Facebook's new revenue model will change that.
- The data also suggest that Facebook's and Google's expansion is tightening the market constraint that Yahoo faces.


## CHAPTER 11: Output and Costs

After studying this chapter, you will be able to:

- Distinguish between the short run and the long run
- Explain and illustrate a firm's short-run product curves
- Explain and derive a firm's short-run cost curves
- Explain and derive a firm's long-run average cost curve

Behind the scenes of your favorite Starbucks coffee shop, many economic decisions have been made that affect the firm's cost of production. Starbucks has decided how much to produce, how many people toemploy, and how much and what type of equipmentto use. How does a firm make these decisions?
We are going to answer this question in this chapter. And in Economics in the News at the end of thechapter, we'll look at how recent expansion decisions by Starbucks affect the firm's production costs. But first, we'll study the costs of a simpler, smaller firm, Campus Sweaters, a (fictional) producer of knitwear.

## Decision Time Frames

People who operate firms make many decisions, and all of their decisions are aimed at achieving one overriding goal: maximum attainable profit. But not all decisions are equally critical. Some decisions are big ones. Once made, they are costly (or impossible) to reverse. If such a decision turns out to be incorrect, it might lead to the failure of the firm. Other decisions are small. They are easily changed. If one of these decisions turns out to be incorrect, the firm can change its actions and survive.
The biggest decision that an entrepreneur makes is in what industry to establish a firm. For most entrepreneurs, their background knowledge and interests drive this decision. But the decision also depends on profit prospects - on the expectation that total revenue will exceed total cost.
Cindy has decided to set up Campus Sweaters. She has also decided the most effective method of organizing the firm. But she has not decided the quantity to produce, the factors of production to hire, or the price to charge for sweaters.
Decisions about the quantity to produce and the price to charge depend on the type of market in which the firm operates. Perfect competition, monopolistic competition, oligopoly, and monopoly all confront the firm with different problems.
Decisions about how to produce a given output do not depend on the type of market in which the firmoperates. All types of firms in all types of markets make similar decisions about how to produce.
The actions that a firm can take to influence the relationship between output and cost depend on how soon the firm wants to act. A firm that plans to change its output rate tomorrow has fewer options than one that plans to change its output rate six months or six years in the future.
To study the relationship between a firm's output decision and its costs, we distinguish between two decision time frames:

- The short run
- The long run


## The Short Run

The short run is a time frame in which the quantity of at least one factor of production is fixed. For mostfirms, capital, land, and entrepreneurship are fixed factors of production and labor is the variable factor of production. We call the fixed factors of production the firm'splant. In the short run, a firm's plant is fixed.
For Campus Sweaters, the fixed plant is its factory building and its knitting machines. To increase output in the short run, a firm must increase the quantity of a variable factor of production, which is usually labor. So to produce more output, Campus Sweaters must hire more labor and operate its knitting machines for more hours a day. Short-run decisions are easily reversed. The firm can increase or decrease its output in the short run byincreasing or decreasing the amount of labor it hires.

## The Long Run

The long run is a time frame in which the quantities of all factors of production can be varied. That is, the long run is a period in which the firm can change its plant. To increase output in the long run, a firm can change its plant as well as the quantity of labor it hires.
Campus Sweaters can decide whether to install more knitting machines, use a new type of machine, reorganize its management, or hire more labor. Long-run decisions are not easily reversed. Once a plant decision is made, the firm usually must live with it for some time. To emphasize this fact, we call the past expenditure on a plant that has no resale value a sunk cost. A sunk cost is irrelevant to the firm's current decisions. The only costs that influence its current decisions arethe short-run cost of changing its labor inputs and thelong-run cost of changing its plant.

## Short-Run Technology Constraint

To increase output in the short run, a firm must increase the quantity of labor employed. We describe the relationship between output and the quantity of labor employed by using three related concepts:

1. Total product
2. Marginal product
3. Average product

These product concepts can be illustrated either by product schedules or by product curves. Let's look first at the product schedules.

## Product Schedules

Table 11.1 shows some data that describe Campus Sweaters' total product, marginal product, and aver- age product. The numbers tell us how the quantity of sweaters produced increases as Campus Sweaters employs more workers. The numbers also tell us about the productivity of the labor that CampusSweaters employs.

Focus first on the columns headed "Labor" and "Total product." Total product is the maximum output that a given quantity of labor can produce. You can see from the numbers in these columns that as Campus Sweaters employs more labor, total product increases. For example, when 1 worker is employed, total product is 4 sweaters a day, and when 2 workers are employed, total product is 10 sweaters a day. Each increase in employment increases total product.
The marginal product of labor is the increase in total product that results from a one-unit increase in the quantity of labor employed, with all other inputs remaining the same. For example, in Table 11.1, when Campus Sweaters increases employment from 2 to 3 workers and does not change its capital, the marginal product of the third worker is 3 sweaterstotal product increases from 10 to 13 sweaters.
Average product tells how productive workers are on average. The average product of labor is equal to total product divided by the quantity of labor employed. For example, in Table 11.1, the average product of 3 workers is 4.33 sweaters per worker- 13 sweaters a day divided by 3 workers.
If you look closely at the numbers in Table 11.1, you can see some patterns. As Campus Sweaters hires more labor, marginal product increases initially, and then begins to decrease. For example, marginal product increases from 4 sweaters a day for the first worker to 6 sweaters a day for the second worker and then decreases to 3 sweaters a day for the third worker.
Average product also increases at first and then decreases. You can see the relationships between the quantity of labor hired and the three product concepts more dearly by looking at the product curves.

## Product Curves

The product curves are graphs of the relationshipsbetween employment and the three product concepts you've just studied. They show how total product, marginal product, and average product change as employment changes. They also show the relationships amongthe three concepts. Let's look at the product curves.

## Total Product Curve

Figure 11.1 shows Campus Sweaters' total product curve, $T P$, which is a graph of the total product schedule. Points $A$ through $F$ correspond to rows $A$ through Fin Table 11.1. To graph the entire total product curve, we vary labor by hours rather than whole days.
Notice the shape of the total product curve. As employment increases from zero to 1 worker a day, the curve becomes steeper. Then, as employment increases to 3, 4, and 5 workers a day, the curve becomes less steep.
The total product curve is similar to the production possibilities frontier (explained in Chapter 2). It separates the attainable output levels from those that are unattainable. All the points that lie above the curveare unattainable. Points that lie below the curve, inthe orange area, are attainable, but they are inefficient - they use more labor than is necessary to produce a given output. Only the points on the total product curve are technologically efficient.

FIGURE 11.1 Total Product Curve


The total product curve, $T P$, is based on the data in Table 11.1. The total product curve shows how the quantity of sweaters produced changes as the quantity of labor employed changes. For example, 2 workers can produce 10 sweaters a day (point $C$ ). Points $A$ through $F$ on the curve correspond to the rows of Table 11.1. The total product curve separates attainable outputs from unattainable outputs. Points below the TP curve are inefficient.

## Marginal Product Curve

Figure 11.2 shows Campus Sweaters' marginal product of labor. Part (a) reproduces the total product curve from Fig. 11.1 and part (b) shows the marginalproduct curve, MP. In part (a), the orange bars illustrate the marginal product of labor. The height of a bar measures marginal product. Marginal product is also measured by the slope of the total product curve. Recall that the slope of a curve is the change in the value of the variable measured on the $y$-axis-output-divided by the change in the variable measured on the $x$-axis- labor-as we move along the curve. A one-unit increase in labor, from 2 to 3 workers, increases out- putfrom 10 to 13 sweaters, so the slope from point $C$ to point $D$ is 3 sweaters per additional worker, the same as the marginal product we've just calculated.
Again varying the amount of labor in the smallest units possible, we can draw the marginal product curve shown in Fig. II.2(b). The height of this curve measures the slope of the total product curve at a point. Part (a) shows that an increase in employment from 2 to 3 workers increases output from 10 to 13 sweaters (an increase of 3). The increase in output of 3 sweaters appears on the $y$-axis of part (b) as the marginal product of going from 2 to 3 workers. We plot that marginal product at the midpoint between 2 and 3 workers. Notice that the marginal product shown in Fig. I 1.2(b) reaches a peak at 1.5 workers, and at that
point, marginal product is 6 sweaters per additional worker. The peak occurs at 1.5 workers because the total product curve is steepest when employment increases from 1 worker to 2 workers.

## TABLE 11.1 Total Product, Marginal Product, and Average Product

|  | Labor (workers per dayl | Total product (sweaters per day) | Marginal product (sweaters per additionol worker) | Average product (sweaters per worker) |
| :---: | :---: | :---: | :---: | :---: |
| A | 0 | 0 |  |  |
| B | 1 | 4 |  | 4.00 |
| C | 2 | 10 |  | 5.00 |
| D | 3 | 13 |  | 4.33 |
| E | 4 | 15 |  | 3.75 |
| F | 5 | 16 |  | 3.20 |

Total product is the total amount produced. Marginal product is the change in total product that results from a one-unit increase in labor. For example, when labor increases from 2 to 3 workers a day (row $C$ to row $D$ ), total product increases from 10 to 13 sweaters a day. The marginal product of going from 2 to 3 workers is 3 sweaters. Average product is total product divided by the quantity of labor employed. For example, the average product of 3 workers is 4.33 sweaters per worker ( 13 sweaters a day divided by 3 workers).

The total product and marginal product curves differ across firms and types of goods. GM's productcurves are different from those of PennPower, whosecurves in turn are different from those of Campus Sweaters. But the shapes of the product curves are similar because almost every production process hastwo features:

- Increasing marginal returns initially
- Diminishing marginal returns eventually

Increasing Marginal Returns Increasingmarginal returns occur when the marginal product of an additional worker exceeds the marginal product of the previous worker. Increasing marginal returns arisefrom increased specialization and division of labor inthe production process.

For example, if Campus Sweaters employs one worker, that person must learn all the aspects of sweater production: running the knitting machines, fixing breakdowns, packaging and mailing sweaters, buying and checking the type and color of the wool. All these tasks must be performed by that one person.
If Campus Sweaters hires a second person, the two workers can specialize in different parts of the production process and can produce more than twice as much as one worker. The marginal product of the second worker is greater than the marginal product of the first worker. Marginal returns are increasing.

Diminishing Marginal Returns Most production processes experience increasing marginal returns initially, but all production processes eventually reach a point of diminishing marginal returns. Diminishing marginal returns occur when the marginal product of an additional worker is less than the marginal product of the previous worker. Diminishing marginal returns arise from the fact that more and more workers are using the same capital and working in the same space. As more workersare added, there is less and less for the additional workers to do that is productive. For example, if Campus Sweaters hires a third worker, output increases but not by as much as it did when it hiredthe second worker. In this case, after two workers arehired, all the gains from specialization and the division of labor have been exhausted. By hiring a thirdworker, the factory produces more sweaters, but the equipment is being operated closer to its limits. There are even times when the third worker has nothing to do because the machines are running without the need for further attention. Hiring more and more workers continues to increase output but by successively smaller amounts. Marginal returns are diminishing. This phenomenon is such a pervasive one that it is called a "law" - the law of diminishing returns. The law of diminishing returns states that

As a firm uses more of a variable factor of production with a given quantity of the fixed factor of production, the marginal product of the variable factor eventually diminishes. You are going to return to the law of diminishing returns when we study a firm's costs, but before we do that, let's look at the average product of labor and the average product curve.

## Average Product Curve

Figure 11.3 illustrates Campus Sweaters' average product of labor and shows the relationship between average product and marginal product. Points $B$ through $F$ on the average product curve AP correspond to those same rows in Table 11.1. Average product increases from 1 to 2 workers (its maximum value at point $C$ ) but then decreases as yet more workers are employed. Notice also that average product is largest when average product and marginal product are equal. That is, the marginal product curve cuts the average product curve at the point of maximum average product. For the number of workers at which marginal product exceeds average product, average product is increasing. For the number of workers at which marginal product is less than average product, average product is decreasing.

The relationship between the average product and marginal product is a general feature of the relationship between the average and marginal values of any variable-even your grades.

FIGURE 11.3 Average Product


The figure shows the average product of labor and the connection between average product and marginal product. With 1 worker, marginal product exceeds average product, so average product is increasing. With 2 workers, marginal product equals average product, so average product is at its maximum. With more than 2 workers, marginal product is less than average product, so average product is decreasing.

## Short-Run Cost

To produce more output in the short run, a firm mustemploy more labor, which means that it must increaseits costs. We describe the relationship between outputand cost by using three cost concepts:

- Total cost
- Marginal cost
- Average cost


## Total Cost

A firm's total cost (TC) is the cost of all the factors of production it uses. We separate total cost into total fixed cost and total variable cost. Total fixed cost (TFC) is the cost of the firm's fixed factors. For Campus Sweaters, total fixed cost includes the cost of renting knitting machines and normal profit, which is the opportunity cost of Cindy's entrepreneurship (see Chapter 10, p. 263). The quantities of fixed factors don't change as out- put changes, so total fixed cost

FIGURE 11.2 Total Product and Marginal Product

(a) Total product

(b) Marginal product

Marginal product is illustrated by the orange bars. For example, when labor increases from 2 to 3 workers a day, marginal product is the orange bar whose height is 3 sweaters. (Marginal product is shown midway between the quantities of labor to emphasize that marginal product results from changing the quantily of labor.) The steeper the slope of the total product curve (TP) in part (a), the larger is marginal product (MP) in part (b). Marginal product increases to a maximum (in this example when 1.5 workers a day are employed) and then declines-diminishing marginal product.

## ECONOMICS IN ACTION

## How To Pull Up Your Average

Do you want to pull up your average grade? Then make sure that your grade this semester is better than your current average! This semester is your marginal semester. If your marginal grade exceeds your average grade (like the second semester in the figure), your average will rise. If your marginal grade equals your average grade (like the third semester in the figure), your average won't change. If your marginal grade is below your average grade (like the fourth semester in the figure), your average will fall.
The relationship between your marginal and average grades is exactly the same as that between marginal product and average product.

is the same at all outputs. Total variable cost (TVC) is the cost of the firm's variable factors. For Campus Sweaters, labor is the variable factor, so this component of cost is its wage bill. Total variable cost changes as output changes Total cost is the sum of total fixed cost and total variable cost. That is,

$$
T C=T F C+T V C .
$$

The table in Fig. 11.4 shows total costs. Campus Sweaters rents one knitting machine for $\$ 25$ a day, soits TFC is $\$ 25$. To produce sweaters, the firm hires labor, which costs $\$ 25$ a day. TVC is the number of workers multiplied by $\$ 25$. For example, to produce 13 sweaters a day, in row $D$, the firm hires 3 workers and TVCis $\$ 75$. TCis the sum of TFC and TVC, soto produce 13 sweaters a day, TC is $\$ 100$. Check the calculations in the other rows of the table.
Figure 11.4 shows Campus Sweaters' total cost curves, which graph total cost against output. The TFC curve is horizontal because total fixed cost ( $\$ 25$ a day) does not change when output changes. The purple TVC curve and the blue TC curve both slope upward because to increase output, more labor must be employed, which increases total variable
cost. Total fixed cost equals the vertical distance between the TVC and TC curves. Let's now look at a firm's marginal cost.

FIGURE 11.4 Total Cost Curves


| Labor (workers per day) |  | Outpur (sweaters per day) | Total fixed cost <br> (TFC) | $\qquad$ | Total cost (TC) |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | (dollars per day) |
| A | 0 |  | 0 | 25 | 0 | 25 |
| B | 1 | 4 | 25 | 25 | 50 |
| C | 2 | 10 | 25 | 50 | 75 |
| D | 3 | 13 | 25 | 75 | 100 |
| E | 4 | 15 | 25 | 100 | 125 |
| F | 5 | 16 | 25 | 125 | 150 |

Campus Sweaters rents a knitting machine for \$25 a day, so this cost is the firm's total fixed cost. The firm hires workers at a wage rate of $\$ 25$ a day, and this cost is its total variable cost. For example, in row $D$, Campus Sweaters employs 3 workers and its total variable cost is $3 \times \$ 25$, which equals $\$ 75$. Total cost is the sum of total fixed cost and total variable cost. For example, when Campus Sweaters employs 3 workers, total cost is $\$ 100$-total fixed cost of $\$ 25$ plus total variable cost of $\$ 75$.
The graph shows Campus Sweaters' total cost curves. Total fixed cost is constant - the TFC curve is a horizontalline. Total variable cost increases as output increases, so the TVC curve and the TC curve increase as output increases. The vertical distance between the TC curve and the TVC curve equals total fixed cost, as illustrated by the two arrows.

## Marginal Cost

Figure 11.4 shows that total variable cost and total cost increase at a decreasing rate at small outputs but eventually, as output increases, total variable cost and total cost increase at an increasing rate. To understand this pattern in the change in total costas output increases, we need to use the concept of marginal cost.
A firm's marginal cost is the increase in total cost that results from a one-unit increase in output. Wecalculate marginal cost as the increase in total costdivided by the increase in output. The table in Fig. 11.5 shows this calculation. When, for example, output increases from 10 sweaters to 13 sweaters, total cost increases from $\$ 75$ to $\$ 100$. The change in output is 3 sweaters, and the change in total cost is $\$ 25$. The marginal cost of one of those 3 sweaters is ( $\$ 25 \ldots, 3$ ), which equals $\$ 8.33$.
Figure 11.5 graphs the marginal cost data in the table as the red marginal cost curve, MC. This curveis U-shaped because when Campus Sweaters hires asecond worker, marginal cost decreases, but when ithires a third, a fourth, and a fifth worker, marginal cost successively increases.
At small outputs, marginal cost decreases as output increases because of greater specialization and the division of labor. But as output increases further, marginal cost eventually increases because of the law of diminishing returns. The law of diminishing returns means that the output produced by each additional worker is successively smaller. To produce an additional unit of output, ever more workers are required, and the cost of producing the additional unit of out- put-marginal cost-must eventually increase. Marginal cost tells us how total cost changes as output increases. The final cost concept tells us whatit costs, on average, to produce a unit of output. Let's now look at Campus Sweaters' average costs.

## Average Cost

There are three average costs of production:

1. Average fixed cost
2. Average variable cost
3. Average total cost

Average fixed cost (AFC) is total fixed cost per unit of output. Average variable cost (AVC) is totalvariable cost per unit of output. Average total cost (ATC) is total cost per unit of output. The average cost concepts are calculated from the total cost concepts as follows:

$$
T C=T F C+T V C .
$$

Divide each total cost term by the quantity produced, $Q$, to get

$$
\frac{T C}{Q}=\frac{T F C}{Q}+\frac{T V C}{Q}
$$

or

$$
A T C=A F C+A V C
$$

The table in Fig. 11.5 shows the calculation of average total cost. For example, in row C , output is 10 sweaters. Average fixed cost is ( $\$ 25 \ldots, . . .10$ ), which equals $\$ 2.50$, average variable cost is ( $\$ 50 \ldots, . . .10$ ), which equals $\$ 5.00$, and average total cost is ( $\$ 75 \ldots, 10$ ), which equals $\$ 7.50$. Note that average total cost is equal to average fixed cost ( $\$ 2.50$ ) plus average variable cost (\$5.00).
Figure 11.5 shows the average cost curves. The green average fixed cost curve (AFC) slopes down- ward. As output increases, the same constant total fixed cost is spread over a larger output. The blue average total cost curve (ATC) and the purple average variable cost curve (AVC) are U-shaped. The vertical distance between the average total cost and average variable cost curves is equal to average fixed cost - as indicated by the two arrows. That distance shrinks as output increases because average fixed cost declines with increasing output.

FIGURE 11.5 Marginal Cost and Average Costs


|  | Labor (workers per day) | Output <br> (sweaters per day) | Total fixed cost <br> (TFC) | Total variable cost (TVC) | Total cost (TC) | Marginal <br> cost <br> (MC) <br> (dollars per | Average fixed cost (AFC) | Average variable cost (AVC) | Average total cost (ATC) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | (dollars per day) |  |  | additional sweater) | (dollars per sweater) |  |  |
| A | 0 | 0 | 25 | 0 | 25 |  | - | - | - |
| B | 1 | 4 | 25 | 25 | 50 |  | 6.25 | 6.25 | 12.50 |
| C | 2 | 10 | 25 | 50 | 75 |  | 2.50 | 5.00 | 7.50 |
| D | 3 | 13 | 25 | 75 | 100 |  | 1.92 | 5.77 | 7.69 |
| $E$ | 4 | 15 | 25 | 100 | 125 |  | 1.67 | 6.67 | 8.33 |
| $F$ | 5 | 16 | 25 | 125 | 150 |  | 1.56 | 7.81 | 9.38 |

## Marginal Cost and Average Cost

The marginal cost curve ( MC ) intersects the averagevariable cost curve and the average total cost curve at their minimum points. When marginal cost is less than average cost, average cost is decreasing, and when marginal cost exceeds average cost, average cost is increasing. This relationship holds for both the ATC curve and the AVC curve. It is another example of the
relationship you saw in Fig. 11.3 for average product and marginal product and in your average and marginal grades.

## Why the Average Total Cost Curve is U-Shaped

Average total cost is the sum of average fixed cost and average variable cost, so the shape of the ATC curve combines the shapes of the AFC and AVC curves. The $U$ shape of the ATC curve arises from the influence of two opposing forces:

1. Spreading total fixed cost over a larger output
2. Eventually diminishing returns

When output increases, the firm spreads its total fixed cost over a larger output and so its average fixed cost decreases-its AFC curve slopes downward.

Diminishing returns means that as output increases, ever-larger amounts of labor are needed to produce an additional unit of output. So as output increases, average variable cost decreases initially but eventually increases, and the AVC curve slopes upward. The AVC curve is U-shaped.
The shape of the ATC curve combines these two effects. Initially, as output increases, both average fixed cost and average variable cost decrease, so average total cost decreases. The ATC curve slopes downward.
But as output increases further and diminishing returns set in, average variable cost starts to increase. With average fixed cost decreasing more quickly than average variable cost is increasing, the ATC curve continues to slope downward. Eventually, average variable cost starts to increase more quickly thanaverage fixed cost decreases, so average total cost starts to increase. The ATC curve slopes upward.

## Cost Curves and Product Curves

The technology that a firm uses determines its costs. A firm's cost curves come directly from its product curves. You've used this link in the tables in which we have calculated total cost from the total product schedule and information about the prices of the factors of production. We're now going to get a clearer view of the link between the product curves and the cost curves. We'll look first at the link between total cost and total product and then at the links between the average and marginal product and cost curves.

Total Product and Total Variable Cost Figure 11.6 shows the links between the firm's total product curve, TP, and its total variable cost curve, TVC. The graph is a bit unusual in two ways. First, it measures two variables on the x-axis-labor and variablecost. Second, it graphs the TVC curve but with variable cost on the $x$-axis and output on the $y$-axis. The graph can show labor and cost on the x -axis because variable cost is proportional to labor. One worker costs \$25 a day. Graphing output against labor gives the TP curve and graphing variable cost against output gives the TVC curve.

FIGURE 11.6 Total Product and Total Variable Cost


The figure shows the total product curve, $T P$, as a graph of output (sweaters per day) plotted against labor (workers per day). It also shows the total variable cost curve, TVC, as a graph of total variable cost (dollars per day) against output. The only difference between the TVC curve here and that in Fig. 11.4 is that we've switched the $x$-axis and $y$-axis.

Average and Marginal Product and Cost Figure 11.7 shows the links between the firm's average and marginal product curves and its average and marginal cost curves. The upper graph shows the average product curve, AP, and the marginal product curve MP - like those in Fig. 11.3. The lower graph shows the average variable cost curve, $A V C$, and the marginal cost curve, MC-like those in Fig. 11.5.
As labor increases up to 1.5 workers a day (upper graph), output increases to 6.5 sweaters a day (lower graph). Marginal product and average product rise and marginal cost and average variable cost fall. At the point of maximum marginal product, marginal cost is at a minimum. As labor increases from 1.5 workers to 2 workers a day (upper graph), output increases from 6.5 sweaters to 10 sweaters a day (lower graph). Marginal product falls and marginal cost rises, but average product continues to rise and average variable cost continues to fall. At the point of maximum average product, aver- age variable cost is at a minimum. As labor increases further, output increases. Average product diminishes and average variable cost increases.

## Shifts in the CostCurves

The position of a firm's short-run cost curves depends on two factors:

- Technology
- Prices of factors of production


## FIGURE 11.7 Average and Marginal Product

 Curves and Cost Curves

A firm's MP curve is linked to its MC curve. If, as the firm increases its labor from 0 to 1.5 workers a day, the firm's marginal product rises, its marginal cost falls. If marginal product is at a maximum, marginal cost is at a minimum. If, as the firm hires more labor, its marginal product diminishes, its marginal cost rises.

A firm's AP curve is linked to its AVC curve. If, as the firm increases its labor to 2 workers a day, its average product rises, its average variable cost falls. If average product is at a maximum, average variable cost is at a minimum. If, as the firm hires more labor, its average product diminishes, its average variable cost rises.

Technology A technological change that increases productivity increases the marginal product and aver- age product of labor. With a better technology, the same factors of production can produce more output, so the technological advance lowers the costs of production and shifts the cost curves downward.
For example, advances in robot production techniques have increased productivity in the automobileindustry. As a result, the product curves of Chrysler, Ford, and GM have shifted upward and their cost curves have shifted downward. But the relationships between their product curves and cost curves have not changed. The curves are still linked in the way shown in Figs. 11.6 and 11.7. Often, as in the case of robots producing cars, a technological advance results in a firm using more capital, a fixed factor, and less labor, a variable factor.

Another example is the use of ATMs by banks to dispense cash. ATMs, which are fixed capital, have replaced tellers, which are variable labor. Such a technological change decreases total cost but increases fixed costs and decreases variable cost. This change in the mix of fixed cost and variable cost means that at small outputs, average total cost might increase, while at large outputs, average total cost decreases.

Prices ofFactors of Production An increase in the price of a factor of production increases the firm's costs and shifts its cost curves. How the curves shift depends on which factor price changes.
An increase in rent or some other component of fixed cost shifts the TFC and AFC curves upward and shifts the TC curve upward but leaves the AVC and TVC curves and the MC curve unchanged. For example, if the interest expense paid by a trucking company increases, the fixed cost of transportation services increases.
An increase in wages, gasoline, or another component of variable cost shifts the TVC and AVC curves upward and shifts the MC curve upward but leaves the AFC and TFC curves unchanged. For example, if truck drivers' wages or the price of gasoline increases, the variable cost and marginal cost of transportation services increase.

You've now completed your study of short-run costs. All the concepts that you've met are summarized in a compact glossary in Table 11.2.

## Long-Run Cost

We are now going to study the firm's long-run costs. In the long run, a firm can vary both the quantity of labor and the quantity of capital, so in the long run, all the firm's costs are variable.
The behavior of long-run cost depends on the firm's production function, which is the relationship between the maximum output attainable and thequantities of both labor and capital.

TABLE 11.2 A Compact Glossary of Costs

| Term | Symbol | Definition | Equation |
| :---: | :---: | :---: | :---: |
| Fixed cost |  | Cost that is independent of the output level; cost of a fixed factor of production |  |
| Variable cost |  | Cost that varies with the output level; cost of a variable factor of production |  |
| Total fixed cost | TFC | Cost of the fixed factors of production |  |
| Total variable cost | TVC | Cost of the variable factors of production |  |
| Total cost | TC | Cost of all factors of production | $T C=T F C+T V C$ |
| Output (total product) | TP | Total quantity produced (output $Q$ ) |  |
| Marginal cost | MC | Change in total cost resulting from a oneunit increase in total product | $M C=\Delta T C \div \Delta Q$ |
| Average fixed cost | AFC | Total fixed cost per unit of output | $A F C=T F C \div Q$ |
| Average variable cost | AVC | Total variable cost per unit of output | $A V C=T V C \div Q$ |
| Average total cost | ATC | Total cost per unit of output | $A T C=A F C+A V C$ |

## The Production Function

Table 11.3 shows Campus Sweaters' production function. The table lists total product schedules for four different quantities of capital. The quantity of capital identifies the plant size. The numbers for plant 1 are for a factory with 1 knitting machine - the case we've just studied. The other three plants have 2, 3, and 4 machines. If Campus Sweaters uses plant 2 with 2 knitting machines, the various amounts of labor can produce the outputs shown in the second column of the table. The other two columns show the outputs of yet larger quantities of capital. Each column of the table could be graphed as a total product curve for each plant. Diminishing Returns Diminishing returns occur with each of the four plant sizes as the quantity of laborincreases. You can check that fact by calculating themarginal product of labor in each of the plants with 2,3 , and 4 machines. With each plant size, as the firm increases the quantity of labor employed, the marginal product of labor (eventually) diminishes.

Diminishing Marginal Product of Capital Diminishing returns also occur with each quantity of labor as the quantity of capital increases. You can check that fact by calculating the marginal product of capital at a given quantity of labor. The marginal product of capital is the change in total product divided by the change in capital when the quantity of labor is constant-equivalently, the change in output resulting from a one-unit increase in the quantity of capital. For example, if Campus Sweaters has 3 workers and increases its capital from 1 machine to 2 machines, output increases from 13 to 18 sweaters a day. The marginal product of the second machine is 5 sweaters a day. If Campus Sweaters continues
to employ 3 workers and increases the number of machines from 2 to 3 , output increases from 18 to 22 sweaters a day. The marginal product of the third machine is 4 sweaters a day, down from 5 sweaters a day for the second machine. Let's now see what the production function implies for long-run costs.

## TABLE 11.3 The Production Function

|  | Output <br> (sweaters per day) |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
| Labor <br> (werkers per day) | Plant 1 | Plant 2 | Plant 3 | Plant 4 |
| 1 | 4 | 10 | 13 | 15 |
| 2 | 10 | 15 | 18 | 20 |
| 3 | 13 | 18 | 22 | 24 |
| 4 | 15 | 20 | 24 | 26 |
| 5 | 16 | 21 | 25 | 27 |
| Knitting machines <br> (number) | $\mathbf{1}$ | $\mathbf{2}$ | 3 | 4 |

The table shows the total product data for four quantities of capital (plant sizes). The greater the plant size, the larger is the output produced by any given quantity of labor. For a given plant size, the marginal product of labor diminishes as more labor is employed. For a given quantity of labor, the marginal product of capital diminishes as the quantity of capital used increases.

## Short-Run and Long-Run Cost

$A B$ before, Campus Sweaters can hire workers for $\$ 25$ a day and rent knitting machines for $\$ 25$ a day. Using these factor prices and the data in Table 11.3, we can calculate the average total cost and graph theATC curves for factories with $1,2,3$, and 4 knitting machines. We've already studied the costs of a factorywith 1 machine in Figs. 11.4 and 11.5. In Fig. 11.8, the average total cost curve for that case is $A T C_{1}$. Figure 11.8 also shows the average total cost curve for a factory with 2 machines, $A T C_{2}$, with 3 machines, $A T C_{3}$, and with 4 machines, $A^{2} C_{4}$. You can see, in Fig. 11.8, that the plant size has a big effect on the firm's average total cost.

In Fig. 11.8, two things stand out:

1. Each short-run ATC curve is U-shaped.
2. For each short-run ATC curve, the larger the plant, the greater is the output at which average total cost is at a minimum.

FIGURE 11.8 Short-Run Costs of Four Different Plants


Each short-run ATC curve is U-shaped because, as the quantity of labor increases, its marginal product initially increases and then diminishes. This pattern in the marginal product of labor, which we examined in some detail for the plant with 1 knitting machine on pp. 292-293, occurs at all plant sizes.
The minimum average total cost for a larger plant occurs at a greater output than it does for a smaller plant because the larger plant has a higher total fixed cost and therefore, for any given output, a higher average fixed cost.
Which short-run ATC curve a firm operates on depends on the plant it has. In the long run, the firm can choose its plant and the plant it chooses is the one that enables it to produce its planned output atthe lowest average total cost.
To see why, suppose that Campus Sweaters plans to produce 13 sweaters a day. In Fig. 11.8, with 1 machine, the average total cost curve is ATC and the average total cost of 13 sweaters a day is $\$ 7.69$ a sweater. With 2 machines, on $A T C_{2}$, average total cost is $\$ 6.80$ a sweater. With 3 machines, on $A T C_{3}$, average total cost is $\$ 7.69$ a sweater, the same as with 1 machine. Finally, with 4 machines, on $A T C_{4}$, aver- age total cost is $\$ 9.50$ a sweater. The economically efficient plant for producing a given output is the one that has the lowest average total cost. For Campus Sweaters, the economically efficient plant to use to produce 13 sweaters a day isthe one with 2 machines. In the long run, Cindy chooses the plant that minimizes average total cost. When a firm is producing a given output at the least possible cost, it is operating on its long-run average cost curve.

The long-run average cost curve is the relationship between the lowest attainable average total cost and output when the firm can change both the plant ituses and the quantity of labor it employs. The long-run average cost curve is a planning curve. It tells the firm the plant
and the quantity of labor to use at each output to minimize average cost. Once the firm chooses a plant, the firm operates on the short-run cost curves that apply to thatplant.

## The Long Run Average Cost Curve

Figure 11.9 shows how a long-run average cost curveis derived. The long-run average cost curve LRAC consists of pieces of the four short-run ATC curves. For outputs up to 10 sweaters a day, average total cost is the lowest on $A T C_{1}$. For outputs between 10 and 18 sweaters a day, average total cost is the lowest on $A T C_{2}$. For outputs between 18 and 24 sweaters a day, average total cost is the lowest on $\mathrm{ATC}_{3}$. And for out- puts in excess of 24 sweaters a day, average total cost is the lowest on $A_{4}$. The piece of each ATC curve with the lowest average total cost is highlighted in dark blue in Fig. 11.9. This dark blue scallop-shaped curve made up of the pieces of the four ATC curves is the LRAC curve.

## Economies and Diseconomies of Scale

Economies of scale are features of a firm's technology that make average total cost fall as output increases. When economies of scale are present, the LRAC curve slopes downward. In Fig. 11.9, Campus Sweaters has economies of scale for outputs up to 15 sweaters a day. Greater specialization of both labor and capital is the main source of economies of scale. For example, if GM produces 100 cars a week, each worker must per- form many different tasks and the capital must be general-purpose machines and tools. But if GM produces 10,000 cars a week, each worker specializes in a small number of tasks, uses task-specific tools, and becomes highly proficient.

Diseconomies of scale are features of a firm's technology that make average total cost rise as outputincreases. When diseconomies of scale are present, the LRAC curve slopes upward. In Fig. 11.9, Campus Sweaters experiences diseconomies of scaleat outputs greater than 15 sweaters a day. The challenge of managing a large enterprise is the main source of diseconomies of scale.

Constant returns to scale are features of afirm'stechnology that keep average total cost constant as output increases. When constant returns to scale are present, the LRAC curve is horizontal.

Economies of Scale at Campus Sweaters The economies of scale and diseconomies of scale at Campus Sweaters arise from the firm's productionfunction in Table 11.3. With 1 machine and 1 worker, the firm produces 4 sweaters a day. With 2 machines and 2 workers, total cost doubles but output more than doubles to 15 sweaters a day, so average cost decreases and Campus Sweaters experiences economies of scale. With 4 machines and 4 workers, total cost doubles again but output less than doubles to 26 sweaters a day, so average cost increases and the firm experiences diseconomies of scale.

FIGURE 11.9 Long-Run Average Cost Curve


The long-run average cost curve traces the lowest attainable ATC when both labor and capital change. The green arrows highlight the output range over which each plant achieves the lowest ATC. Within each range, to change the quantity produced, the firm changes the quantity of labor it employs.

Along the LRAC curve, economies of scale occur if average cost falls as output increases; diseconomies of scale occur if average cost rises as output increases. Minimum efficient scale is the output at which average cost is lowest, 15 sweaters a day.

Minimum Efficient Scale A firm's minimum efficient scale is the smallest output at which longrun averagecost reaches its lowest level. At Campus Sweaters, the minimum efficient scale is 15 sweaters a day. The minimum efficient scale plays a role in deter- mining market structure. In a market in which the minimum efficient scale is small relative to market demand, the market has room for many firms, and the market is competitive. In a market in which the minimum efficient scale is large relative to market demand, only a small number of firms, and possibly only one firm, can make a profit and the market is either an oligopoly or monopoly. We will return to this idea in the next three chapters.

## ECONOMIC ANALYSIS

- Starbucks can increase output by hiring more labor, or by increasing its plant size. Or it can both increase its plant size and hire more labor.
- Starbucks can increase its plant size by either replacing an existing café with a larger one, or expanding the number of cafés.
- The decision turns on comparing costs, and Starbucks has figured that it minimizes cost by expanding the number of cafés and hiring more labor.
- We don't know Starbucks' costs, but we can gain insight into the firm's decision with an example.
- The table shows an assumed total product schedule for a Starbucks café. It also shows Starbucks' total cost (TC), marginal cost (MC), and average total cost, (ATC).
- Figure 1 graphs the marginal cost and average total cost curves.
- If Starbucks wants to increase production in a café to above 1,000 coffees per day, marginal cost rises sharply.
- But if Starbucks opens a new café, a given quantity of labor can produce a greater output.
- With a bigger capacity, fixed cost increases, so at low output levels, average total cost also increases.
- But at higher output levels, because average fixed cost decreases, average total cost also decreases.
- Figure 2 shows Starbucks original ATC curve, ATC 0 , and the new ATC curve, ATC 1 , which shows average total cost when the firm has added one more café.


Figure 1 Starbucks' Short-Run Cost Curves

|  | Labor (workers per day) | Output (coffees per day | Total cost (TC) (dollars per day) | Marginal cost (MC) | Average total cost (ATC) |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  | (dollars per coffee) |  |
| A | 0 | 0 | 1,000 |  | - |
| B | 10 | 400 | 1,800 |  | 4.50 |
| C | 20 | 1,000 | 2,600 |  | 2.60 |
| D | 30 | 1,300 | 3,400 |  | 2.62 |
| E | 40 | 1,500 | 4,200 |  | 2.80 |
| F | 50 | 1,600 | 5,000 |  | 3.13 |

- Increasing output with a larger plant size avoids the sharply rising marginal cost of the original café.
- In this example, Starbucks can now hire more labor to operate two cafés and average total cost falls as output increases above 1,000 coffees per day.
- Figure 2 also shows Starbucks long-run average cost curve (LRAC).
- If the firm wants to expand output yet further and avoid the rising costs along $A T C_{1}$, it can open additional cafés and move along its long-run average cost curve.


Figure 2 Starbucks' Long-Run Cost Curve

## Chapter 12: Perfect Competition

After studying this chapter, you will be able to:

- Define perfect competition
- Explain how a firm makes its output decision
- Explain how price and output are determined in perfect competition
- Explain why firms enter and leave a market
- Predict the effects of technological change in a competitive market
- Explain why perfect competition is efficient

A million "apps" have been created for smartphones and tablets. Most of these apps are the work ofindividuals in intense competition with each other. No single app writer can influence the price of an app, but each writer can and must decide how many hours to work and how many apps to produce. In this chapter, we study producers who, like small app developers, are in intense competition- in perfectcompetition. At the end of the chapter, in Economics in the News, we apply the perfect competition model to the highly competitive market in apps.

## What is Perfect Competition?

The firms that you study in this chapter face the force of raw competition. We call this extreme form of competition perfect competition. Perfect competition is a market in which

- Many firms sell identical products to many buyers.
- There are no restrictions on entry into the market.
- Established firms have no advantage over new ones.
- Sellers and buyers are well informed about prices.

Farming, fishing, wood pulping and paper milling, the manufacture of paper cups and shopping bags, grocery and fresh flower retailing, photo finishing, lawn services, plumbing, painting, dry cleaning, and laundry services are all examples of highly competitive industries.

## How Perfect Competition Arises

Perfect competition arises if the minimum efficientscale of a single producer is small relative to the market demand for the good or service. In this situation, there is room in the market for many firms. A firm'sminimum efficient scale is the smallest output at which long-run average cost reaches its lowest level. (See Chapter 11, p. 301.)
In perfect competition, each firm produces a good that has no unique characteristics, so consumers don't care which firm's good they buy.

## Price Takers

Firms in perfect competition are price takers. A price taker is a firm that cannot influence the market price because its production is an insignificant part of the total market. Imagine that you are a wheat farmer in Kansas. You have a thousand acres planted-which sounds like a lot.

If the market price of wheat is $\$ 4$ a bushel, then that is the highest price you can get for your wheat. Ask for $\$ 4.10$ and no one will buy from you. Offer it for $\$ 3.90$ and you'll be sold out in a flash and have given away 10¢ a bushel. You take the market price.

## Economic Profit and Revenue

A firm's goal is to maximize economicprofit, which is equal to total revenue minus total cost. Total cost isthe opportunity cost of production, which includes normal profit. (See Chapter 10, p. 263.) A firm's total revenue equals the price of its output multiplied by the number of units of output sold(price $X$ quantity). Marginal revenue is the change in total revenue that results from a one-unit increase in the quantity sold. Marginal revenue is calculated by dividing the change in total revenue by the change in the quantity sold. Figure 12.1 illustrates these revenue concepts. In part (a), the market demand curve, $D$, and market supply curve, $S$, determine the market price. The market price is $\$ 25$ a sweater. Campus Sweaters is just one of many producers of sweaters, so the best itcan do is to sell its sweaters for $\$ 25$ each.

Total Revenue is equal to the price multiplied by the quantity sold. In the table in Fig. 12.1, if Campus Sweaters sells 9 sweaters, its total revenue is $\$ 225$ (9 X \$25).
Figure 12.1(b) shows the firm's total revenue curve (TR), which graphs the relationship between total revenue and the quantity sold. At point $A$ on the $T R$ curve, the firm sells 9 sweaters and has total revenue of $\$ 225$. Because each additional sweater sold bringsin constant amount-\$25-the total revenue curve is an upward-sloping straight line.

Marginal Revenue is the change in total revenue that results from a one-unit increase in quantity sold. In the table in Fig. 12.1, when the quantity sold increases from 8 to 9 sweaters, total revenue increases from $\$ 200$ to $\$ 225$, so marginal revenue is $\$ 25$ a sweater. Because the firm in perfect competition is a price taker, the change in total revenue that results from a one-unit increase in the quantity sold equals the market price. In perfect competition, the firm's marginal revenue equals the market price. Figure 12.1(c) shows the firm's marginal revenue curve (MR) as the horizontal line at the market price.

Demand for the Firm's Product The firm can sell any quantity it chooses at the market price. So the demand curve for the firm's product is a horizontal line at the market price, the same as the firm's marginal revenue curve.

FIGURE 12.1 Demand, Price, and Revenue in Perfect Competition


| Quantity sold (Q) (sweaters per day) | Price (P) (dollars per sweater) | $\begin{gathered} \text { Total } \\ \text { revenue } \\ (T R=P \times Q) \\ (\text { dollars }) \end{gathered}$ | Marginal revenue ( $M R=\Delta T R / \Delta Q$ ) (dollars per additional sweater) |
| :---: | :---: | :---: | :---: |
| 8 | 25 | 200 | 25 |
| 9 | 25 | 225 |  |
| 10 | 25 | 250 |  |

In part (a), market demand and market supply determine the market price (and quantity). Part (b) shows the firm's total revenue curve (TR). Point A corresponds to the second row of the table-Campus Sweaters sells 9 sweaters at $\$ 25$ a sweater, so total revenue is $\$ 225$. Part (c) shows the firm's marginal revenue curve (MR). This curve is also the demand curve for the firm's sweaters. The demand for sweaters from Campus Sweaters is perfectly elastic at the market price of $\$ 25$ a sweater.

A horizontal demand curve illustrates a perfectly elastic demand, so the demand for the firm's productis perfectly elastic. A sweater from Campus Sweatersis a perfect substitute for a sweater from any other factory. But the market demand for sweaters is not perfectly elastic: Its elasticity depends on the substitutability of sweaters for other goods and services.

## The Firm's Decisions

The goal of the competitive firm is to maximize economic profit, given the constraints it faces. To achieve its goal, a firm must decide

1. How to produce at minimum cost
2. What quantity to produce
3. Whether to enter or exit a market

You've already seen how a firm makes the first decision. It does so by operating with the plant that minimizes long-run average cost - by being on its long-run average cost curve. We'll now see how the firm makes the other two decisions. We start by looking at the firm's output decision.

## The Firm's Output Decision

A firm's cost curves (total cost, average cost, and marginal cost) describe the relationship between its output and costs (see Chapter 11, pp. 291-297). And a firm's revenue curves (total revenue and marginal revenue) describe the relationship between its output and
revenue (pp. 310-311). From the firm's cost curves and revenue curves, we can find the output that maximizes the firm's economic profit.
Figure 12.2 shows how to do this for Campus Sweaters. The table lists the firm's total revenue and total cost at different outputs, and part (a) of the figure shows the firm's total revenue curve, $T R$, and total cost curve, $T C$. These curves are graphs of numbers in the first three columns of the table. Economic profit equals total revenue minus total cost. The fourth column of the table in Fig. 12.2 shows the economic profit made by Campus Sweaters, and part (b) of the figure graphs thesenumbers as its economic profit curve, $E P$.
Campus Sweaters maximizes its economic profit by producing 9 sweaters a day: Total revenue is $\$ 225$, total cost is $\$ 183$, and economic profit is $\$ 42$. No other output rate achieves a larger profit. At outputs of less than 4 sweaters and more than 12 sweaters a day, Campus Sweaters would incur aneconomic loss. At either 4 or 12 sweaters a day, Campus Sweaters would make zero economic profit,called a break-even point.
FIGURE 12.2 Total Revenue, Total Cost, and Economic Profit


| Quantity <br> $(\boldsymbol{Q})$ <br> (sweaters <br> per day) | Total <br> revenue <br> (TR) <br> (dollars) | Total <br> cost <br> (TC) <br> (dollars) | Economic <br> profit <br> (TR - TC) <br> (dollors) |
| :---: | :---: | :---: | :---: |
| 0 | 0 | 22 | -22 |
| 1 | 25 | 45 | -20 |
| 2 | 50 | 66 | -16 |
| 3 | 75 | 85 | -10 |
| 4 | 100 | 100 | 0 |
| 5 | 125 | 114 | 11 |
| 6 | 150 | 126 | 24 |
| 7 | 175 | 141 | 34 |
| 8 | 200 | 160 | 40 |
| 9 | $\mathbf{2 2 5}$ | 183 | 42 |
| 10 | 250 | 210 | 40 |
| 11 | 275 | 245 | 30 |
| 12 | 300 | 300 | 0 |
| 13 | 325 | 360 | -35 |

The table lists Campus Sweaters' total revenue, total cost, and economic profit. Part (a) graphs the total revenue and total cost curves, and part (b) graphs economic profit.

Campus Sweaters makes maximum economic profit, \$42 a day ( $\$ 225-\$ 183)$, when it produces 9 sweaters a day. At outputs of 4 sweaters and 12 sweaters a day, Campus Sweaters makes zero economic profit-these are breakeven points. At an output less than 4 sweaters and greater than 12 sweaters a day, Campus Sweaters incurs an economic loss.

## Marginal Analysis and the Supply Decision

Another way to find the profit-maximizing output isto use marginal analysis, which compares marginal revenue, $M R$, with marginal cost, $M C$. As output increases, the firm's marginal revenue is constant but its marginal cost eventually increases.

If marginal revenue exceeds marginal cost ( $M R>M C$ ), then the revenue from selling one more unit exceeds the cost of producing it and an increase in output increases economic profit. If marginal revenue is less than marginal cost ( $M R<M C$ ), then the revenue from selling one more unit is less than the cost of producing that unit and a decrease in output increases economic profit. If marginal revenue equals marginal cost ( $M R=M C$ ), then the revenue from selling one more unit equals the cost incurred to produce that unit. Economic profit is maximized and either an increase or a decrease in output decreases economic profit. Figure 12.3 illustrates these propositions. If Campus Sweaters increases its output from 8 sweaters to 9 sweaters a day, marginal revenue (\$25) exceeds marginal cost (\$23), so by producing the 9th sweater economic profit increases by $\$ 2$ from $\$ 40$ to $\$ 42$ a day. The blue area in the figure shows the increase in economic profit when the firm increases production from 8 to 9 sweaters per day.
If Campus Sweaters increases its output from 9 sweaters to 10 sweaters a day, marginal revenue ( $\$ 25$ ) is less than marginal cost ( $\$ 27$ ), so by producing the 10th sweater, economic profit decreases. The last column of the table shows that economic profit decreases from $\$ 42$ to $\$ 40$ a day. The red area in the figure shows the economic loss that arises from increasing production from 9 to 10 sweaters a day.
Campus Sweaters maximizes economic profit by producing 9 sweaters a day, the quantity at which marginal revenue equals marginal cost.
A firm's profit-maximizing output is its quantity supplied at the market price. The quantity supplied at a price of $\$ 25$ a sweater is 9 sweaters a day.
If the price were higher than $\$ 25$ a sweater, the firm would increase production. If the price were lower than $\$ 25$ a sweater, the firm would decrease production. These profitmaximizing responses to different market prices are the foundation of the law of supply: Other things remaining the same, the higher the market price of a good, the greater is the quanti1y supplied of that good.

## Temporary Shutdown Decision

You've seen that a firm maximizes profit by producing the quantity at which marginal revenue (price) equals marginal cost. But suppose that at this quantity, price is less than average total cost. In this case, the firm incurs an economic loss. Maximum profit is a loss (a minimum loss). What does the firm do?
If the firm expects the loss to be permanent, it goes out of business. But if it expects the loss to be temporary, the firm must decide whether to shut down temporarily and produce no output, or to keep producing. To make this decision, the firmcompares the loss from shutting down with the loss from producing and takes the action that minimizes its loss.

Loss Comparisons A firm's economic loss equals total fixed cost, TFC, plus total variable cost minustotal revenue. Total variable cost equals average variable cost, $A V C$, multiplied by the quantity produced, Qand total revenue equals price, $P$, multiplied by the quantity $Q$. So:

$$
\text { Economic loss }=T F C+(\mathrm{AVC}-\mathrm{P}) \times Q .
$$

FIGURE 12.3 Profit-Maximizing Output


| Quantity (Q) (sweaters per day) | Total revenue (TR) (dollars) | Marginal revenue (MR) (dollars per additional sweater) | Total cost (TC) (dollars) | Marginal cost (MC) (dollars per additional sweater) | Economic profit <br> (TR - TC) (dollars) |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 7 | 175 | . 25 | 141 | . . 19 | 34 |
| 8 | 200 | 25 | 160 | . 23 | 40 |
| 9 | 225 | . . 25 | 183 | . . 27 | 42 |
| 10 | 250 | . . 25 | 210 | . 35 | 40 |
| 11 | 275 |  | 245 |  | 30 |

The firm maximizes profit by producing the output at which marginal revenue equals marginal cost and marginal cost is increasing. The table and figure show that marginal cost equals marginal revenue and economic profit is maximized when Campus Sweaters produces 9 sweaters a day. The table shows that if Campus Sweaters increases output from 8 to 9 sweaters, marginal cost is $\$ 23$, which is less than the marginal revenue of $\$ 25$. If output increases from 9 to 10 sweaters, marginal cost is $\$ 27$, which exceeds the marginal revenue of $\$ 25$. If marginal revenue exceeds marginal cost, an increase in output increases economic profit. If marginal revenue is less than marginal cost, an increase in output decreases economic profit. If marginal revenue equals marginal cost, economic profit is maximized.

If the firm shuts down, it produces no output ( $Q=0$ ). The firm has no variable costs and no revenue but it must pay its fixed costs, so its economic loss equals total fixed cost. If the firm produces, then in addition to its fixed costs, it incurs variable costs. But it also receives revenue. Its economic loss equals total fixed cost-the loss when shut down-plus total variable cost minus total revenue. If total variable cost exceeds total revenue, this loss exceeds total fixed cost and the firm shuts down. Equivalently, if average variable cost exceeds price, this loss exceeds total fixed cost and the firm shuts down.

The Shutdown Point A firm's shutdown point is the price and quantity at which it is indifferent between producing and shutting down. The shutdown pointoccurs at the price and the quantity at which aver- age variable cost is a minimum. At the shutdown point, the firm is minimizing its loss and its loss equals total fixed cost. If the price falls below mini- mum average variable cost, the firm shuts down temporarily and continues to incur a loss equal to total fixed cost. At prices above minimum average variable cost but below average total cost, the firm produces the loss-minimizing output and incurs a loss, but a loss that is less than total fixed cost.
Figure 12.4 illustrates the firm's shutdown decision and the shutdown point that we've just described for Campus Sweaters.
The firm's average variable cost curve is AVC and the marginal cost curve is MC. Average variable cost has a minimum of $\$ 17$ a sweater when output is 7 sweaters a day. The MC curve intersects the AVCcurve at its minimum. (We explained this relationshipin Chapter 11; see pp. 292-293.)
The figure shows the marginal revenue curve $M R$ when the price is $\$ 17$ a sweater, a price equal to minimum average variable cost. Marginalrevenue equals marginal cost at 7 sweaters a day, so this quantity maximizes economic profit (minimizes economic loss). The ATC curve shows that the firm's average total cost of producing 7 sweaters a day is $\$ 20.14$ a sweater. The firm incurs a loss equal to $\$ 3.14$ a sweater on 7 sweaters a day, so its loss is $\$ 22$ a day. The table in Fig. 12.2 shows that Campus Sweaters' loss equals its total fixed cost.

## The Firm's Supply Curve

A perfectly competitive firm's supply curve shows how its profit-maximizing output varies as the market price varies, other things remaining the same. The supply curve is derived from the firm's marginal cost curve and average variable cost curves. Figure 12.5 illustrates the derivation of the supply curve.
When the price exceeds minimum average variable cost (more than $\$ 17$ ), the firm maximizes profit by producing the output at which marginal cost equals price. If the price rises, the firm increases its out- put-it moves up along its marginal cost curve.
When the price is less than minimum average variable cost (less than $\$ 17$ a sweater), the firm maximizes profit by temporarily shutting down and producing no output. The firm produces zero output at all prices below minimum average variable cost.

FIGURE 12.4 The Shutdown Decision


The shutdown point is at minimum average variable cost. At a price below minimum average variable cost, the firm shuts down and produces no output. At a price equal to minimum average variable cost, the firm is indifferent between shulting down and producing no output or producing the output at minimum average variable cost. Either way, the firm minimizes its economic loss and incurs a loss equal to total fixed cost.
When the price equals minimum average variable cost, the firm maximizes profit either by temporarily shutting down and producing no output or by producing the output at which average variable cost is a minimum-the shutdown point, $T$. The firm never produces a quantity between zero and the quantity at the shutdown point $T$ (a quantity greater than zero and less than 7 sweaters a day).
The firm's supply curve in Fig. 12.5(b) runs along the $y$-axis from a price of zero to a price equal to minimum average variable cost, jumps to point $T$, andthen, as the price rises above minimum average variable cost, follows the marginal cost curve.

## Output, Price, and Profit in the Short Run

To determine the price and quantity in a perfectly competitive market, we need to know how market demand and market supply interact. We start by studying a perfectly competitive market in the short run. The short run is a situation in which the number of firms is fixed.
figure 12.5 A Firm's Supply Curve

(a) Marginal cost and average variable cost

(b) Campus Sweaters' short-run supply curve

Market Supply in the Short Run
The short-run market supply curve shows the quantity supplied by all the firms in the market at each price when each firm's plant and the number of firms remain the same.

You've seen how an individual firm's supply curve is determined. The market supply curve is derived from the individual supply curves. The quantity supplied by the market at a given price is the sum of the quantities supplied by all the firms in the market at that price. Figure 12.6 shows the supply curve for the competitive sweater market. In this example, the market consists of 1,000 firms exactly like Campus Sweaters. At each price, the quantity supplied by the market is 1,000 times the quantity supplied by a single firm.
The table in Fig. 12.6 shows the firm's and the market's supply schedules and how the market sup ply curve is constructed. At prices below $\$ 17$ a sweater, every firm in the market shuts down; the quantity supplied by the market is zero. At $\$ 17$ a sweater, each firm is indifferent between shutting down and producing nothing or operating and producing 7 sweaters a day. Some firms will shut down, and others will supply 7 sweaters a day. The quantitysupplied by each firm is either O or 7 sweaters, and the quantity supplied by the market is between O (all firms shut down) and 7,000 (all firms produce 7 sweaters a day each).
The market supply curve is a graph of the market supply schedules, and the points on the supply curve $A$ through $D$ represent the rows of the table.
To construct the market supply curve, we sum the quantities supplied by all the firms at each price.
Each of the 1,000 firms in the market has a supply schedule like Campus Sweaters. At prices below $\$ 17$ a sweater, the market supply curve runs along the axis. At $\$ 17$ a sweater, the market supply curve is horizontal-supply is perfectly elastic. As the price rises above $\$ 17$ a sweater, each firm increases its quantity supplied and the quantity supplied by the market increases by 1,000 times that of one firm.

## Short-Run Equilibrium

Market demand and short-run market supply determine the market price and market output. Figure 12.7 (a) shows a short-run equilibrium. The short-run supply curve, $S$, is the same as SMin Fig. 12.6. If the market demand curve is $D_{1}$, the market price is $\$ 20$ a sweater. Each firm takes this price as given and produces its profit-maximizing output, which is 8 sweaters a day. Because the market has 1,000 identical firms, the market output is 8,000 sweaters a day.

## A Change in Demand

Changes in demand bring changes to short-run market equilibrium. Figure 12.7(b) shows these changes.
If demand increases and the demand curve shifts rightward to $D_{2}$, the market price rises to $\$ 25$ a sweater. At this price, each firm maximizes profit by increasing its output to 9 sweaters a day. The market output increases to 9,000 sweaters a day.
If demand decreases and the demand curve shifts leftward to $D_{3}$, the market price falls to \$17. At this price, each firm maximizes profit by decreasing itsoutput. If each firm produces 7 sweaters a day, themarket output decreases to 7,000 sweaters a day.

FIGURE 12.6 Short-Run Market Supply Curve


|  | Quantity <br> supplied by <br> Crice <br> (dollars <br> (swere sweaters <br> per day) | Quantity <br> supplied by <br> market <br> (sweaters <br> per day) |  |
| :---: | :---: | :---: | :---: |
| A | 17 | 0 or 7 | 0 to 7,000 |
| B | 20 | 8 | 8,000 |
| C | 25 | 9 | 9,000 |
| D | 31 | 10 | 10,000 |

The market supply schedule is the sum of the supply schedules of all the individual firms. A market that consists of 1,000 identical firms has a supply schedule similar to that of one firm, but the quantity supplied by the market is 1,000 times as large as that of the one firm (see the table). The market supply curve is $S_{M}$. Points $A, B, C$, and $D$ correspond to the rows of the table. At the shutdown price of $\$ 17$ a sweater, each firm produces either 0 or 7 sweaters a day and the quantity supplied by the market is between 0 and 7,000 sweaters a day. The market supply is perfectly elastic at the shutdown price.
If the demand curve shifts farther leftward than D3, the market price remains at \$17 a sweater because the market supply curve is horizontal at that price. Some firms continue to produce 7 sweaters a day, and others temporarily shut down. Firms are indifferent between these two activities, and whichever they choose, they incur an economicloss
equal to total fixed cost. The number of firms continuing to produce is just enough to satisfy the market demand at a price of $\$ 17$ a sweater.

## Profits and Losses in the Short Run

In short-run equilibrium, although the firm produces the profit-maximizing output, ii: does not necessarily end up making an economic profit. It might do so, but it might alternatively break even or incur an economic loss. Economic profit (or loss) per sweater is price, $P$, minus average total cost, ATC So economic profit (or loss) is ( $P-A T C$ ) $X Q$. If price
FIGURE 12.7 Short-Run Equilibrium

(a) Equilibrium

In part ( a ), the market supply curve is $S$ and the market demand curve is $D_{1}$. The market price is $\$ 20$ a sweater.
At this price, each firm produces 8 sweaters a day and the market produces 8,000 sweaters a day.

In part (b), if the market demand increases to $D_{2}$, the

(b) Change in equilibrium
price rises to $\$ 25$ a sweater. Each firm produces 9 sweaters a day and market output is 9,000 sweaters. If market demand decreases to $D_{3}$, the price falls to $\$ 17$ a sweater and each firm decreases its output. If each firm produces 7 sweaters a day, the market output is 7,000 sweaters a day. equals average total cost, a firm breaks even - the entrepreneur makes normal profit. If price exceeds average total cost, a firm makes an economic profit. If price is less than average total cost, a firm incurs an economic loss. Figure 12.8 shows these three possible short-run profit outcomes for Campus Sweaters.
These outcomes correspond to the three different levels of market demand that we've just examined.

## Three Possible Short-Run Outcomes

Figure 12.5(a) corresponds to the situation in Fig. 12.7(a) where the market demand is D1. The equilibrium price of a sweater is $\$ 20$ and the firm produces 8 sweaters a day.Average total cost is $\$ 20$ a sweater. Price equals average total cost (ATC), so the firm breaks even (makes zero economic profit).
Figure 12.5(b) corresponds to the situation in Fig. 12.7(b) where the market demand is $D_{2}$. The equilibrium price of a sweater is $\$ 25$ and the firm produces sweaters a day. Here, price exceeds average total cost, so the firm makes an economic profit. Its economic profit is $\$ 42$ a day, which equals $\$ 4.67$ per sweater ( $\$ 25.00-\$ 20.33$ ) multiplied by 9 , the profit-maximizing
number of sweaters produced. The blue rectangle shows this economic profit. The heightof that rectangle is profit per sweater, $\$ 4.67$, and the length is the quantity of sweaters produced, 9 a day. So the area of the rectangle is economic profit of $\$ 42$ a day.
Figure 12.S(c) corresponds to the situation in Fig. 12.7(b) where the market demand is D3. Theequilibrium price of a sweater is $\$ 17$. Here, the price is less than average total cost, so the firm incurs an economic loss. Price and marginal revenue are $\$ 17$ a sweater, and the profit-maximizing (in this case, loss- minimizing) output is 7 sweaters a day. Total revenue is $\$ 119$ a day ( $7 \times 17$ ). Average total cost is $\$ 20.14$ a sweater, so the economic loss is $\$ 3.14$ per sweater ( $\$ 20.14-\$ 17.00$ ). This loss per sweater multiplied by the number of sweaters is $\$ 22$. The red rectangle shows this economic loss. The height of that rectangle is economic loss per sweater, $\$ 3.14$, and the length is the quantity of sweaters produced, 7 a day. So the area of the rectangle is the firm's economic loss of $\$ 22$ a day. If the price dips below $\$ 17$ a sweater, the firm temporarily shuts down and incurs an economic loss equal to total fixed cost.

FIGURE 12.8 Three Short-Run Outcomes for the Firm

(a) Break even

(b) Economic profit

(c) Economic loss

In the short run, the firm might break even (make zero economic profit), make an economic profit, or incur an economic loss. In part (a), the price equals minimum average total cost. At the profit-maximizing output, the firm breaks even and makes zero economic profit. In part (b), the market price is $\$ 25$ a sweater. At the profit-maximizing output,
the price exceeds average total cost and the firm makes an economic profit, which is equal to the area of the blue rectangle. In part (c), the market price is \$17 a sweater. At the profit-maximizing output, the price is below minimum average total cost and the firm incurs an economic loss, which is equal to the area of the red rectangle.

## ECONOMICS IN ACTION

## Production Cutback and Temporary Shutdown

The high price of gasoline and anxiety about unemployment and future incomes brought a decrease in the demand for luxury goods including high-end motorcycles such as HarleyDavidsons.
Harley-Davidson's profit-maximizing response to the decrease in demand was to cut production and lay off workers. Some of the production cuts andlay- offs were temporary and some were permanent.
Harley-Davidson's bike production plant in York County, Pennsylvania, was temporarily shut down in the summer of 2008 because total revenue was insufficient to cover total variable cost.

The firm also permanently cut its workforce by 300 people. This permanent cut was like that at Campus Sweaters when the market demand for sweaters decreased from D1 to D3 in Fig. 12.7(b).

## Output, Price, and Profit in the Long Run

In short-run equilibrium, a firm might make an economic profit, incur an economic loss, or break even. Although each of these three situations is a short-run equilibrium, only one of them is a long-run equilibrium. The reason is that in the long run, firms can enter or exit the market.

## Entry and Exit

Entry occurs in a market when new firms come intothe market and the number of firms increases. Exit occurs when existing firms leave a market and the number of firms decreases. Firms respond to economic profit and economic loss by either entering or exiting a market. New firms enter a market in which existing firms are making an economic profit. Firms exit a market in which they are incurring an economic loss. Temporary economic profit and temporary economic loss don't trigger entry and exit. It's the prospect of persistent economic profit or loss that triggers entry and exit.
Entry and exit change the market supply, which influences the market price, the quantity produced by each firm, and its economic profit (or loss).
If firms enter a market, supply increases and the market supply curve shifts rightward. The increase in supply lowers the market price and eventually eliminates economic profit. When economic profit reaches zero, entry stops.
If firms exit a market, supply decreases and the market supply curve shifts leftward. The market pricerises and economic loss decreases. Eventually, economic loss is eliminated and exit stops. To summarize:

- New firms enter a market in which existing firms are making an economic profit.
- As new firms enter a market, the market price falls and the economic profit of each firm decreases.
- Firms exit a market in which they are incurring an economic loss.
- As firms leave a market, the market price rises and the economic loss incurred by the remaining firms decreases.
- Entry and exit stop when firms make zero economic profit.


## A Closer Look at Entry

The sweater market has 800 firms with cost curves like those in Fig. 12.9(a). The market demand curve is $D$, the market supply curve is $S_{1}$, and the price is $\$ 25$ a sweater in Fig. 12.9(b). Each firmproduces 9 sweaters a day and makes an economic profit.

This economic profit is a signal for new firms to enter the market. As entry takes place, supply increases and the market supply curve shifts rightward toward $S^{*}$. As supply increases with no changein demand, the market price gradually falls from $\$ 25$ to $\$ 20$ a sweater. At this lower price, each firm makes zero economic profit and entry stops.

Entry results in an increase in market output, but each firm's output decreases. Because the price falls, each firm moves down its supply curve and produces less. Because the number of firms increases, the market produces more.

## A Closer Look at Exit

The sweater market has 1,200 firms with cost curves like those in Fig. 12.9(a). The market demand curve is $D$, the market supply curve is $S_{2}$, and the price is $\$ 17$ a sweater in Fig. 12.9(b). Each firm produces 7 sweaters a day and incurs an economic loss.

This economic loss is a signal for firms to exit the market. As exit takes place, supply decreases and the market supply curve shifts leftward toward $S^{*}$. As supply decreases with no change in demand, the market price gradually rises from $\$ 17$ to $\$ 20$ a sweater. At this higher price, losses are eliminated, each firm makes zero economic profit, and exit stops. Exit results in a decrease in market output, but each firm's output increases. Because the price rises, each firm moves up its supply curve and producesmore. Because the number of firms decreases, themarket produces less.
figure 12.9 Entry, Exit, and Long-Run Equilibrium

(a) Campus Sweaters

Each firm has cost curves like those of Campus Sweaters in part (a). The market demand curve is $D$ in part (b).

When the market supply curve in part (b) is $S_{1}$, the price is $\$ 25$ a sweater. In part (a), each firm produces 9 sweaters a day and makes an economic profit. Profit triggers the entry of new firms and as new firms enter, the market supply curve shifts rightward, from $S_{1}$ toward $S^{*}$. The price falls from $\$ 25$ to $\$ 20$ a sweater, and the quantity produced increases from 7,200 to 8,000 sweaters. Each firm decreases its output to 8

(b) The sweater market
sweaters a day and its economic profit falls to zero.
When the market supply curve is $S_{2}$, the price is $\$ 17$ a sweater. In part (a), each firm produces 7 sweaters a day and incurs an economic loss. Loss triggers exit and as firms exit, the market supply curve shifts leffward, from $S_{2}$ toward $S^{\star}$. The price rises from $\$ 17$ to $\$ 20$ a sweater, and the quantity produced decreases from 8,400 to 8,000 sweaters. Each firm increases its output from 7 to 8 sweaters a day and its economic profit rises to zero.

## ECONOMICS IN ACTION

Entry and exit

An example of entry and falling prices occurred during the 1980s and 1990s in the personal computer market. When IBM introduced its first PC in 1981, IBM had little competition. The price was $\$ 7,000$ (about $\$ 16,850$ in today's money) and IBM made a large economic profit selling the new machine.
Observing IBM's huge success, new firms such as Gateway, NEC, Dell, and a host of others entered the market with machines that were technologically identical to IBM's. In fact, they were so similar that they came to be called "clones." The massive wave of entry into the personal computer market increased the market supply and lowered the price. The economic profit for all firms decreased.
Today, a \$400 computer is vastly more powerful than its 1981 ancestor that cost 42 times as much.
The same PC market that saw entry during the 1980s and 1990s has seen some exit more recently. In 200I, IBM, the firm that first launched the PC, announced that it was exiting the market. The intense competition from Gateway, NEC, Dell, and others that entered the market following IBM's lead has lowered the price and eliminated the economic profit. So IBM now concentrates on servers and other parts of the computer market. IBM exited the PC market because it was incur- ring economic losses. Its exit decreased market supply and made it possible for the remaining firms in the market to make zero economic profit.
International Harvester, a manufacturer of farm equipment, provides another example of exit. For decades, people associated the name "International Harvester" with tractors, combines, and other farm machines. But International Harvester wasn't the only maker of farm equipment. The market became intensely competitive, and the firm began to incur economic losses. Now the firm has a new name, Navistar International, and it doesn't make tractors any more. After years of economic losses andshrinking revenues, it got out of the farm-machine business in 1985 and started to make trucks.
International Harvester exited because it was incurring an economic loss. Its exit decreased supply and made it possible for the remaining firms in the market to break even.

## Long-Run Equilibrium

You've now seen how economic profit induces entry, which in turn eliminates the profit. You've also seen how economic loss induces exit, which in turn eliminates the loss. When economic profit and economic loss have been eliminated and entry and exit have stopped, a competitive market is in long-run equilibrium.

You've seen how a competitive market adjusts toward its long-run equilibrium. But a competitive market is rarely in a state of long-run equilibrium. Instead, it is constantly and restlessly evolving toward long-run equilibrium. The reason is that the market is constantly bombarded with events that change the constraints that firms face.
Markets are constantly adjusting to keep up with changes in tastes, which change demand, and changes in technology, which change costs.
In the next sections, we're going to see how a competitive market reacts to changing tastes and technology and how the market guides resources to their highest-valued use.

## Changes in Demand and Supply as technology Advances

The arrival of high-speed Internet service increasedthe demand for personal computers and the demand for music and movie downloads. At the same time, the arrival of these technologies decreased the demand for the retail services of record stores.
What happens in a competitive market when the demand for its product changes? The perfect competition model can answer this question.

## An Increase in Demand

Producers of computer components are in long-run equilibrium making zero economic profit when the arrival of the high-speed Internet brings an increase in the demand for computers and the components from which they are built. The equilibrium price of a component rises and producers make economic profits. New firms start to enter the market. Supply increases and the price stops rising and then begins to fall. Eventually, enough firms have entered for thesupply and the increased demand to be in balance at a price that enables the firms in the market to return to zero economic profit - long-run equilibrium. Figure 12.10 illustrates. In the market in part (a), demand is $D_{0}$, supply is $S_{0}$, price is $P_{0}$, and market output is $Q_{0}$. At the firm in part (b), profit is maximized with marginal revenue, $M R_{0}$, equal to marginal cost, $M C$, at output $q_{0}$. Economic profit is zero.
Market demand increases and the demand curve shifts rightward to $D_{1}$, in Fig. 12.10(a). The price rises to $P_{1}$, and the quantity supplied increases from to Q1 as the market moves up along its short-run supply curve $S_{0}$. In Fig. 12.10(b), the firm maximizes profit by producing qi, where marginal revenue $M R_{1}$ equals $M C$. The market is now in short-run equilibrium in which each firm makes an economic profit.
The economic profit brings entry and short-run supply increases - the market supply curve starts toshift rightward. The increase in supply lowers the price and firms decrease output-move down alongtheir marginal cost or supply curve in Fig. 12.10(b).
Eventually, entry shifts the supply curve to S 1 in Fig. 12.10(a). The market price has returned to its originallevel, $P_{0}$. At this price, each firm produces $q_{0}$, the same as the quantity produced before the increase in demand. Market output is $\mathrm{Q}_{2}$ in a long-run equilibrium.
The difference between the initial long-runequilibrium and the new long-run equilibrium is the number of firms in the market. An increase in demand has increased the number of firms. In the process of moving from the initial equilibrium to the new one, each firm makes an economic profit.

## A Decrease in Demand

A decrease in demand triggers a similar response to the one you've just studied but in the opposite direction. A decrease in demand brings a lower price, economic losses, and exit. Exit decreases supply, which raises the price to its original level and economic profit returns to zero in a new long-run equilibrium. Economics intheNews on page 323 looksatan example.

FIGURE 12.10 An Increase in Demand

(a) Market

A market starts out in long-run competitive equilibrium. Part (a) shows the market demand curve $D_{0}$, the market supply curve $S_{0}$, the market price $P_{0}$, and the equilibrium quantity $Q_{0}$. Each firm sells its output at the price $P_{0}$, so its marginal revenue curve is $M R_{0}$ in part (b). Each firm produces $q_{0}$ and makes zero economic profit.

Market demand increases from $D_{0}$ to $D_{1}$ in part (a) and the market price rises to $P_{1}$. Each firm maximizes profit by increasing its output to $q_{1}$ in part (b), and the market output

(b) Firm
increases to $Q_{1}$ in part (d). Firms now make economic profits. New firms enter the market, and as they do so, the market supply curve gradually shifts rightward, from $S_{0}$ toward $S_{1}$. This shift gradually lowers the market price from $P_{1}$ back to $P_{0}$. While the price is above $P_{0}$, firms make economic profits, so new firms keep entering the market. Once the price has returned to $P_{0}$, each firm makes zero economic profit and there is no incentive for firms to enter. Each firm produces $q_{0}$, and the market output is $Q_{2}$.

## Technology Advances Change Supply

We've studied the effects of technological change on demand and to isolate those effects we've kept the individual firm's cost curves unchanged. But new technologies also lower production costs. We now study those effects of advancing technology. Starting from a long-run equilibrium, when a new technology becomes available that lowers production costs, the first firms to use it make economic profit. But as more firms begin to use the new technology, market supply increases and the price falls. At first, new-technology firms continue to make positive economic profits, so more enter. But firms that continue to use the old technology incur economic losses. Why? Initially they were making zero economic profit and now with the lower price they incur economic losses. So old-technology firms exit.
Eventually, all the old-technology firms have exited and enough new-technology firms have entered to increase the market supply to a level thatlowers the price to equal the minimum average total cost using the new technology. In this situation, all the firms, all of which are now new-technology firms, are making zero economic profit.
Figure 12.11 illustrates the process that we've just described. Part (a) shows the market demand and supply curves and market equilibrium. Part (b) showsthe cost and revenue curves for a firm using the original old technology. Initially these are the only firms. Part (c) shows the cost and revenue curves for a firmusing a new technology after it becomes available.
In part (a), the demand curve is $D$ and initially the supply curve is $S_{0}$, so the price is $P o$ and the equilibrium quantity is $Q_{0}$. In part (b), marginal revenue is $M R_{0}$ and each firm produces $q_{0}$
where $M R_{0}$ equals $M C_{\text {old }}$. Economic profit is zero and firms are producing atminimum average total cost on the curve $A T C_{\text {old }}$.
When a new technology becomes available, average total cost and marginal cost of production fall, and firms that use the new technology produce with the average total cost curve $A T C_{\text {New }}$ and marginal cost curve $M C_{\text {New }}$ in part (c).
When one firm adopts the new technology, it is too small to influence supply, so the price remains at $P_{0}$ and the firm makes an economic profit. But economic profit brings entry of new-technology firms. Market supply increases and the price falls.
FIGURE 12.10 An Increase in Demand


(a) Market

A market starts out in long-run competitive equilibrium. Part (a) shows the market demand curve $D_{0}$, the market supply curve $S_{0}$, the market price $P_{0}$, and the equilibrium quantity $Q_{0}$. Each firm sells its output at the price $P_{0}$, so its marginal revenue curve is $M R_{0}$ in part (b). Each firm produces $q_{0}$ and makes zero economic profit.

Market demand increases from $D_{0}$ to $D_{1}$ in part (a) and the market price rises to $P_{1}$. Each firm maximizes profit by increasing its output to $q_{1}$ in part (b), and the market output

(b) Firm
increases to $Q_{1}$ in part (a). Firms now make economic profits. New firms enter the market, and as they do so, the market supply curve gradually shifts rightward, from $S_{0}$ toward $S_{1}$. This shift gradually lowers the market price from $P_{1}$ back to $P_{0}$. While the price is above $P_{0}$, firms make economic profits, so new firms keep entering the market. Once the price has returned to $P_{0}$, each firm makes zero economic profit and there is no incentive for firms to enter. Each firm produces $q_{0}$, and the market output is $Q_{2}$.

With price below $P_{0}$, old-technology firms incur an economic loss and exit. With price above $P_{1}$, new-technology firms make an economic profit and enter. When a new longrun equilibrium is achieved, the old-technology firms have gone. The number of new-technology firms that have entered have shifted the supply curve to $S_{1}$. The price is $P_{1}$, marginal revenue is $M R_{1}$, and each firm in Fig. 12.11(c) produces $q_{1}$ using the new technology where $M R_{1}$ equals $M C_{\text {New }}$. Technological change brings only temporary gains to producers. But the lower prices and better products that technological advances bring are permanent gains forconsumers.

## Competition and Efficiency

You've seen how firms in perfect competition decide the quantity to produce in the short run and in thelong run. You've also seen how these individual decisions determine the
market supply that interacts with market demand to determine the equilibrium price and quantity.
We're now going to use what you've learned to gain a deeper understanding of why competition achieves an efficient allocation of resources.

## Efficient Use of Resources

Resource use is efficient when we produce the goods and services that people value most highly (seeChapter 2, pp. 73-75, and Chapter 5, p. 150). If it is possible to make someone better off without anyone else becoming worse off, resources are not being usedefficiently. For example, suppose we produce a computer that no one wants and no one will ever use and, at the same time, some people are clamoring for morevideo games. If we produce fewer computers and reallocate the unused resources to produce more video games, some people will be better off and no one will be worse of $£$ So the initial resource allocation was inefficient.
We can test whether resources are allocated efficiently by comparing marginal social benefit and marginal social cost. In the computer and video games example, the marginal social benefit of a video game exceeds its marginal social cost; the marginal social cost of a computer exceeds its marginal socialbenefit. So by producing fewer computers and more video games, we move resources toward a higher- valued use.

## Choices, equilibrium, and Efficiency

We can use what you have learned about the decisions of consumers and firms and equilibrium in a competitive market to describe an efficient use of resources.

Choices Consumers allocate their budgets to get the most value possible out of them. We derive a consumer's demand curve by finding how the bestbudget allocation changes as the price of good changes. So consumers get the most value out of their resources at all points along their demand curves. If the people who consume a good or service are the only ones who benefit from it, then the market demand curve measures the benefit to the entiresociety and is the marginal social benefit curve.
Competitive firms produce the quantity that maximizes profit. We derive the firm's supply curve byfinding the profit-maximizing quantity at each price. So firms get the most value out of their resources at all points along their supply curves. If the firms that pro-duce a good or service bear all the costs of producing it, then the market supply curve measures the marginal cost to the entire society and the market supply curve is the marginal social cost curve.

Equilibriumand Efficiency Resources areused efficiently when marginal social benefit equals marginal social cost. Competitive equilibrium achieves this efficient outcome because, with no externalities, priceequals marginal social benefit for consumers, and price equals marginal social cost for producers.
The gains from trade are the sum of consumer surplus and producer surplus. The gains from trade for consumers are measured by consumer surplus, which is the area below the demand curve and above the price paid. (See Chapter 5, p. 147.) The gains from trade for producers are measured by producer surplus, which is the area above the supply curve and
below the price received. (See Chapter 5, p. 149.) The total gains from trade equals total surplus-the sum of consumersurplus and producer surplus. When the market for a good or service is in equilibrium, the gains from trade are maximized.

Efficiency in the Sweater Market Figure 12.12 illustrates the efficiency of perfect competition in the sweater market. Part (a) shows the market, and part
(b) shows Campus Sweaters.

In part (a), consumers get the most value from their budgets at all points on the market demand curve, $D=M S B$. Producers get the most value from their resources at all points on the market supply curve, $S=M S C$. At the equilibrium quantity and price, marginal social benefit equals marginal social cost, and resources are allocated efficiently. Total surplus (the sum of producer surplus and consumer surplus) is maximized.
In part (b) Campus Sweaters (and every other firm) makes zero economic profit, and each firm has the plant that enables it to produce at the lowest possible average total cost. Consumers are as well off as possible because the good cannot be produced at a lower cost and the equilibrium price equals that least possible cost.

FIGURE 12.12 Efficiency of Perfect Competition

(a) The sweater market

In part (a), market demand, D, and market supply, S, determine the equilibrium price and quantity. Consumers have made the best available choices on the demand curve, and firms are producing at least cost on the supply curve.

(b) Campus Sweaters

Marginal social benefit, MSB, equals marginal social cost, MSC, so resources are used efficiently. In part (b), Campus Sweaters produces at the lowest possible long-run average total cost and makes zero economic profit.

When firms in perfect competition are away from long-run equilibrium, either entry or exit moves the market toward the situation depicted in Fig. 12.12. During this process, the market is efficient because marginal social benefit equals marginal social cost. But it is only in long-run equilibrium that economic profit is driven to zero and consumers pay the lowest feasible price.

You've now completed your study of perfect competition. Economics in the News on pp. 328329 gives you an opportunity to use what you have learned to understand the market for smartphone and tablet computer "apps."
Although many markets approximate the model of perfect competition, many do not. In Chapter 13, westudy markets at the opposite extreme of marketpower: monopoly. Then in the following chapters we'll study markets that lie between perfect competition and monopoly. In Chapter 14, we study monopolistic competition and in Chapter 15, we study oligopoly. When you have completed this study, you'll have a tool kit that will enable you to understand the variety of real-worldmarkets.

## ECONOMIC ANALYSIS

- The iPhone, iPad, and Android smartphones and tablet computers have created a large demand for apps.
- Although apps are not like corn or sweaters and come in thousands of varieties, the market for apps is highly competitive and we can use the perfect competition model to explain what is happening in that market.
- The market began to operate in 2008 , when the first app developers got to work using a software develop. ment kit made available by Apple.
- During 2009 through 2014, the number of iPhones and Android smartphones increased dramatically. By the end of 2013, 420 million iPhones and 750 million Android phones had been sold.
- The increase in the number of devices in use increased the demand for apps.
- Thousands of developers, most of them individuals, saw a profit opportunity and got to work creating apps. Their entry into the market increased the supply of apps.
- But the demand for apps kept growing and despite the entry of more developers, profit opporfunities remained.
- Figure 1 illustrates the market for apps. In 2013, the demand for apps was $D_{0}$ and the supply was $S_{0}$. The equilibrium price was $P_{0}$ and the quantity was $Q_{0}$.
- Figure 2 illustrates the cost and revenue curves of an individual app developer. With marginal revenue $M R$ and marginal cost $M C$, the developer maximizes profit by producing an app that sells $q_{0}$ units.
- Average total cost of an app (on the ATC curve) is less than the price, so the developer makes an economic profit.
- Economic profit brings entry, so in Fig. 1, supply increases in 2014 to $S_{1}$. But the demand for apps also keeps increasing and in 2014 the demand curve is $D_{1}$.
- The equilibrium quantity increases to $Q_{1}$, and this quantity is produced by an increased number of devel-opers-each producing qo units and each continuing to make an economic profit.
- The developer's cost curves in Fig. 2 are unchanged, but as development tools improve, development costs will fall and the cost curves will shift downward, which will further increase supply.
- At some future date, market supply will increase by enough to eliminate economic profit and the market for apps will be in long-run equilibrium. That date is unknown but likely to be a long way off.


Figure 1 The Market for Apps


Figure 2 An Individual App Developer

## CHAPTER 13: Monopoly

After studying this chapter, you will be able to:

- Explain how monopoly arises
- Explain how a single-price monopoly determines its output and price
- Compare the performance and efficiency of single-price monopoly and competition
- Explain how price discrimination increases profit
- Explain how monopoly regulation influences output, price, economic profit, and efficiency
Google and Microsoft are big players in the markets for Web search and advertising and for computer operating systems, markets that are obviously not perfectly competitive.
In this chapter, we study markets dominated by one big firm. We call such a market monopoly. We studythe performance and the efficiency of monopoly andcompare it with perfect competition.
In Economics in the News at the end of the chapter, we look at the remarkable success of Google and ask whether Google is serving the social interest or violating U.S. and European antitrust laws.


## Monopoly and How It Arises

A monopoly is a market with a single firm that produces a good or service with no close substitutes and that is protected by a barrier that prevents other firms from entering that market.

## How monopoly Arises

Monopoly arises for two key reasons:

- No close substitutes
- Barrier to entry

No Close Substitutes If a good has a close substitute, even though only one firm produces it, that firm effectively faces competition from the producers of the substitute. A monopoly sells a good or service that has no good substitutes. Tap water and bottled water are close substitutes for drinking, but tap water has no effective substitutes for showering or washing a car and a local public utility that supplies tap water is a monopoly.

Barrier to Entry A constraint that protects a firm from potential competitors is called a barrier to entry. There are three types of barrier to entry:

- Natural
- Ownership
- Legal

Natural Barrier to Entry A natural barrier to entry creates a natural monopoly: a market in which economies of scale enable one firm to supply the entire market at the lowest possible
cost. The firms that deliver gas, water, and electricity to our homes are examples of natural monopoly.
Figure 13.1 illustrates a natural monopoly. The market demand curve for electric power is $D$, and the long-run average cost curve is $L R A C$. Economies of scale prevail over the entire length of the $L R A C$ curve. At a price of 5 cents per kilowatt-hour, the quantity demanded is 4 million kilowatt-hours and one firm can produce that quantity at a cost of 5 cents per kilowatt-hour. If two firms shared the market equally, it would cost each of them 10 cents perkilowatt-hour to produce a total of 4 million kilowatt-hours.

Ownership Barrier to Entry An ownership barrier to entry occurs if one firm owns a significant portion of a key resource. An example of this type of monopoly occurred during the last century when De Beers controlled up to 90 percent of the world's supply of diamonds. (Today, its share is only 65 percent.)

Legal Barrier to Entry A legal barrier to entry creates a legal monopoly: a market in which competition and entry are restricted by the granting of a public franchise, government license, patent, or copyright.
A public franchise is an exclusive right granted to a firm to supply a good or service. An example is the US Postal Service, which has the exclusive right to carry first-class mail. Agovernment licensecontrols entry into particular occupations, professions, and industries. Examples of this type of barrier to entry occur in medicine, law, dentistry, school teaching, architecture, and many other professional services. Licensing does not always create a monopoly, but it does restrict competition. A patent is an exclusive right granted to the inventor of a product or service. A copyright is an exclusive right granted to the author or composer of a literary, musical, dramatic, or artistic work. Patents and copyrights are valid for a limited time period that varies from country to country. In the United States, a patent is valid for 20 years. Patents encourage the invention of new products and production methods. They also stimulate innovation - the use of new inventions - by encouraging inventors to publicize their discoveries and offer them for use under license. Patents have stimulated innovations in areas as diverse as soybean seeds, pharmaceuticals, memory chips, and video games.

## ECONOMICS IN ACTION

## Information-Age Monopolies

Information-age technologies have created three big natural monopolies-firms with large plant costs butalmost zero marginal cost, so they experience economies of scale.
These firms are Microsoft, Google, and Facebook. The operating system of 87 percent of personal computers is some version of Windows; Google performs 67 percent of Internet searches and 58 percent of Web browsing is done using Chrome; and Facebook has a 50 percent share of the social media market.
These same information-age technologies have also destroyed monopolies. FedEx, UPS, the fax machine, and e-mail have weakened the monopoly of the U.S. Postal Service; and the satellite dish has weakened cable television monopolies. (Ref: Chart on page 337)

FIGURE 13.1 Natural Monopoly


The marker demand curve for electric power is $D$, and the long-run average cost curve is LRAC. Economies of scale exist over the entire LRAC curve. One firm can distribute 4 million kilowatt-hours at a cost of 5 cents a kilowatt-hour. This same total output costs 10 cents a kilowatt-hour with two firms. One firm can meet the market demand at a lower cost than two or more firms can. The market is a natural monopoly.

## Monopoly Price-Setting Strategies

A major difference between monopoly and competition is that a monopoly sets its own price. In doing so, the monopoly faces a market constraint: To sell a larger quantity, the monopoly must set a lower price. There are two monopoly situations that create two pricing strategies:

- Single price
- Price discrimination

Single Price A single-price monopoly is a firm that must sell each unit of its output for the same price to all its customers. De Beers sells diamonds (of a given size and quality) for the same price to all its customers. If it tried to sell at a low price to some customers and at a higher price to others, only the low-price customers would buy from De Beers. Others would buy from De Beers' low-price customers. De Beers is a single-price monopoly.

Price Discrimination When firm practices price discrimination; it sells different units of a good or service for different prices. Many firms price discriminate.

Microsoft sells its Windows and Office software atdifferent prices to different buyers. Computer manufacturers who install the software on new machines, students and teachers, governments, and businesses all pay different prices. Pizza producers offer a second pizza for a lower price than the first one. These areexamples of price discrimination.
When a firm price discriminates, it looks as though it is doing its customers a favor. In fact, it is charging the highest possible price for each unit sold and making the largest possible profit.

## A Single-Price Monopoly's Output and Price Decision

To understand how a single-price monopoly makes its output and price decision, we must first study the link between price and marginal revenue.

Price and Marginal Revenue Because in a monopoly there is only one firm, the demand curve facing the firm is the market demand curve. Let's look at Bobbie's Barbershop, the sole supplier of haircuts in Cairo, Nebraska. The table in Fig. 13.2 shows the market demand schedule. At a price of $\$ 20$, Bobbie sells no haircuts. The lower the price, the more haircuts per hour she can sell. For example, at $\$ 12$, consumers demand 4 haircuts per hour (row $E$ ). Total revenue (TR) is the price ( $P$ ) multiplied by the quantity sold $(Q)$. For example, in row $D$, Bobbie sells 3 haircuts at $\$ 14$ each, so total revenue is $\$ 42$. Marginal revenue $(M R)$ is the change in total revenue ( $\Delta T R$ ) resulting from a one-unit increase in the quantity sold. For example, if the price falls from $\$ 16$ (row $C$ ) to $\$ 14$ (row $D$ ), the quantity sold increases from 2 to 3haircuts. Total revenue increases from $\$ 32$ to $\$ 42$, so the change in total revenue is \$10. Because the quantity sold increases by 1 haircut, marginal revenue equals the change in total revenue and is $\$ 10$.
Marginal revenue is placed between the two rows toemphasize that marginal revenue relates to the change in the quantity sold.
Figure 13.2 shows the market demand curve and marginal revenue curve ( $M R$ ) and also illustrates the calculation we've just made. Notice that at each level of output, marginal revenue is less than price-the marginal revenue curve lies below the demand curve. Why is marginal revenue less than price? It is because when the price is lowered to sell one more unit, two opposing forces affect total revenue. The lower price results in a revenue loss on the original units sold and a revenue gain on the increased quantity sold. For example, at a price of $\$ 16$ a haircut, Bobbiesells 2 haircuts (point C). If she lowers the price to $\$ 14$, she sells 3 haircuts and has a revenue gain of $\$ 14$ on the third haircut. But she now receives only $\$ 14$ on each of the first 2 haircuts- $\$ 2$ less than be- fore. As a result, she loses $\$ 4$ of revenue on the first 2 haircuts. To calculate marginal revenue, she must deduct this amount from the revenue gain of $\$ 14$. So marginal revenue is $\$ 10$, which is less than the price.

FIGURE 13.2 Demand and Marginal Revenue


|  | Price <br> (P) <br> (dollars per haircut) | Quantity demanded (Q) (haircuts per hour) | Total revenue $(T R=P \times Q)$ (dollars) | Marginal revenue $(M R=\Delta T R / \Delta Q)$ (dollars per haircut) |
| :---: | :---: | :---: | :---: | :---: |
| A | 20 | 0 | 0 | 18 |
| B | 18 | 1 | 18 |  |
| $C$ | 16 | 2 | 32 | . 10 |
| D | 14 | 3 | 42 | 6 |
| E | 12 | 4 | 48 | 2 |
| $F$ | 10 | 5 | 50 |  |

The table shows the demand schedule. Total revenue $(T R)$ is price multiplied by quantity sold. For example, in row $C$, the price is $\$ 16$ a haircut, Bobbie sells 2 haircuts, and total revenue is $\$ 32$. Marginal revenue (MR) is the change in total revenue that results from a one-unit increase in the quantity sold. For example, when the price falls from $\$ 16$ to $\$ 14$ a haircut, the quantity sold increases from 2 to 3 , an increase of 1 haircut, and total revenue increases by $\$ 10$. Marginal revenue is $\$ 10$. The demand curve and the marginal revenue curve, $M R$, are based on the numbers in the table and illustrate the calculation of marginal revenue when the price falls from $\$ 16$ to $\$ 14$ a haircut.

## Marginal Revenue and Elasticity

A single-price monopoly's marginal revenue is related to the elasticity of demand for its good. The demand for a good can be elastic (the elasticity is greater than1), inelastic (the elasticity is less than 1), or unit elastic (the elasticity is equal to 1). Demand is elastic if a 1 percent fall in the price brings a greater than 1 per-cent increase in the quantity demanded. Demand is inelastic if a 1 percent fall in the price brings a less than 1 percent increase in the quantity demanded.
Demand is unit elastic if a 1 percent fall in the price brings a 1 percent increase in the quantity demanded. (See Chapter 4, pp. 122-124.)
If demand is elastic, a fall in the price brings an increase in total revenue - the revenue gain from the increase in quantity sold outweighs the revenue loss from the lower price and marginal revenue is positive. If demand is inelastic, a fall in the price brings a decrease in total revenue-the revenue gain from the increase in quantity sold is outweighed by the revenue loss from the lower price-and marginal revenue is negative. If demand is unit elastic, total revenue does not change-the revenue gain from the increase in the quantity sold offsets the revenue loss from the lower price - and marginal revenue is zero. (See Chapter 4, p. 126.)
Figure 13.3 illustrates the relationship between marginal revenue, total revenue, and elasticity. As theprice gradually falls from $\$ 20$ to $\$ 10$ a haircut, the quantity demanded increases from O to 5 haircuts anhour. Over this output range, marginal revenue is positive in part (a), total revenue increases in part (b), and the demand for haircuts is elastic. As the price falls from $\$ 10$ to $\$ 0$ a haircut, the quantity of hair- cuts demanded increases from 5 to 10 an hour. Over this output range, marginal revenue is negative in part (a), total revenue decreases in part (b), and the demand for haircuts is inelastic. When the price is $\$ 10$ a haircut, marginal revenue is zero in part (a), total revenue is at a maximum in part (b), and the demand for haircuts is unit elastic.

In Monopoly, Demand Is Always Elastic The relationship between marginal revenue and elasticity ofdemand that you've just discovered implies that a profit-maximizing monopoly never produces an out-put in the inelastic range of the market demand curve. If it did so, it could charge a higher price, pro- duce a smaller quantity, and increase its profit. Let'snow look at a monopoly's price and output decision.

FIGURE 13.3 Marginal Revenue and Elasticity

(a) Demand and marginal revenue curves

(b) Total revenue curve

In part (a\}, the demand curve is $D$ and the marginal revenue curve is $M R$. In part (b), the total revenue curve is $T R$. Over the range $O$ to 5 haircuts an hour, a price cut increases total revenue, so marginal revenue is positive-as shown by the bars on the left. Demand is elastic. Over the range 5 to 10 haircuts an hour, a price cut decreases total revenue, so marginal revenue is negative-as shown by the bars on the right. Demand is inelastic. At 5 haircuts an hour, total revenue is maximized and marginal revenue is zero. Demand is unit elastic.
Price and OutputDecision

## Price and Output Decision

A monopoly sets its price and output at the levels that maximize economic profit. To determine this price and output level, we need to study the behavior of both cost and revenue as output varies. A monopoly faces the same types of technology and cost constraints as a competitive firm, so its costs (total cost, average cost, and marginal cost) behave just like those of a firm in perfect competition. And a monopoly's revenues (total revenue, price, and marginal revenue) behave in the way we've just described. Table 13.1 provides information about Bobbie's costs, revenues, and economic profit, and Fig. 13.4 shows the same information graphically.

Maximizing Economic Profit You can see in Table and Fig. 13.4(a) that total cost (TC) and total revenue (TR) both rise as output increases, but TCrises at an increasing rate and $T R$ rises at a decreasing rate. Economic profit, which equals $T R$ minus $T C$, increases at small output levels, reaches a maximum, and then decreases. The maximum profit ( $\$ 12$ ) occurs when Bobbie sells 3 haircuts for $\$ 14$ each. If she sells 2 haircuts for $\$ 16$ each or 4 haircuts for $\$ 12 e a c h$, her economic profit will be only $\$ 8$.

Marginal Revenue Equals Marginal Cost You can see Bobbie's marginal revenue (MR) and marginal cost (MC) in Table 13.1 and Fig. 13.4(b).
When Bobbie increases output from 2 to 3 haircuts, $M R$ is $\$ 10$ and $M C$ is $\$ 6$. MR exceeds MC by $\$ 4$ and Bobbie's profit increases by that amount. If Bobbie increases output yet further, from 3 to 4 haircuts, $M R$ is $\$ 6$ and $M C$ is $\$ 10$. In this case, $M C$ exceeds $M R$ by $\$ 4$, so profit decreases by that amount. When $M R$ exceeds $M C$, profit increases if output increases. When $M C$ exceeds $M R$, profit increases if output decreases. When $M C$ equals $M R$, profit is maximized. Figure 13.4(b) shows the maximum profit as price (on the demand curve $D$ ) minus average total cost (on the ATC curve) multiplied by the quantity produced-the rectangle.

Maximum Price the Market Will Bear Unlike a firm in perfect competition, a monopoly influences the price of what it sells. But a monopoly doesn't set the price at the maximum possible price. At the maxi- mum possible price, the firm would be able to sell only one unit of output, which in general is less than the profit-maximizing quantity. Rather, a monopoly produces the profit-maximizing quantity and sells that quantity for the highest price it can get.
All firms maximize profit by producing the output at which marginal revenue equals marginal cost. For a competitive firm, price equals marginal revenue, so price also equals marginal cost. For a monopoly, price exceeds marginal revenue, so price also exceeds marginal cost.
A monopoly charges a price that exceeds marginal cost, but does it always make an economic profit? In Fig. 13.4(b), Bobbie produces 3 haircuts an hour. Her average total cost is $\$ 10$ (on the ATC curve) and her price is $\$ 14$ (on the $D$ curve), so her profit per haircut is $\$ 4$ (\$14 minus $\$ 10$ ). Bobbie's economic profit is shown by the area of the blue rectangle, which equals the profit per haircut (\$4) multiplied by the number of haircuts (3), for a total of $\$ 12$.


If firms in a perfectly competitive market make a positive economic profit, new firms enter. That does not happen in monopoly. Barriers to entry prevent new firms from entering the market, so a monopoly can make a positive economic profit and might continue to do so indefinitely. Sometimes that economic profit is large, as in the international diamond business.

Bobbie makes a positive economic profit. But suppose that Bobbie's landlord increases the rent on her salon. If Bobbie pays an additional $\$ 12$ an hour for rent, her fixed cost increases by $\$ 12$ an hour. Her marginal cost and marginal revenue don't change, so her profit-maximizing output remains at 3 haircuts anhour. Her profit decreases by $\$ 12$ an hour to zero. If Bobbie's salon rent increases by more than $\$ 12$ an hour, she incurs an economic loss. If this situation were permanent, Bobbie would go out of business.

## Single-Price Monopoly and Competition Compared

Imagine a market that is made up of many small firms operating in perfect competition. Then imagine that a single firm buys out all these small firms and creates a monopoly. What will happen in this market? Will the price rise or fall? Will the quantity produced increase or decrease? Will economic profit increase or decrease? Will either the original competitive situation or the new monopoly situation be efficient?
These are the questions we're now going to answer. First, we look at the effects of monopoly on the price and quantity produced. Then we turn to the questions about efficiency.

FIGURE 13.4 A Monopoly's Output and Price

(b) Demand and marginal revenue and cost curves

In part (a), economic profit is the vertical distance equal to total revenue ( $T R$ ) minus total cost ( $T C$ ) and it is maximized at 3 haircuts an hour.

In part (b), economic profit is maximized when marginal cost (MC) equals marginal revenue (MR). The profit-maximizing output is 3 haircuts an hour. The price is determined by the demand curve (D) and is $\$ 14$ a haircut. The average total cost of a haircut is $\$ 10$, so economic profit, the blue rectangle, is $\$ 12$-the profit per haircut (\$4) multiplied by 3 haircuts.

## Comparing Price and Output

Figure 13.5 shows the market we'll study. The market demand curve is $D$. The demand curve is the same regardless of how the industry is organized. But the supply side and the equilibrium are different in monopoly and competition. First, let's look at thecase of perfect competition.

Perfect Competition Initially, with many small perfectly competitive firms in the market, the market supply curve is $S$. This supply curve is obtained by summing the supply curves of all the individual firms in the market.
In perfect competition, equilibrium occurs where the supply curve and the demand curve intersect. Theprice is $P c$, and the quantity produced by the industryis $Q c$ - Each firm takes the price Pcand maximizes its profit by producing the output at which its own marginal cost equals the price. Because each firm is a small part of the total industry, there is no incentive for any firm to try to manipulate the price by varying its output.

Monopoly Now suppose that this industry is taken over by a single firm. Consumers do not change, so the market demand curve remains the same as in the case of perfect competition. But now the monopoly recognizes this demand curve as a constraint on the price at which it can sell its output. The monopoly's marginal revenue curve is $M R$. The monopoly maximizes profit by producing the quantity at which marginal revenue equals marginal cost. To find the monopoly's marginal cost curve, first recall that in perfect competition, the market supply curve is the sum of the supply curves of the firms in the industry. Also recall that each firm's supply curve is its marginal cost curve (see Chapter 12, p. 315). So when the market is taken over by a single firm, the competitive market's supply curve becomes the monopoly's marginal cost curve. To remind you of this fact, the supply curve is also labeled $\mathrm{M}_{\mathrm{c}}$.
The output at which marginal revenue equals marginal cost is $Q_{M .}$. This output is smaller than the competitive output $Q_{c}$. And the monopoly charges the price $P_{M}$ which is higher than $P_{C}$. We have established that. Compared to a perfectly competitive market, a single-price monopoly produces a smaller output and charges a higher price. We've seen how the output and price of a monopoly compare with those in a competitive market. Let's now compare the efficiency of the two types ofmarket.

## Efficiency Comparison

Perfect competition (with no externalities) is efficient. Figure 13.6(a) illustrates the efficiency of perfect competition and serves as a benchmark against which to measure the inefficiency of monopoly. Along the demand and marginal social benefit curve ( $D=M S B$ ), consumers are efficient. Along the supply curve and marginal social cost curve ( $S=M S C$ ), producers are efficient. In competitive equilibrium, the price is $P c$, the quantity is $Q c$, and marginal social benefit equals marginal social cost. Consumer surplus is the green triangle underthe demand curve and above the equilibrium price (seeChapter 5, p. 147). Producer surplus is the blue area above the supply curve and below the equilibrium price (see Chapter 5, p. 149). Total surplus (consumer surplus and producer surplus) is maximized.

FIGURE 13.5 Monopoly's Smaller Output and Higher Price


A competifive market produces the quantity $Q_{C}$ at price $P_{C}$.
A single-price monopoly produces the quantity $Q_{M}$ at which marginal revenue equals marginal cost and sells that quantity for the price $P_{M}$. Compared to perfect competition, a single-price monopoly produces a smaller output and charges a higher price.
Also, in long-run competitive equilibrium, entry and exit ensure that each firm produces its output at the minimum possible long-run average cost.
To summarize: At the competitive equilibrium, marginal social benefit equals marginal social cost; total surplus is maximized; firms produce at the lowest possible long-run average cost; and resource use is efficient.
Figure 13.6(b) illustrates the inefficiency of monopoly and the sources of that inefficiency. A monopoly produces and sells its output for $\mathrm{P}_{\mathrm{M}}$. The smaller output and higher price drive a wedge between marginal social benefit and marginal social cost and create a deadweight loss. The gray triangle shows the deadweight loss and its magnitude is a measure of the inefficiency of monopoly.
Consumer surplus shrinks for two reasons. First, consumers lose by having to pay more for the good. This loss to consumers is a gain for monopoly and increases the producer surplus. Second, consumers lose by getting less of the good, and this loss is part of the deadweightloss.
Although the monopoly gains from a higher price, it loses some producer surplus because it produces a smaller output. That loss is another part of the dead- weight loss.
A monopoly produces a smaller output than perfect competition and faces no competition, so it does not produce at the lowest possible long-run average cost. As a result, monopoly
damages the consumer interest in three ways: A monopoly produces less, increases the cost of production, and raises the priceby more than the increased cost of production.

## FIGURE 13.6 Inefficiency of Monopoly


(b) Monopoly

In perfect competition in part (a), output is $\mathrm{Q}_{\mathrm{c}}$ and the price is $P_{c}$. Marginal social benefit (MSB) equals marginal social cost (MSC); total surplus, the sum of consumer surplus (triangle) and producer surplus, is maximized; and in the long run, firms produce at the lowest possible
average cost. Monopoly in part (b) produces $Q_{M}$ andraises the price to $P_{M}$. Consumer surplus shrinks, the monopoly gains, and a deadweight loss (the shaded triangle) arises.

## Redistribution of Surpluses

You've seen that monopoly is inefficient because marginal social benefit exceeds marginal social cost and there is deadweight loss-a social loss. But monopoly also brings a redistribution of surpluses.
Some of the lost consumer surplus goes to the monopoly. In Fig. 13.6, the monopoly takes the difference between the higher price, $P_{M}$ and the competitive price, $P_{C}$ on the quantity sold, $\mathrm{Q}_{\mathrm{m}}$. So the monopoly takes that part of the consumer surplus. This portion of the loss of consumer surplus is not a loss to society. It is a redistribution from consumers to the monopoly producer.

## Rent Seeking

You've seen that monopoly creates a deadweight loss and is inefficient. But the social cost of monopoly can exceed the deadweight loss because of an activity called rent seeking. Any surplus - consumer surplus, producer surplus, or economic profit - is called economic rent. The pursuit of wealth by capturing economic rent is called rentseeking.
You've seen that a monopoly makes its economic profit by diverting part of consumer surplus to itself-by converting consumer surplus into economic profit. So the pursuit of economic profit by a monopoly is rent seeking. It is the attempt to capture consumer surplus.
Rent seekers pursue their goals in two main ways:
They might

- Buy a monopoly
- Create a monopoly

Buy a Monopoly To rent seek by buying a monopoly, a person searches for a monopoly that is for sale at a lower price than the monopoly's economic profit. Trading of taxicab licenses is an example of this type of rent seeking. In some cities, taxicabs are regulated. The city restricts both the fares and the number of taxis that can operate so that operating a taxi results in economic profit. A person who wants to operate a taxi must buy a license from someone who already has one. People rationally devote time and effort to seeking out profitable monopoly businesses to buy. In the process, they use up scarce resources that could otherwise have been used to pro- duce goods and services. The value of this lost production is part of the social cost of monopoly. The amount paid for a monopoly is not a social cost because the payment is just a transfer of an existing producer surplus from the buyer to the seller.

Create a Monopoly Rent seeking by creating a monopoly is mainly a political activity. It takes the form of lobbying and trying to influence the political process. Such influence might be sought by making campaign contributions in exchange for legislative support or by indirectly seeking to influence political outcomes through publicity in the media or more direct contacts with politicians and bureaucrats. An example of a monopoly created in this
way is the government-imposed restrictions on the quantities of textiles that may be imported into the United States.
Another is a regulation that limits the number of oranges that may be sold in the United States. These are regulations that restrict output and increase price.
This type of rent seeking is a costly activity that uses up scarce resources. Taken together, firms spend billions of dollars lobbying Congress, state legislators, and local officials in the pursuit of licenses and laws that create barriers to entry and establish a monopoly.

## Rent-Seeking Equilibrium

Barriers to entry create monopoly. But there is no barrier to entry into rent seeking. Rent seeking is like perfect competition. If an economic profit is avail- able, a new rent seeker will try to get some of it. And competition among rent seekers pushes up the price that must be paid for a monopoly, to the point at which the rent seeker makes zero economic profit by operating the monopoly. For example, competition for the right to operate a taxi in New York City leads to a price of more than $\$ 100,000$ for a taxi license, which is sufficiently high to eliminate the economic profit made by a taxi operator.
Figure 13.7 shows a rent-seeking equilibrium. The • cost of rent seeking is a fixed cost that must be added to a monopoly's other costs. Rent seeking and rent- seeking costs increase to the point at which no economic profit is made. The average total cost curve, which includes the fixed cost of rent seeking, shifts upward until it just touches the demand curve.
Economic profit is zero. Ithas been lost in rent seeking.
Consumer surplus is unaffected, but the dead- weight loss from monopoly is larger; The deadweightloss now includes the original deadweight loss triangle plus the lost producer surplus, shown by the enlarged gray area in Fig. 13.7.

## Price Discrimination

You encounter price discrimination - selling a good or service at a number of different prices-when you travel, go to the movies, get your hair cut, visit an art museum or theme park, or buy pizza. These are all examples of firms with market power, setting the prices of an identical good or service at different levels for different customers.
Not all price differences are price discrimination: they reflect differences in production costs. For example, real-time meters for electricity enable power utilities to charge a different price at peak-load times thanduring the night. But it costs more per kilowatt-hour to generate electricity at peak-load times so this pricedifference reflects production cost differences and is not price discrimination.
At first sight, price discrimination appears to be inconsistent with profit maximization. Why would a movie theater allow children to see movies at a dis- count? Why would a hairdresser charge students and senior citizens less? Aren't these firms losing profit by being nice to their customers? The answer, as you are about to discover, is that price discrimination is profitable: It increases economic profit.
But to be able to price discriminate, the firm must sell a product that cannot be resold; and it must be possible to identify and separate different buyer types.

FIGURE 13.7 Rent-Seeking Equilibrium


With competitive rent seeking, a single-price monopoly uses all its economic profit to maintain its monopoly. The firm's rent-seeking costs are fixed costs. They add to total fixed cost and to average total cost. The ATC curve shifts upward until, at the profi-maximizing price, the firm breaks even.

## Two Ways of Price Discriminating

Firms price discriminate in two broad ways. They discriminate

- Among groups of buyers
- Among units of a good

Discriminating Among Groups of Buyers People differ in the value they place on a good-their marginal benefit and willingness to pay. Some of these differences are correlated with features such as age, employment status, and other easily distinguished characteristics.
When such a correlation is present, firms can profit by price discriminating among the different groups of buyers.
For example, salespeople and other business travelers know that a face-to-face sales meeting with a customer might bring a large and profitable order. So for these travelers, the marginal benefit from a trip is large and the price that such a traveler is willing topay for a trip is high. In contrast, for a leisure traveler, any of several different trips and even no trip at all are options. So for leisure travelers, the marginal benefit of a trip is small and the price that such a traveler is willing to pay for a trip is low. Because the price that business travelers are willing to pay exceeds what leisure travelers are willing to pay, it is possible for an airline to price discriminate between these two groups and increase its profit. We'll return to this example of price discrimination below.

Discriminating Among Units of Good Everyone experiences diminishing marginal benefit, so if all the units of the good are sold for a single price, buyersend up with a consumer surplus equal to the value they get from each unit minus the price paid for it.
A firm that price discriminates by charging a buyer one price for a single item and a lower price for a second or third item can capture some of the consumersurplus. Buy one pizza and get a second one for a lower price is an example of this type of price dis-crimination.

## Increasing Profit and Producer Surplus

By getting buyers to pay a price as close as possible to their maximum willingness to pay, a monopoly captures the consumer surplus and converts it into producer surplus. And more producer surplus means more economic profit. To see why more producer surplus means more economic profit, recall some definitions. With total revenue TR and total cost TC,

$$
\text { Economic profit = } T R-T C
$$

Producer surplus is total revenue minus the area under the marginal cost curve. But the area under the marginal cost curve is total variable cost, TVC. So producer surplus equals total revenue minus TVC, or

Producer surplus $=T R-T V C$.
You can see that the difference between economic profit and producer surplus is the same as the difference between TC and TVC. But TC minus TVC equals total fixed cost, TFC So Economic profit = Producer surplus -TFC.
For a given level of total fixed cost, anything that increases producer surplus also increases economic profit.
Let's now see how price discrimination works by looking at a price-discriminating airline.

## A Price-Discriminating Airline

Inter-City Airlines has a monopoly on passenger flights between two cities. Figure 13.8 shows the market demand curve, $D$, for travel on this route. It also shows Inter-City Airline's marginal revenue curve, $M R$, and marginal cost curve, MC. Inter- City's marginal cost is a constant $\$ 40$ per trip. (It is easier to see how price discrimination works for a firm with constant marginal cost.)

Single-Price Profit Maximization As a single-price monopoly, Inter-City maximizes profit by producing the quantity of trips at which MR equals $M C$, which is 8,000 trips a week, and charging $\$ 120$ a trip. Witha marginal cost of $\$ 40$ a trip, producer surplus is $\$ 80$ a trip, and Inter-City's producer surplus is $\$ 640,000$ a week, shown by the area of the rectangle. Inter-City's customers enjoy a consumer surplus shown by the area of the .green triangle.

## FIGURE 13.8 A Single Price of Air Travel



Inter-City Airlines has a monopoly on an air route with a market demand curve 0 . Inter-City's marginal cost, MC, is \$40 per trip. As a single-price monopoly, Inter-City's marginal revenue curve is $M R$. Profit is maximized by selling 8,000 trips a week at $\$ 120$ a trip. Producer surplus is $\$ 640,000$ aweek-the rectangle - and Inter-City's customers enjoy a consumer surplus - the triangle.

Discrimination Between Two Types of Travelers Inter-City surveys its customers and discovers that they are all business travelers. It also surveys people who are not its customers and discovers that they are mainly people who travel for leisure. These people travel by bus or car, but would travel by air at a low fare. Inter-City would like to attract some of these travelers and knows that to do so, it must offer a fare below the current $\$ 120$ a trip. How can it do that? Inter-City digs more deeply into its survey results and discovers that its current customers always plan their travel less than two weeks before departure. In contrast, the people who travel by bus or car know their travel plans at least two weeks ahead of time.
Inter-City sees that it can use what it has discovered about its current and potential new customers to separate the two types of travelers into two markets: one market for business travel and another for leisure travel. Figure 13.9 shows Inter-City's two markets. Part (a), the market for business travel, is the same as Fig. 13.8 Part (b) shows the market for leisure travel. No leisure traveler is willing to pay the business fare of $\$ 120$ a trip, so at that price, the quantity demanded in part (b) is zero. The demand curve $D_{L}$ is the demand for travel on this route after satisfying the demand of business travelers. Inter-City's marginal cost remains at $\$ 40$ a trip, so its marginal revenue curve is $M R_{L}$ Inter-City maximizes profit by setting the leisure fare at $\$ 80$ a trip and attracting 4,000leisure travelers a week. Inter-City's producer surplus
increases by \$160,000 a week-the area of the blue rectangle in Fig. 13.9(b)-and leisure travelers enjoy a consumer surplus-the area of the green triangle.
Inter-City announces its new fare schedule: no restrictions, \$120 and 14-day advance purchase, $\$ 80$. Inter-City increases its passenger count by 50 percent and increases its producer surplus by $\$ 160,000$.

FIGURE 13.9 Price Discrimination

(a) Business travel

Inter-City separates its market into two types of travel: business travel with no restrictions in part (a) and leisure travel that requires a 14-day advance purchase in part (b). For business travel, the profit-maximizing price is $\$ 120$ a trip with 8,000 trips a week. For leisure travel, the profit-maximizing price is $\$ 80$ a trip with 4,000 trips a week.

Discrimination Among Several Types of Travelers Pleased with the success of its price discrimination between business and leisure travelers, Inter-City sees that it might be able to profit even more by dividing its customers into a larger number of types. So it does another customer survey, which reveals that some business travelers are willing to pay $\$ 160$ for a fully-refundable, unrestricted ticket while others are willing to pay only $\$ 120$ for a nonrefundable ticket. So applying the same principles as it used to discriminate between business and leisure travelers, Inter-City now discriminates between business travelers who want a refundable ticket and those who want a non-refundable ticket.
Another survey of leisure travelers reveals that they fall into two groups: those who are able to plan 14 days ahead and others who can plan 21 days ahead. So Inter-City discriminates between these two groups with two fares: an $\$ 80$ and a $\$ 60$ fare. By offering travelers four different fares, the air- line increases its producer surplus and increases its economic profit. But why only four fares? Why not keep looking for ever more traveler types and offer even more fares?

(b) Leisure travel

Inter-City continues to make the same producer surplus on business travel as it did with a single price and businesstravelers continue to enjoy the same consumer surplus. But in part (b), Inter-City sells 4,000 trips to leisure travelers, which increases its producer surplus-the rectangle - and increases consumer surplus - the triangle.

Perfect Price Discrimination Firms try to capture an ever larger part of consumer surplus by devising ahost of special conditions, each one of which appeals to a tiny segment of the market but at the same time excludes others from taking advantage of a lower price. The more consumer surplus a firm is able to capture, the closer it gets to the extreme case called perfect price discrimination, which occurs if a firm can sell each unit of output for the highest price someone is willing to pay for it. In this extreme (hypothetical) case consumer surplus is eliminated and captured as producer surplus.
With perfect price discrimination, something special happens to marginal revenue-the market demand curve becomes the marginal revenue curve. The reason is that when the monopoly cuts the price to sell a larger quantity, it sells only the marginal unit at the lower price. All the other units continue to be sold for the highest price that each buyer is willing to pay. So for the perfect price discriminator, marginal revenue equals price and the market demand curve becomes the monopoly's marginal revenue curve.
With marginal revenue equal to price, Inter-City can obtain even greater producer surplus by increasing output up to the point at which price (and marginal revenue) equals marginal cost.
So Inter-City seeks new travelers who will not pay as much as $\$ 60$ a trip but who will pay more than $\$ 40$, its marginal cost. Inter-City offers a variety of vacation specials at different low fares that appeal only to new travelers. Existing customers continue to pay the higher fares and some, with further perks and frills that have no effect on cost, are induced to pay fares going all the way up to $\$ 200$ a trip.

With all these special conditions and fares, Inter- City increases its output to the quantity demanded at marginal cost, extracts the entire consumer surplus on that quantity, and maximizes economic profit.
Figure 13.10 shows the outcome with perfect price discrimination and compares it with the single-price monopoly outcome. The range of business-class fares extract the entire consumer surplus from this group. The new leisure-class fares going down to $\$ 40$ a trip attract an additional 8,000 travelers and take the entire consumer surplus of leisure travelers. Inter- City makes the maximum possible economic profit.

FIGURE 13.10 Perfect Price Discrimination


Dozens of fares discriminate among many different types of business travelers, and many new low fares with restrictions appeal to leisure travelers. With perfect price discrimination, the market demand curve becomes Inter-City's marginal revenve curve. Producer surplus is maximized when the lowest fare equals marginal cost. Inter-City sells 16,000 trips and makes the maximum possible economic profit.

## Efficiency and Rent Seeking with Price Discrimination

With perfect price discrimination, output increases tothe point at which price equals marginal cost. This out- put is identical to that of perfect competition. Perfect price discrimination pushes consumer surplus to zero but increases the monopoly's producer surplus to equal the total surplus in perfect competition. With perfect price discrimination, no deadweight loss is created, so perfect price discrimination achieves efficiency.
The more perfectly the monopoly can price discriminate, the closer its output is to the competitive output and the more efficient is the outcome.

But the outcomes of perfect competition and perfect price discrimination differ. First, the distribution of the total surplus is not the same. In perfect competition, total surplus is shared by consumers and producers, while with perfect price discrimination, the monopoly takes it all. Second, because the monopoly takes all the total surplus, rent seeking is profitable.
People use resources in pursuit of economic rent, and the bigger the rents, the more resources are used in pursuing them. With free entry into rent seeking,the long-run equilibrium outcome is that rent seekers use up the entire producer surplus.
Real-world airlines are as creative as Inter-City Airlines, as you can see in the cartoon! Disney Corporation is creative too in extracting consumersurplus, as Economics inAction shows. We next study some key monopoly policy issues.

## ECONOMICS IN ACTION

## Attempting Perfect Price Discrimination

If you want to spend a day at Disney World, it will cost you $\$ 99$. You can spend a second consecutive day for an extra $\$ 89$. A third day will cost you $\$ 86$. But for a fourth day, you'll pay only $\$ 20$ and for more days all the way up to 10, you'll pay only $\$ 10$ a day. The Disney Corporation hopes that it has read your willingness to pay correctly and not left you with too much consumer surplus.


Disney's Ticket Prices

## Monopoly Regulation

Natural monopoly presents a dilemma. With economies of scale, it produces at the lowest possible cost. But with market power, it has an incentive to raise the price above the competitive price and pro- duce too little-to operate in the self-interest of themonopolist and not in the social interest.

Regulation - rules administered by a government agency to influence prices, quantities, entry, and other aspects of economic activity in a firm or industry - is a possible solution to this dilemma.
To implement regulation, the government establishes agencies to oversee and enforce the rules. For example, the Surface Transportation Board regulates prices on interstate railroads, some trucking and bus lines, and water and oil pipelines. By the 1970s, almost a quarter of the nation's output was produced by regulated industries (far more than just natural monopolies) and a process of deregulation began.

Deregulation is the process of removing regulation of prices, quantities, entry, and other aspects of economic activity in a firm or industry. During the past 30 years, deregulation has occurred in domestic air transportation, telephone service, interstate trucking, and banking and financial services. Cable TV was deregulated in 1984, re-regulated in 1992, and deregulated again in 1996.
Regulation is a possible solution to the dilemma presented by natural monopoly but not a guaranteedsolution. There are two theories about how regulation actually works: the social interest theory and the capture theory.
The social interest theory is that the political and regulatory process relentlessly seeks out inefficiency and introduces regulation that eliminates deadweightloss and allocates resources efficiently.
The capture theory is that regulation serves the self-interest of the producer, who captures the regulator and maximizes economic profit. Regulation that benefits the producer but creates a deadweight loss gets adopted because the producer's gain is large and visible while each individual consumer's loss is small and invisible. No individual consumer has an incentive to oppose the regulation, but the producer has a big incentive to lobby for it. We're going to examine efficient regulation that serves the social interest and see why it is not a simple matter to design and implement such regulation.

## Efficient Regulation of a Natural Monopoly

A cable TV company is a natural monopoly-it cansupply the entire market at a lower price than two or more competing firms can. Cox Communications, based in Atlanta, provides cable TV to households in 20 states. The firm has invested heavily in satellite receiving dishes, cables, and control equipment and so has large fixed costs. These fixed costs are part of the firm's average total cost. Its average total cost decreases as the number of households served increases because the fixed cost is spread over a larger number of households.
Unregulated, Cox produces the quantity that maximizes profit. Like all single-price monopolies, the profit-maximizing quantity is less than the efficient quantity, and underproduction results in a dead weight loss.
How can Cox be regulated to produce the efficient quantity of cable TV service? The answer is by being regulated to set its price equal to marginal cost, known as the marginal cost pricing rule. The quantity demanded at a price equal to marginal cost is the efficient quantity - the quantity at which marginal benefit equals marginal cost.

Figure 13.11 illustrates the marginal cost pricing rule. The demand curve for cable TV is $D$. Cox'smarginal cost curve is MC. That marginal cost curve is (assumed to be) horizontal at $\$ 10$ per household per month-that is, the cost of providing each additional household with a month of cable programming is $\$ 10$. The efficient outcome occurs if the price is regulated at $\$ 10$ per household per month with 10 million households served. But there is a problem: At the efficient output, average total cost exceeds marginal cost, so a firm that uses marginal cost pricing incurs an economic loss. A cable TV company that is required to use a marginal cost pricing rule will not stay in business for long. How can the firm cover its costs and, at the same time, obey a marginal cost pricing rule?
There are two possible ways of enabling the firm to cover its costs: price discrimination and a two-part price (called a two-part tariff).

FIGURE 13.11 Regulating a Natural Monopoly


A natural monopoly cable TV supplier faces the demand curve $D$. The firm's marginal cost is constant at $\$ 10$ per household per month, as shown by the curve labeled MC. The long-run average cost curve is LRAC.

Unregulated, as a profit-maximizer, the firm serves 5 million households at a price of $\$ 60$ a month. An efficient marginal cost pricing rule sets the price at $\$ 10$ a month. The monopoly serves 10 million households and incurs an economic loss. A second-best average cost pricing rule sets the price at $\$ 30$ a month. The monopoly serves 8 million households and earns zero economic profit.
For example, Verizon offers plans at a fixed monthly price that give access to the cellphone network and unlimited free calls. The price of a call(zero) equals Verizon's marginal cost of a call.

Similarly, a cable TV operator can charge a one-time connection fee that covers its fixed cost and thencharge a monthly fee equal to marginal cost.

## Second-Best Regulation of a Natural Monopoly

A natural monopoly cannot always be regulated to achieve an efficient outcome. There are two possibleways of enabling a regulated monopoly to avoid aneconomic loss:

- Average cost pricing
- Government subsidy

Average Cost Pricing The average cost pricing rule sets price equal to average total cost. With this rulethe firm produces the quantity at which the average total cost curve cuts the demand curve. This rule results in the firm making zero economic profit- breaking even. But because for a natural monopoly average total cost exceeds marginal cost, the quantity produced is less than the efficient quantity and a deadweight loss arises. Figure 13.11 illustrates the average cost pricing rule. The price is $\$ 30$ a month and 8 million households get cable TY.

Government Subsidy A government subsidy is a direct payment to the firm equal to its economic loss. To pay a subsidy, the government must raise the revenue by taxing some other activity. You saw in Chapter 6 that taxes themselves generate deadweightloss.

And the Second-Best Is ... Which is the better option, average cost pricing or marginal cost pricing with a government subsidy? The answer depends on the relative magnitudes of the two deadweight losses.Average cost pricing generates a deadweight loss in the market served by the natural monopoly. A subsidy generates deadweight losses in the markets for the items that are taxed to pay for the subsidy. The smaller deadweight loss is the second-best solution toregulating a natural monopoly. Making this calculation in practice is too difficult, so average cost pricingis generally preferred to a subsidy.
Implementing average cost pricing presents the regulator with a challenge because it is not possible to be sure what a firm's costs are. So regulators use one of two practical rules:

- Rate of return regulation
- Price cap regulation

Rate of Return Regulation Under rate of return regulation, a firm must justify its price by showing that its return on capital doesn't exceed a specified target rate. This type of regulation can end up serving the self-interest of the firm rather than the social interest. The firm's managers have an incentive to inflate costs by spending on items such as private jets, free baseball tickets (disguised as public relations expenses), and lavish entertainment. Managersalso have an incentive to use more capital than the efficient amount. The rate of return on capital is regulated but not the total return on capital, and the greater the amount of capital, the greater is the total return.

## FIGURE 13.12 Price Cap Regulation



A natural monopoly cable TV supplier faces the demand curve $D$. The firm's marginal cost is constant at $\$ 10$ per household per month, as shown by the curve labeled MC. The long-run average cost curve is LRAC.

Unregulated, the firm serves 5 million households at a price of $\$ 60$ a month. A price cap sets the maximum price at $\$ 30$ a month. The firm has an incentive to minimize cost and serve the quantity of households that demand service at the price cap. The price cap regulation lowers the price and increases the quantity.

Price Cap Regulation For the reason that we've just examined, rate of return regulation is increasingly being replaced by price cap regulation. A price cap regulation is a price ceilinga rule that specifies the highest price the firm is permitted to set. This type of regulation gives a firm an incentive to operate efficiently and keep costs under control. Price cap regulation has become common for the electricity and telecommunications industries and is replacing rate of return regulation.
To see how a price cap works, let's suppose that the cable TV operator is subject to this type of regulation. Figure 13.12 shows that without regulation, the firm maximizes profit by serving 5 million house- holds and charging a price of $\$ 60$ a month. If a pricecap is set at $\$ 30$ a month, the firm is permitted to sell any quantity it chooses at that price or at a lower price. At 5 million households, the firm now incurs an economic loss. Itcan decrease the loss by increasing output to 8 million households. To increase out- put above 8 million households, the firm would have to lower the price and again it would incur a loss. So the profit-maximizing quantity is 8 million households - the same as with average cost pricing.

Notice that a price cap lowers the price and increases output. This outcome is in sharp contrast to the effect of a price ceiling in a competitive market that you studied in Chapter 6 (pp. 166-168). Thereason is that in a monopoly, the unregulated equilibrium output is less than the competitive equilibrium output, and the price cap regulation replicates the conditions of a competitive market.
In Fig. 13.12, the price cap delivers average cost pricing. In practice, the regulator might set the cap too high. For this reason, price cap regulation is often combined with earnings sharing regulation-a regulation that requires firms to make refunds to customers when profits rise above a target level.

## ECONOMIC ANALYSIS

- Google gets its revenue by selling advertisements associated with search keywords.
- Google sells keywords based on a combination of will-ingness-to-pay and the number of clicks an advertisement receives, with bids starting at 5 cents per click.
- Google has steadily improved its search engine and refined and simplified its interface with both searchers and advertisers to make searches more powerful and advertising more effective.
- Figure 1 shows Google's extraordinary success in terms of its revenue, cost, and profit.
- Google could have provided a basic search engine with none of the features of today's Google.
- If Google had followed this strategy, people seeking information would have used other search engines and advertisers would have been willing to pay lower prices for Google ads.
- Google would have faced the market described in Fig. 2 and earned a small economic profit.
- Instead, Google improved its search engine and the effectiveness of advertising. The demand for Google ads increased.
- By selling keywords to the highest bidder, Google is able to achieve perfect price discrimination.


Figure 2 Basic Search Engine

- Figure 3 shows the consequences of Google's successful strategy. With perfect price discrimination, Google's producer surplus is maximized. Google produces the efficient quantity of search and advertising by accepting ads at prices that exceed or equal marginal cost.
- Google does not appear to be acting against the social interest: There is no antitrust case to answer.


Figure 1 Google's Revenue, Cost, and Profit


Figure 3 Google with AdWords and Other Features

## CHAPTER 14: Monopolistic Competition

After studying this chapter, you will be able to:

- Define and identify monopolistic competition
- Explain how a firm in monopolistic competition determines its price and output in the short run and the long run
- Explain why advertising costs are high and why firms in monopolistic competition use brand names

At tennis-warehouse.com, you have a choice of racquets made by 19 producers. The top five have 265 different racquets to choose among. Tennis racquet producers compete, but each has a monopoly on its own special kind of racquet-the market is an example of monopolistic competition. The model of monopolistic competition helps us to understand the competition that we see every day. And in Economics in the News, at the end of thechapter, we apply the model to the market for tennis racquets.

## What is Monopolistic Competition?

You have studied perfect competition, in which a large number of firms produce at the lowest possble cost, make zero economic profit, and are efficient. You've also studied monopoly, in which a single firm restricts output, produces at a higher cost and price than in perfect competition, and is inefficient. Most real-world markets are competitive but not perfectly competitive, because firms in these markets have some power to set their prices, as monopolies do. We call this type of market monopolistic competition.

## Monopolistic competition is a market structure in which

- A large number of firms compete.
- Each firm produces a differentiated product.
- Firms compete on product quality, price, and marketing.
- Firms are free to enter and exit the industry.


## Large Number of Firms

In monopolistic competition, as in perfect competition, the industry consists of a large number of firms. The presence of a large number of firms has three implications for the firms in the industry.

Small Market Share In monopolistic competition, each firm supplies a small part of the total industry output. Consequently, each firm has only limited power to influence the price of its product. Each firm's price can deviate from the average price of other firms by only a relatively smallamount.

Ignore Other Firms A firm in monopolistic competition must be sensitive to the average market price of the product, but the firm does not pay attention toany one individual competitor. Because all the firms are relatively small, no one firm can dictate market conditions, and so no one firm's actions directly affect the actions of the other firms.

Collusion Impossible Firms in monopolistic competition would like to be able to conspire to fix a higher price-called collusion. But because the number of firms in monopolistic competition is large, coordination is difficult and collusion is not possible.

## Product Differentiation

A firm practices product differentiation if it makes a product that is slightly different from the products of competing firms. A differentiated product is one that is aclose substitute but not a perfect substitute for the products of the other firms. Some people are willing to pay more for one variety of the product, so when itsprice rises, the quantity demanded of that variety decreases, but it does not (necessarily) decrease to zero. For example, Adidas, Asics, Diadora, Etonic, Fila, New Balance, Nike, Puma, and Reebok all make differentiated running shoes. If the price of Adidas running shoes rises and the prices of the other shoes remain constant, Adidas sells fewer shoes and the other producers sell more. But Adidas shoes don't disappear unless the price rises by a large enough amount.

## Competing on Quality, Price, and Marketing

Product differentiation enables a firm to compete with other firms in three areas: product quality, price, and marketing.

Quality The quality of a product is the physical attributes that make it different from the products of other firms. Quality includes design, reliability, the service provided to the buyer, and the buyer's ease of access to the product. Quality lies on a spectrum that runs from high to low. Some firms - such as Dell Computer Corp. - offer high-quality products. Theyare well designed and reliable, and the customer receives quick and efficient service. Other firms offer a lower-quality product that is poorly designed, that might not work perfectly, and that is not supported by effective customer service.

Price Because of product differentiation, a firm in monopolistic competition faces a downwardsloping demand curve. So, like a monopoly, the firm can setboth its price and its output. But there is a tradeoff between the product's quality and price. A firm that makes a high-quality product can charge a higher price than a firm that makes a low-quality product.

Marketing Because of product differentiation, a firm in monopolistic competition must market its product. Marketing takes two main forms: advertising and packaging. A firm that produces a high-quality product wants to sell it for a suitably high price. To be able to do so, it must advertise and package its product in a way that convinces buyers that they are getting the higher quality for which they are paying a higher price. For example, pharmaceutical companies advertise and package their brand-name drugs to persuade buyers that these items are superior to the lower-priced generic alternatives. Similarly, a lowquality producer uses advertising and packaging to persuade buyers that although the quality is low, the low price more than compensates for this fact.

## Entry and Exit

Monopolistic competition has no barriers to prevent new firms from entering the industry in the long run. Consequently, a firm in monopolistic competition cannot make an economic profit in the long run. When existing firms make an economic profit, new firms enter the industry. This entry lowers prices and eventually eliminates economic profit. When firms incur economic losses, some firms leave the industry in the long run. This exit increases prices and eventually eliminates the economicloss. In long-run equilibrium, firms neither enter nor leave the industry and the firms in the industry make zero economic profit.

Examples of Monopolistic Competition Economics in Action below shows 10 industries that are good examples of monopolistic competition. These industries have a large number of firms, which is shown in parentheses. In the market for audio and video equipment, the 4 largest firms pro- duce only 41 percent of the industry's total sales and the 20 largest firms produce 72 percent of total sales. The number on the right is the Herfindahl- Hirschman Index (see Chapter 10, Pg. 272).

## Price and Output in Monopolistic Competition

Suppose you've been hired by VF Corporation, the firm that owns Nautica Clothing Corporation, to manage the production and marketing of Nautica jackets. Think about the decisions that you must make at Nautica. First, you must decide on the design and quality of jackets and on your marketing program. Second, you must decide on the quantity of jackets to produce and the price at which to sell them.
We'll suppose that Nautica has already made its decisions about design, quality, and marketing and now we'll concentrate on the output and pricingdecisions. We'll study quality and marketing decisions in the next section.
For a given quality of jackets and marketing activity, Nautica faces given costs and market conditions. Givenits costs and the demand for its jackets, how does Nautica decide the quantity of jackets to produce and the price at which to sell them?

## The Firm's Short-Run Output and Price Decision

In the short run, a firm in monopolistic competitionmakes its output and price decision just like a monopoly firm does. Figure 14.1 illustrates this decision for Nautica jackets. The demand curve for Nautica jackets is D. This demand curve tells us the quantity of Nauticajackets demanded at each price, given the prices of other jackets. It is not the demand curve for jackets in general. The MR curve shows the marginal revenue curve associated with the demand curve for Nautica jackets. It is derived just like the marginal revenue curve of a single-price monopoly that you studied in Chapter 13.
The ATC curve and the MC curve show the aver- age total cost and the marginal cost of producing Nautica jackets. Nautica's goal is to maximize its economic profit. To do so, it produces the output at which marginalrevenue equals marginal cost. In Fig. 14.1, this output is 125 jackets a day. Nautica charges the price that buyers are willing to pay for this quantity, which is determined by the demand curve. This price is $\$ 75$ per jacket. When Nautica produces 125 jackets a day, its average total cost is $\$ 25$ per jacket and it makes an
economic profit of $\$ 6,250$ a day ( $\$ 50$ per jacket multiplied by 125 jackets a day). The blue rectangle shows Nautica's economic profit.

Nautica maximizes profit by producing the quantity at which marginal revenue equals marginal cost, 125 jackets a day, and charging the price of $\$ 75$ a jacket. This price exceeds the average total cost of $\$ 25$ a jacket, so the firm makes an economic profit of $\$ 50$ a jacket. The rectangle illustrates economic profit, which equals $\$ 6,250$ a day ( $\$ 50$ a jacket multiplied by 125 jackets a day).

FIGURE 14.1 Economic Profit in the Short Run


## Profit Maximizing Might Be Loss Minimizing

Figure 14.1 shows that Nautica is making a large economic profit. But such an outcome is not inevitable. Afirm might face a level of demand for its product that is too low for it to make an economic profit. Excite@Home was such a firm. Offering high- speed Internet service over the same cable that pro- vides television, Excite@Home hoped to capture a large share of the Internet portal market in competition with AOL, MSN, and a host of other providers. Figure 14.2 illustrates the situation facing Excite@Home in 2001. The demand curve for its portal service is $D$, the marginal revenue curve is $M R$, the average total cost curve is $A T C$, and the marginalcost curve is $M C$. Excite@Home maximized profit- equivalently, it minimized its loss-by producing the output at which marginal revenue equals marginal cost. In Fig. 14.2, this output is 40,000 customers.
Excite@Home charged the price that buyers were willing to pay for this quantity, which was determined by the demand curve and which was $\$ 40$ a month. With 40,000 customers, Excite@Home's average total cost was \$50 per customer, so it incurred an economic loss of $\$ 400,000$ a month ( $\$ 10$ ) a customer multiplied by 40,000 customers). The red rectangle shows Excite@Home's economic loss.

So far, the fl.rm in monopolistic competition looks like a single-price monopoly. It produces the quantity atwhich marginal revenue equals marginal cost and then charges the price that buyers are willing to pay for that quantity, as determined by the demand curve. The keydifference between monopoly and monopolistic competition lies in what happens next when firms either make an economic profit or incur an economic loss.

FIGURE 14.2 Economic Loss in the Short Run


Profit is maximized where marginal revenue equals marginal cost. The loss-minimizing quantity is 40,000 customers. The price of $\$ 40$ a month is less than the average total cost of $\$ 50$ a month, so the firm incurs an economicloss of $\$ 10$ a customer. The rectangle illustrates economic loss, which equals $\$ 400,000$ a month ( $\$ 10$ a customer multiplied by 40,000 customers).

## Long-Run: Zero Economic Profit

A firm like Excite@Home is not going to incur an economic loss for long. Eventually, it goes out of business. Also, there is no restriction on entry into monopolistic competition, so if firms in an industry are making economic profit, other firms have an incentive to enter that industry. As the Gap and other firms start to make jackets similar to those made by Nautica, the demand for Nautica jackets decreases. The demand curve for Nautica jackets and the marginal revenue curve shift leftward. As these curves shift leftward, the profitmaximizing quantity and price fall. Figure 14.3 shows the long-run equilibrium. The demand curve for Nautica jackets and the marginal revenue curve have shifted leftward. The firm produces 75 jackets a day and sells them for $\$ 25$ each. At this output level, average total cost is also \$25 per jacket.

So Nautica is making zero economic profit on its jackets. When all the firms in the industry are making zero economic profit, there is no incentive fornew firms to enter.
If demand is so low relative to costs that firms incur economic losses, exit will occur. As firms leave an industry, the demand for the products of the remaining firms increases and their demand curvesshift rightward. The exit process ends when all the firms in the industry are making zero economic profit.

FIGURE 14.3 Output and Price in the
Long Run


Economic profit encourages entry, which decreases the demand for each firm's product. When the demand curve touches the ATC curve at the quantity at which MR equals MC, the market is in long-run equilibrium. The output that maximizes profit is 75 jackets a day, and the price is $\$ 25$ per jacket. Average total cost is also $\$ 25$ per jacket, so economic profit is zero.

Monopolistic Competition and Perfect Competition
Figure 14.4 compares monopolistic competition and perfect competition and highlights two key differences between them:

- Excess capacity
- Markup

Excess Capacity A firm has excess capacity if it produces less than its efficient scale, which is the quantity at which average total cost is a minimum-the quantity at the bottom of the Ushaped ATC curve. In Fig. 14.4, the efficient scale is 100 jackets a day. Nautica in part (a) produces 75 Nautica jackets a dayand has excess capacity of 25 jackets a day. But if all jackets
are alike and are produced by firms in perfectcompetition, each firm in part (b) produces 100 jackets a day, which is the efficient scale. Average total cost is the lowest possible only in perfect competition. You can see the excess capacity in monopolistic competition all around you. Family restaurants (except for the truly outstanding ones) almost always have some empty tables. You can always get a pizza delivered in less than 30 minutes. It is rare that every pump at a gas station is in use with customers waiting in line. Many real estate agents are ready to help you find or sell a home. These industries are examples of monopolistic competition. The firms have excess capacity. They could sell more by cutting their prices, but they would then incur losses.

FIGURE 14.4 Excess Capacity and Markup

(a) Monopolistic competition

The efficient scale is 100 jackets a day. In monopolistic competition in the long run, because the firm faces a downwardsloping demand curve for its product, the quantily produced is less than the efficient scale and the firm has excess capacity. Price exceeds marginal cost by the amount of the markup.

(b) Perfect competition

In contrast, because in perfect competition the demand for each firm's product is perfectly elastic, the quantity produced in the long run equals the efficient scale and price equals marginal cost. The firm produces at the least possible cost and there is no markup.

Markup A firm's markup is the amount by which price exceeds marginal cost. Figure 14.4(a) showsNautica's markup. In perfect competition, price always equals marginal cost and there is no markup. Figure $14.4(\mathrm{~b})$ shows this case. In monopolistic com- petition, buyers pay a higher price than in perfect competition and also pay more than marginal cost.

## Is Monopolistic Competition Efficient?

Resources are used efficiently when marginal social benefit equals marginal social cost. Price equals marginal social benefit and the firm's marginal cost equals marginal social cost (assuming there are no external benefits or costs). So if the price of a Nautica jacket exceeds the marginal cost of producing it, the quantity of Nautica jackets produced is less than the efficient quantity. And you've just seen that in long-run equilibrium in monopolistic competition, price does exceed marginal cost. So is the quantity produced in monopolistic competition less than the efficient quantity?

Makingthe RelevantComparison Two economists meet in the street, and one asks the other, "How is your husband?" The quick reply is "Compared to what?" This bit of economic wit illustrates a key point: Before we can conclude that something needs fixing, we must check out the available alternatives.
The markup that drives a gap between price and marginal cost in monopolistic competition arises from product differentiation. If there were only one kind of jacket, the total benefit of jackets would almost certainly be lessthan it is with variety. People value variety - not only because it enables each person to select what he or she likes best but also because it provides an external benefit. Most of us enjoy seeing variety in the choices of others. Contrast a scene from the China of the 1960s, when everyone wore a Mao tunic, with the China of today, where everyone wears the clothes of their own choosing. Or contrast a scene from the Germany of the 1930s, when almost everyone who could afford a car owned a first-generation Volkswagen Beetle, with the world of today with its enormous variety of styles and types of automobiles.
If people value variety, why don't we see infinite variety? The answer is that variety is costly. Each different variety of any product must designed, and then customers must be informed about it. These initial costs of design and marketing-called setup costs-mean that some varieties that are too close to others already available are just not worth creating.

The Bottom Line Product variety is both valued and costly. The efficient degree of product variety is the one for which the marginal social benefit of product variety equals its marginal social cost. The loss that arises because the quantity produced is less than the efficient quantity is offset by the gain that arises from having a greater degree of product variety. So compared to the alternative - product uniformity - monopolistic competition might be efficient.

## Product Development and Marketing

When Nautica made its price and output decision that we've just studied, it had already made its product quality and marketing decisions. We'll now look at these decisions and see how they influence the firm's output, price, and economic profit.

## Product Development

The prospect of new firms entering the industry keeps firms in monopolistic competition on their toes! To enjoy economic profits, they must continually seek ways of keeping one step ahead of imitators-other firms who imitate the success of profitable firms.
To maintain economic profit, a firm must either develop an entirely new product, or develop a significantly improved product that provides it with a competitive edge, even if only temporarily. A firm that introduces a new or improved and more differentiated product faces a demand that is less elastic and is able to increase its price and make an economic profit. Eventually, imitators will make close substitutes for the firm's new product and compete away the economic profit arising from an initial advantage. So to restore economic profit, the firm must develop another new or seriously improved product.

Profit-Maximizing Product Development The decision to develop a new or improved product is based on the same type of profit-maximizing calculation that you've already studied. Product development is a costly activity, but it also brings in additional revenues. The firm must balance the cost and revenue at the margin.
The marginal dollar spent on developing a new or improved product is the marginal cost of product development. The marginal dollar that the new or improved product earns for the firm is the marginal revenue of product development. At a low level of product development, the marginal revenue from a better product exceeds the marginal cost. At a high level of product development, the marginal cost of a better product exceeds the marginal revenue.
When the marginal cost and marginal revenue of product development are equal, the firm is under- taking the profit-maximizing amount of product development.

Efficiency and ProductDevelopment Is the profit- maximizing amount of product development also theefficient amount? Efficiency is achieved if the marginal social benefit of a new and improved product equals its marginal social cost. The marginal social benefit of an improved product is the increase in price that consumers are willing to pay for it. The marginal social cost is the amount that the firm must pay to make the improvement. Profit is maximized when marginal revenue equals marginal cost. But in monopolistic competition, marginal revenue is less than price, so product development is probably not pushed to its efficient level.
Monopolistic competition brings many product changes that cost little to implement and are purely cosmetic, such as improved packaging or a new scent in laundry powder. Even when there is a truly improved product, it is never as good as the consumer would like and for which the consumer is willing to pay a higher price. For example, "The Legend of Zelda: Skyward Sword" is regarded as an almost perfect and very cool game, but users complain that it isn't quite perfect. It is a game whose features generate a marginal revenue equal to the marginal cost of creating them.

## Advertising

A firm with a differentiated product needs to ensure that its customers know how its product is different from the competition. A firm also might attempt to create a consumer perception that its product is different, even when that difference is small. Firms use advertising and packaging to achieve this goal.

Advertising Expenditures Firms in monopolistic competition incur huge costs to ensure that buyers appreciate and value the differences between their own products and those of their competitors. So a large proportion of the price that we pay for a good covers the cost of selling it, and this proportion is increasing. Advertising in newspapers and magazines and on radio, television, and the Internet is the main selling cost. But it is not the only one. Selling costs include the cost of shopping malls that look like movie sets, glossy catalogs and brochures, and the salaries, airfares, and hotel bills of salespeople. Advertising expenditures affect the profits of firms, in two ways: They increase costs, and they change demand. Let's look at these effects.

## ECONOMICS IN ACTION

## The Cost of Selling a Pair of Shoes

When you buy a pair of running shoes that cost you $\$ 70$, you're paying $\$ 9$ for the materials from which the shoes are made, $\$ 2.75$ for the services of the Malaysian worker who made the shoes, and $\$ 5.25$ for the production and transportation services of a manufacturing firm in Asia and a shipping company. These numbers total $\$ 17$. You pay $\$ 3$ to the U.S. government in import duty. So we've now accounted for a total of $\$ 20$. Where did the other $\$ 50$ go? It is the cost of advertising, retailing, and other sales and distribution services.

The selling costs associated with running shoes are not unusual. Almost everything that you buy includes a selling cost component that exceeds one half of the total cost. Your clothing, food, electronic items, DVDs, magazines, and even your textbooks cost more to sell than they cost to manufacture.
Advertising costs are only a part, and often a small part, of total selling costs. For example, Nike spends about $\$ 4$ on advertising per pair of running shoes sold.

For the U.S. economy as a whole, there are some 20,000 advertising agencies, which employ more than 200,000 people and have sales of $\$ 45$ billion. These numbers are only part of the total cost of advertising because firms have their own internal advertising departments, the costs of which we can only guess.

But the biggest part of selling costs is not the cost of advertising. It is the cost of retailing services. The retailer's selling costs (and economic profit) are often as much as 50 percent of the price you pay.


Selling Costs and Total Cost Selling costs are fixed costs and they increase the firm's total cost. So like the fixed cost of producing a good, advertising costs per unit decrease as the quantity produced increases.

Figure 14.5 shows how selling costs change a firm's average total cost. The blue curve shows the average total cost of production. The red curve shows the firm's average total cost of production plus advertising. The height of the red area between the two curves shows the average fixed cost of advertising. The total cost of advertising is fixed. But the average cost of advertising decreases as output increases.

Figure 14.5 shows that if advertising increases the quantity sold by a large enough amount, it can lower average total cost. For example, if the quantity sold increases from 25 jackets a day with no advertising to 100 jackets a day with advertising, average total cost falls from $\$ 60$ to $\$ 40 \mathrm{a}$ jacket. The reason is that although the total fixed cost has increased, the greater fixed cost is spread over a greater output, so average total cost decreases.

FIGURE 14.5 Selling Costs and Total Cost


Selling costs such as the cost of advertising are fixed costs. When added to the average total cost of production, selling costs increase average total cost by a greater amount at small outputs than at large outputs. If advertising enables sales to increase from 25 jackets a day to 100 jackets a day, average total cost falls from $\$ 60$ to $\$ 40$ a jacket.

Selling Costs and Demand Advertising and other selling efforts change the demand for a firm's product. But how? Does demand increase or does it decrease? The most natural answer is that advertising increases demand. By informing people about the quality of its products or by persuading people to switch from the products of other firms, a firm might expect to increase the demand for its own products.

But all firms in monopolistic competition advertise, and all seek to persuade customers that they have the best deal. If advertising enables a firm to survive, the number of firms in the market might increase. And to the extent that the number of firmsdoes increase, advertising decreases the demand faced by any one firm. It also makes the demand for any one firm's product more elastic. So advertising can end up not only lowering average total cost but also lowering the markup and the price.
Figure 14.6 illustrates this possible effect of advertising. In part (a), with no advertising, the demand for Nautica jackets is not very elastic. Profit is maximized at 75 jackets per day, and the markup is large. In part (b), advertising, which is a fixed cost, increases average total cost from ATCO to ATC 1 but leaves marginal cost unchanged at MC. Demand becomes much more elastic, the profit-maximizing quantity increases, and the markup shrinks.

## Using Advertising to Signal Quality

Some advertising, like the Roger Federer Rolex watch ad in glossy magazines or the huge number of dollarsthat Coke and Pepsi spend, seems hard to understand. There doesn't seem to be any concrete information about a watch in a tennis player's smile.
And surely everyone knows about Coke and Pepsi. What is the gain from pouring millions of dollars into advertising these well-known colas?
One answer is that advertising is a signal to the consumer of a high-quality product. A signal is an action taken by an informed person (or firm) to send a message to uninformed people. Think about two colas:

Coke and Oke. Oke knows that its cola is not verygood and that its taste varies a lot depending on which cheap batch of unsold cola it happens to buy eachweek. So Oke knows that while it could get a lot of people to try Oke by advertising, they would all quickly discover what a poor product it is and switch back to the cola they bought before. Coke, in contrast, knowsthat its product has a high-quality, consistent taste andthat once consumers have tried it, there is a good chance they'll never drink anything else. On the basis of this reasoning, Oke doesn't advertise but Coke does. And Coke spends a lot of money to make a big splash. Cola drinkers who see Coke's splashy ads know that the fl.rm would not spend so much money advertising if its product were not truly good. So consumers reason that Coke is indeed a really good product. The flashy expensive ad has signaled that Coke is really good without saying anything about Coke.
Notice that if advertising is a signal, it doesn't need any specific product information. It just needs to be expensive and hard to miss. That's what a lot of advertising looks like. So the signaling theory of advertising predicts much of the advertising that we see.

FIGURE 14.6 Advertising and the Markup

(a) No firms advertise

If no firms advertise, demand for each firm's product is low and not very elastic. The profit-maximizing output is small, the markup is large, and the price is high.

(b) All firms advertise

Advertising increases average total cost and shifts the ATC curve upward from $A T C_{0}$ to $A T C_{1}$. If all firms advertise, the demand for each firm's product becomes more elastic. Output increases, the price falls, and the markup shrinks.

## Brand Names

Many firms create and spend a lot of money promoting a brand name. Why? What benefit does a brand name bring to justify the sometimes high cost of establishing it?
The basic answer is that a brand name provides information to consumers about the quality of a product, and is an incentive to the producer to achieve a high and consistent quality standard.
To see how a brand name helps the consumer, think about how you use brand names to get information about quality. You're on a road trip, and it is time to find a place to spend the night. You see road- side advertisements for Holiday Inn, Joe's Motel, and Annie's Driver's Stop. You know about Holiday Inn because you've stayed in it before. You've also seen their advertisements and know what to expect. You have no information at all about Joe's and Annie's. They might be better than the lodgings you do know about, but without that knowledge, you're not going to try them. You use the brand name as information and stay at Holiday Inn.

This same story explains why a brand name pro- vides an incentive to achieve high and consistent quality. Because no one would know whether Joe's and Annie's were offering a high standard of service, they have no incentive to do so. But equally, because every- one expects a given standard of service from HolidayInn, a failure to meet a customer's expectation would almost surely lose that customer to a competitor. SoHoliday Inn has a strong incentive to deliver what it promises in the advertising that creates its brand name.

## Efficiency of Advertising and Brand Names

To the extent that advertising and brand names provide consumers with information about the precise nature of product differences and product quality, they benefit the consumer and enable a better product choice to be made. But the opportunity cost of the additional information must be weighed against the gain to the consumer.
The final verdict on the efficiency of monopolistic competition is ambiguous. In some cases, the gains from extra product variety offset the selling costs and the extra cost arising from excess capacity. The tremendous varieties of books, magazines, clothing, food, and drinks are examples of such gains. It is less easy to see the gains from being able to buy a brandname drug with the identical chemical composition to that of a generic alternative, but many people willingly pay more for the brand-name alternative.
Monopolistic competition is one of the most common market structures that you encounter in your daily life. Economics in the News on pp. 372-373 applies the model of monopolistic competition to the market for tennis racquets and shows why you can expect continual improvement and the introduction of new racquets from Babolat, Wilson, Head, Prince, Dunlop, and other racquet producers.

## ECONOMIC ANALYSIS

- The market for tennis racquets is an example of monopolistic competition.
- Nineteen firms compete in a market with up to 1,000 differentiated racquets.
- Although the racquets are differentiated, most of them are close substitutes for each other.
- Close substitutes have highly elastic demand, so markups are low and economic profit is competed away.
- To make an economic profit, a firm must keep innovating.
- The market for tennis racquets has seen a sequence of innovation: metal frame (Wilson, 1967), oversize frame (Prince, 1976), and graphite frame (Prince, 1980).
- Today's innovation is Babolat's Play Pure Drive electronic racquet described in the news article.
- By creating a substantially differentiated product, Babolat was able to bring to the market a product more clearly differentiated from its competitors.
- The monopolistic competition model explains what is now happening at Babolat and what the future holds.
- Figure 1 shows the market for Babolat's electronic racquet. (The numbers are assumptions.)
- Because Babolat's smart racquet differs from other racquets and has features that users value, the demand curve, $D$, and marginal revenue curve, $M R$, provide a large short-run profit opportunity.
- The marginal cost curve is $M C$ and the average total cost curve is ATC. Babolat maximizes its economic profit by producing the quantity at which marginal revenue equals marginal cost.
- This quantity of racquets can be sold for $\$ 400$ each.
- The blue rectangle shows Babolar's economic profit.
- Because Babolat makes an economic profit, entry will take place. Dunlop, Head, Prince, and Wilson will enter the smart racquet market.
- Figure 2 shows the consequences of entry for Babolat.
- The demand for the Babolat racquet decreases as the market is shared with the other racquets.
- Babolat's profit-maximizing price for the electronic racquet falls, and in the long run economic profit is eliminated.


Figure 1 Economic Profit in the Short Run


Figure 2 Zero Economic Profit in the Long Run

- With zero economic profit, Babolat (along with the other producers) has an incentive to develop an even better differentiated racquet and start the cycle described here again, making an economic profit with a new racquet in the short run.


## CHAPTER 15: Oligopoly

After studying this chapter, you will be able to:

- Define and identify oligopoly
- Use game theory to explain how price and output are determined in oligopoly
- Use game theory to explain other strategic decisions
- Describe the antitrust laws that regulate oligopoly

Chances are that your cellphone service provider is Verizon or AT\&T - if you live in the United States. Two thirds of Americans have plans with these two firms. Similarly, the chip in your computer was made by either Intel or AMD; the battery in your TV remote by Duracell or Energizer; and the air- plane that takes you on a long-distance trip by Boeing or the European firm Airbus.
How does a market work when only two or a handful of firms compete? Toanswer this question, we use the model of oligopoly.
At the end of the chapter, in Economics in the News, we'll look at the market for cellphone service and see how Verizon and AT\&T battle to maximize profit.

## What is Oligopoly?

Oligopoly, like monopolistic competition, lies between perfect competition and monopoly. The firms in oligopoly might produce an identical product and compete only on price, or they might pro- duce a differentiated product and compete on price, product quality, and marketing. Oligopoly is a market structure in which

- Natural or legal barriers prevent the entry of new firms.
- A small number of firms compete.


## Barriers to Entry

Natural or legal barriers to entry can create oligopoly. You saw in Chapter 13 how economies of scale and demand form a natural barrier to entry that can create a natural monopoly. These same factors can create a natural oligopoly.
Figure 15.1 illustrates two natural oligopolies. The demand curve, $D$ (in both parts of the figure), shows the demand for taxi rides in a town. If the average total cost curve of a taxi company is $A T C_{1}$ in part (a), the market is a natural duopoly - an oligopoly market with two firms. You can probably see some examples of duopoly where you live. Some cities have only two taxi companies, two car rental firms, two copy centers, or two college bookstores.
The lowest price at which the firm would remain in business is $\$ 10$ a ride. At that price, the quantity of rides demanded is 60 a day, the quantity that can be provided by just two firms. There is no room in this market for three firms. But if there were only one ':firm, it would make an economic profit and a second firm would enter to take some of the business and economic profit.
If the average total cost curve of a taxi company is $A T C_{2}$ in part (b), the efficient scale of one firm is 20 rides a day. This market is large enough for three firms.

A legal oligopoly arises when a legal barrier to entry protects the small number of firms in a market. A city might license two taxi firms or two bus companies, for example, even though the combination of demand and economies of scale leaves room for more than two firms.
FIGURE 15.1 Natural Oligopoly

(a) Natural duopoly

The minimum average total cost of producing a ride is $\$ 10$, so $\$ 10$ a ride is the lowest possible price that a firm can charge. When a firm produces the efficient scale of 30 rides a day, two firms can satisfy the market demand. This market is

(b) Natural oligopoly with three firms
a natural oligopoly with two firms-a natural duopoly. When the efficient scale of one firm is 20 rides per day, three firms can satisfy the market demand at the lowest possible price. This natural oligopoly has three firms.

## ECONOMICS IN ACTION

## Oligopoly Today

These markets are oligopolies. Although in some of them the number of firms (in parentheses) is large, the share of the market held by the four largest firms (the red bars) is close to 100 percent.

The most concentrated marketsbatteries, glass containers, breakfast cereals, computer hard drives, and light bulbs-are dominated by a handful of firms.

If you want to buy a battery for your TV remote or toothbrush, you'll find it hard to avoid buying a Duracell or an Energizer.


Measures of Concentration

## Small Number of Firms

Because barriers to entry exist, oligopoly consists of a small number of firms, each of which has a large share of the market. Such firms are interdependent, and they face a temptation to cooperate to increase their joint economic profit.

Interdependence With a small number of firms in a market, each firm's actions influence the profits of all the other firms. When Penny Stafford opened her coffee shop in Bellevue, Washington, a nearby Starbucks coffee shop took a hit. Within days, Starbucks began to attract Penny's customers with enticing offers and lower prices. Starbucks survived, but Penny eventually went out of business. Penny Stafford and Starbucks were interdependent.

Temptation to Cooperate When a small number of firms share a market, they can increase their profits by forming a cartel and acting like a monopoly. A cartel is a group of firms acting together-colluding-to limit output, raise price, and increase economic profit. Cartels are illegal, but they do operate in some markets. But for reasons that you'lldiscover in this chapter, cartels tend to break down.

## Examples of Oligopoly

Economics in Action above shows some examples of oligopoly. The dividing line between oligopoly and monopolistic competition is hard to pin down. As a practical matter, we identify oligopoly by looking at concentration ratios, the Herfindahl-Hirschman Index, and information about the geographical scope of the market and barriers to entry. The HHI that divides oligopoly from monopolistic competition is generally taken to be 2,500. An HHI below 2,500 is usually an example of monopolistic competition, and a market in which the HHI exceeds 2,500 is usually an example ofoligopoly.

## Oligopoly Games

Economists think about oligopoly as a game between two or a few players, and to study oligopoly markets they use game theory. Game theory is a set of tools for studying strategic behavior-behavior that takes into account the expected behavior of others and the recognition of mutual interdependence. Game theory was invented by John von Neumann in 1937 and extended by von Neumann and Oskar Morgenstern in 1944 (p. 405). Today, it is one of the majorresearch fields in economics.
Game theory seeks to understand oligopoly as well as other forms of economic, political, social, and even biological rivalries by using a method of analysis specifically designed to understand games of all types, including the familiar games of everyday life (see Talking with Thomas Hubbard on p. 406). To lay the foundation for studying oligopoly games, we first think about the features that all games share.

## What Is a Game?

What is a game? At first thought, the question seems silly. After all, there are many different games. There are ball games and parlor games, games of chance and games of skill. But what is it about all these different activities that makes them games? What do all these games have in common? All games share four common features:

- Rules
- Strategies
- Payoffs
- Outcome

We're going to look at these features of games by playing a game called "the prisoners' dilemma." The prisoners' dilemma game displays the essential features of many games, including oligopoly games, and it gives a good illustration of how game theory works and generates predictions.

## The Prisoners' Dilemma

Art and Bob have been caught red-handed stealing a car. Facing airtight cases, they will receive a sentence of two years each for their crime. During his inter- views with the two prisoners, the district attorney begins to suspect that he has stumbled on the two people who were responsible for a multimillion-dollar bank robbery some months earlier. But this is just a suspicion. He has no evidence on which he can convict them of the greater crime unless he can get themto confess. But how can he extract a confession? Theanswer is by making the prisoners play a game. Thedistrict attorney makes the prisoners play the following game.

Rules Each prisoner (player) is placed in a separate room and cannot communicate with the otherprisoner. Each is told that he is suspected of having carried out the bank robbery and that if both of them confess to the larger crime, each will receive a sentence of 3 years for both crimes. If he alone confesses and his accomplice does not, he will receive only a 1-year sentence while his accomplice will receive a 10-year sentence.

Strategies In game theory, strategies are all the possible actions of each player. Art and Bob each have two possible actions:

1. Confess to the bank robbery.
2. Deny having committed the bank robbery.

Because there are two players, each with two strategies, there are four possible outcomes:

1. Both confess.
2. Both deny.
3. Art confesses and Bob denies.
4. Bob confesses and Art denies.

Payoffs Each prisoner can work out his payoff in each of these situations, and we can tabulate the four possible payoffs for each of the prisoners in what iscalled a payoff matrix for the game. A payoff matrix is, a table that shows the payoffs for every possible action by each player for every possible action by each other player.
Table 15.1 on page 383 shows a payoff matrix for Art and Bob. The squares show the payoffs for each prisoner-the red triangle in each square shows Art's and the blue triangle shows Bob's. If both prisoners confess (top left), each gets a prison term of 3 years. If Bob con- fesses but Art denies (top right), Art gets a IO-year sentence and Bob gets a Iyear sentence. If Art confesses and Bob denies (bottom left), Art gets a l-year sentence and Bob gets a 10-year sentence. Finally, if both of them deny (bottom right), neither
can be convicted of the bank robbery charge but both are sentenced for the car theft-a 2-year sentence.

Outcome The choices of both players determine the outcome of the game. To predict that outcome, we use an equilibrium idea proposed by John Nash of Princeton University (who received the Nobel Prize for Economic Science in 1994 and was the subject of the 2001 movie $A$ Beautiful Mind). In Nash equilibrium, player $A$ takes the best possible action given the action of player $B$ and player $B$ takes the best possible action given the action of player $A$. In the case of the prisoners' dilemma, the Nash equilibrium occurs when Art makes his best choice given Bob's choice and when Bob makes his best choice given Art's choice. To find the Nash equilibrium, we compare all the possible outcomes associated with each choice and eliminate those that are dominated-that are not as good as some other choice. Let's find the Nash equilibrium for the prisoners' dilemma game.

Finding the Nash Equilibrium Look at the situation from Art's point of view. If Bob confesses (top row), Art's best action is to confess because in that case, he is sentenced to 3 years rather than 10 years. If Bob denies (bottom row), Art's best action is still to con-fess because in that case, he receives 1 year rather than 2 years. So Art's best action is to confess.
Now look at the situation from Bob's point of view. If Art confesses (left column), Bob's best action is to confess because in that case, he is sentenced to 3 years rather than 10 years. If Art denies (right column), Bob's best action is still to confess because in that case, he receives 1 year rather than 2 years. So Bob's best action is to confess.
Because each player's best action is to confess, each does confess, each goes to jail for 3 years, and the district attorney has solved the bank robbery. This is the Nash equilibrium of the game.
The Nash equilibrium for the prisoners' dilemma is called a dominant-strategy equilibrium, which is an equilibrium in which the best strategy of each player is to cheat (confess) regardless of the strategy of the otherplayer.

The Dilemma The dilemma arises as each prisoner contemplates the consequences of his decision and puts himself in the place of his accomplice. Each knows that it would be best if both denied. But each also knows that if he denies it is in the best interest of the other to confess. So each considers whether to deny and rely on his accomplice to deny or to confess hoping that his accomplice denies but expecting him to confess. The dilemma leads to the equilibrium of the game. (Ref: Table 15.1 Pg. 383)

A Bad Outcome For the prisoners, the equilibrium of the game, with each confessing, is not the best out- come. If neither of them confesses, each gets only 2 years for the lesser crime. Isn't there some way in which this better outcome can be achieved? It seems that there is not, because the players cannot communicate with each other. Each player can put himself in the other player's place, and so each player can figure out that there is a best strategy for each of them. The prisoners are indeed in a dilemma. Each knows that he can serve 2 years only if he can trust the other to deny. But each prisoner also knows that it is
notin the best interest of the other to deny. So each prisoner knows that he must confess, thereby delivering a bad outcome for both.
The firms in an oligopoly are in a similar situation to Art and Bob in the prisoners' dilemma game. Let's see how we can use this game to understand oligopoly.

## An Oligopoly Price-Fixing Game

We can use game theory and a game like the prisoners' dilemma to understand price fixing, price wars, and other aspects of the behavior of firms in oligopoly. We'll begin with a pricefixing game.
To understand price fixing, we're going to study the special case of duopoly-an oligopoly with twofirms. Duopoly is easier to study than oligopoly with three or more firms, and it captures the essence of all oligopoly situations. Somehow, the two firms must share the market. And how they share it depends on the actions of each. We're going to describe the costs of the two firms and the market demand for the item they produce. We're then going to see how game theory helps us to predict the prices charged and the quantities produced by the two firms in a duopoly.

Cost and Demand Conditions Two firms, Trick and Gear, produce switchgears. They have identical costs. Figure 15.2(a) shows their average total cost curve (ATC) and marginal cost curve (MC). Figure 15.2(b)shows the market demand curve for switchgears (D). The two firms produce identical switchgears, so one firm's switchgear is a perfect substitute for the others', and the market price of each firm's product is identical. The quantity demanded depends on that price-the higher the price, the smaller is the quantity demanded. This industry is a natural duopoly. Two firms can produce this good at a lower cost than either one firm or three firms can. For each firm, average total cost is at its minimum when production is 3,000 units a week. When price equals minimum average total cost, the total quantity demanded is 6,000 units a week, and two firms can just produce that quantity.

Collusion We'll suppose that Trick and Gear enter into a collusive agreement. A collusive agreement is an agreement between two (or more) producers to form a cartel to restrict output, raise the price, and increase profits. Such an agreement is illegal in the United States and is undertaken in secret. The firms in a cartel can pursue two strategies:

- Comply
- Cheat

A firm that complies carries out the agreement. A firm that cheats breaks the agreement to its own benefit and to the cost of the other firm.
Because each firm has two strategies, there are four possible combinations of actions for the firms:

1. Both firms comply.
2. Both firms cheat.
3. Trick complies and Gear cheats.
4. Gear complies and Trick cheats.

FIGURE 15.2 Costs and Demand

(a) Individual firm

(b) Industry

The average total cost curve for each firm is ATC, and the marginal cost curve is MC (part a). Minimum average total cost is $\$ 6,000$ a unit, and it occurs at a production of 3,000 units a week.

Part (b) shows the market demand curve. At a price of $\$ 6,000$, the quantily demanded is 6,000 units per week. The two firms can produce this output at the lowest possible average cost. If the market had one firm, it would be profitable for another to enter. If the market had three firms, one would exit. There is room for only two firms in this industry. It is a natural duopoly.

Colluding to Maximize Profits Let's work out the payoffs to the two firms if they collude to make the maximum profit for the cartel by acting like a monopoly. The calculations that the two firms perform are the same calculations that a monopoly performs. (You can refresh your memory of these calculations by looking at Chapter 13, pp. 340-341.)The only thing that the firms in duopoly must do beyond what a monopoly does is to agree on how much of the total output each of them will produce.
Figure 15.3 shows the price and quantity that maximize industry profit for the duopoly. Part (a) shows the situation for each firm, and part (b) shows the situation for the industry as a whole. The curve labeled $M R$ is the industry marginal revenue curve. This marginal revenue c1;1rve is like that of a single- price monopoly (Chapter 13, p. 338). The curve labeled MC1is the industry marginal cost curve if each firm produces the same quantity of output. This curve is constructed by adding together the outputs of the two firms at each level of marginal cost.
Because the two firms are the same size, at each level of marginal cost the industry output is twice the out-put of one firm. The curve MC1in part (b) is twice asfar to the right as the curve $M C$ in part (a).
To maximize industry profit, the firms in the duopoly agree to restrict output to the rate that makes the industry marginal cost and marginal revenue equal. That output rate, as shown in part (b), is 4,000 units a week. The demand curve shows that the highest price for which the 4,000 switchgears can besold is $\$ 9,000$ each. Trick and Gear agree to chargethis price. To hold the price at $\$ 9,000$ a unit, production must be 4,000 units a week. So Trick and Gear mustagree on output rates for each of them that total 4,000 units a week. Let's suppose that they agree to split the market equally so that each firm produces 2,000 switchgears a week. Because the firms are identical, this division is the most likely.
The average total cost (ATC) of producing 2,000 switchgears a week is $\$ 8,000$, so the profit per unit is $\$ 1,000$ and economic profit is $\$ 2$ million ( 2,000 units $X \$ 1,000$ per unit). The economic profit of each firm is represented by the blue rectangle in Fig. 15.3(a).
We have just described one possible outcome for a duopoly game: The two firms collude to produce the monopoly profit-maximizing output and divide that output equally between themselves. From the industry point of view, this solution is identical to a monopoly. A
duopoly that operates in this way is indistinguishable from a monopoly. The economic profit that is made by a monopoly is the maximum total profit that can be made by the duopoly when the firms collude.
But with price greater than marginal cost, either firm might think of trying to increase profit by cheating on the agreement and producing more than theagreed amount. Let's see what happens if one of the firms does cheat in this way.

## FIGURE 15.3 Colluding to Make Monopoly Profits



One firm Cheats on a Collusive Agreement To set the stage for cheating on their agreement, Trick convinces Gear that demand has decreased and that itcannot sell 2,000 units a week. Trick tells Gear that itplans to cut its price so that it can sell the agreed 2,000 units each week. Because the two firms pro-duce an identical product, Gear matches Trick's pricecut but still produces only 2,000 units a week.
In fact, there has been no decrease in demand. Trick plans to increase output, which it knows will lower the price, and Trick wants to ensure that Gear's output remains at the agreed level. Figure 15.4 illustrates the consequences of Trick's cheating. Part (a) shows Gear (the complier); part (b)shows Trick (the cheat); and part (c) shows the industry as a whole. Suppose that Trick increases output to 3,000 units a week. If Gear sticks to the agreement toproduce only 2,000 units a week, total output is now5,000 a week, and given demand in part (c), the price falls to \$7,500 a unit.
Gear continues to produce 2,000 units a week at a cost of $\$ 8,000$ a unit and incurs a loss of $\$ 500$ a unit, or $\$ 1$ million a week. This economic loss is shown by the red rectangle in part (a). Trick produces 3,000 units a week at a cost of $\$ 6,000$ a unit. With a price of $\$ 7,500$, Trick makes a profit of $\$ 1,500$ a unit and therefore an economic profit of $\$ 4.5$ million. This economic profit is the blue rectangle in part (b).
We've now described a second possible outcome for the duopoly game: One of the firms cheats on the collusive agreement. Inthis case, the industry output islarger than the monopoly output and the industry price is lower than the monopoly price. The total economic profit made by the industry is also smaller than the monopoly's economic profit. Trick (the cheat)makes an economic profit of $\$ 4.5$ million, and Gear(the complier) incurs an economic loss of $\$ 1$ million. The industry makes an economic profit of $\$ 3.5$ mil-lion. This industry profit is $\$ 0.5$ million less than the conomic profit that a monopoly would make, but it is distributed
unevenly. Trick makes a bigger economic profit than it would under the collusive agreement, while Gear incurs an economic loss.
A similar outcome would arise if Gear cheated and Trick complied with the agreement. The industry profit and price would be the same, but in this case, Gear (the cheat) would make an economic profit of $\$ 4.5$ million and Trick (the complier) would incur an economic loss of $\$ 1$ million. Let's next see what happens if both firms cheat.
figure 15.4 One Firm Cheats

(a) Complier

switchgears per week)
(b) Cheat

(c) Industry

One firm, shown in part (a), complies with the agreement and produces 2,000 units. The other firm, shown in part (b), cheats on the agreement and increases its output to 3,000 units a week. Given the market demand curve, shown in part (c), and with a total production of 5,000 units a week, the price falls to $\$ 7,500$ a unit. At this price, the complier in part (a) incurs an economic loss of $\$ 1$ million ( $\$ 500$ per unit $\times 2,000$ units), shown by the rectangle. In part (b), the cheat makes an economic profit of $\$ 4.5$ million ( $\$ 1,500$ perunit $x 3,000$ units), shown by the rectangle.

Both Firms Cheat Suppose that both firms cheat and that each firm behaves like the cheating firm that wehave just analyzed. Each tells the other that it is unable to sell its output at the going price and that itplans to cut its price. But because both firms cheat, each will propose a successively lower price. As long as price exceeds marginal cost, each firm has an incentive to increase its production-to cheat. Only when price equals marginal cost is there no further incentive to cheat. This situation arises when the price has reached $\$ 6,000$. At this price, marginal cost equals price. Also, price equals minimum average total cost. At a priceless than $\$ 6,000$, each firm incurs an economic loss.
At a price of $\$ 6,000$, each firm covers all its costsand makes zero economic profit. Also, at a price of $\$ 6,000$, each firm wants to produce 3,000 units a week, so the industry output is 6,000 units a week. Given the demand conditions, 6,000 units can be sold at a price of $\$ 6,000$ each.
Figure 15.5 illustrates the situation just described. Each firm, in part (a), produces 3,000 units a week, and its average total cost is a minimum ( $\$ 6,000$ per unit). The market as a whole, in part (b), operates atthe point at which the market demand curve (D)intersects the industry marginal cost curve (MC1), Each firm has lowered its price and increased its out- put to try to
gain an advantage over the other firm. Each has pushed this process as far as it can without incurring an economic loss.
We have now described a third possible outcome of this duopoly game: Both firms cheat. If both firms cheat on the collusive agreement, the output of each firm is 3,000 units a week and the price is $\$ 6,000$ a unit. Each firm makes zero economic profit.

The Payoff Matrix Now that we have described the strategies and payoffs in the duopoly game, we cansummarize the strategies and the payoffs in the form of the game's payoff matrix. Then we can find the Nash equilibrium.
Table 15.2 sets out the payoff matrix for this game. It is constructed in the same way as the payoff matrix for the prisoners' dilemma in Table 15.1. The squares show the payoffs for the two firms - Gear and Trick. In this case, the payoffs are profits. (For the prisoners' dilemma, the payoffs were losses.)
The table shows that if both firms cheat (top left), they achieve the perfectly competitive outcome - each firm makes zero economic profit. If both firms comply (bottom right), the industry makes the monopoly profit and each firm makes an economic profit of \$2 million. The top right and bottom left squares show the payoff if one firm cheats while the other complies. The firm that cheats makes an economic profit of $\$ 4.5$ million, and the one that complies incurs a loss of $\$ 1$ million.

Nash Equilibrium in the Duopolists' Dilemma The duopolists have a dilemma like the prisoners' dilemma. Do they comply or cheat? To answer this question, we must find the Nash equilibrium.
FIGURE 15.5 Both Firms Cheat

(a) Individual firm
(b) Industry

If both firms cheat by increasing production, the collusive agreement collapses. The limit to the collapse is the competitive equilibrium. Neither firm will cut its price below $\$ 6,000$ (minimum average total cost) because to do so will result in losses. In part (a), each firm produces 3,000 units a week at an average total cost of $\$ 6,000$. In part (b), with a total production of 6,000 units, the price falls to $\$ 6,000$. Each firm now makes zero economic profit. This output and price are the ones that would prevail in a competitive industry.

Each square shows the payoffs from a pair of actions. For example, if both firms comply with the collusiveagreement, the payoffs are recorded in the bottomright square. The darker triangle shows Gear's payoff, and the bottom triangle shows Trick's. In Nash equilibrium, both firms cheat.

TABLE 15.2 Duopoly Payoff Matrix


Look at things from Gear's point of view. Gear reasons as follows: Suppose that Trick cheats. If I comply, I will incur an economic loss of $\$ 1$ million.
If I also cheat, I will make zero economic profit. Zero is better than minus \$ I million, so I'm better off If I cheat. Now suppose Trick complies. If I cheat, I will make an economic profit of $\$ 4.5$ million, and if I comply, I will make an economic profit of $\$ 2$ mil- lion. A $\$ 4.5$ million profit is better than a $\$ 2$ millionprofit, so I'm better off if I cheat. So regardless of whether Trick cheats or complies, it pays Gear to cheat. Cheating is Gear's best strategy. Trick comes to the same conclusion as Gear because the two firms face an identical situation. So both firms cheat. The Nash equilibrium of the duopoly game is that both firms cheat. And although the industry has only two firms, they charge the same price and produce the same quantity as those in a competitive industry. Also, as in perfect competition, each firm makes zero economic profit.
Economics in Action (on this page) and Economics in the News (p. 391) look at some other prisoners' dilemma games. But not all games are prisoners'dilemmas, as you'll now see.

## A Game of Chicken

The Nash equilibrium for the prisoners' dilemma isunique: both players cheat (confess). Not all games have a unique equilibrium, and one that doesn't is a game called "chicken."

An Example of the Game of Chicken A graphic, if disturbing, version of "chicken" has two cars racing toward each other. The first driver to swerve and avoid a crash is the "chicken." The payoffs are a big loss for both if no one "chickens out;" zero for both if both "chicken out;" and zero for the chicken and a gain for the one who stays the course. If player 1 swerves, player 2's best strategy is to stay the course; and if player 1 stays the course, player 2's best strategy is to swerve.

## ECONOMICS IN ACTION

## A Game in the Market for Tissues

Anti-Viral Kleenex and Puffs Plus Lotion didn't get developed because Kimberly-Clark (Kleenex) and P\&G (Puffs) were thinking about helping you cope with a miserable cold. These new-style tissues and other innovations in the quality of facial tissues are the product of a costly research and development ( $\mathrm{R} \& \mathrm{D}$ ) game.

The table below illustrates the game (with hypothetical numbers). Each firm can spend either $\$ 25$ million or nothing on R\&D. If neither firm spends, KimberlyClark makes an economic profit of $\$ 70$ million and P\&G of $\$ 30$ million (bottom right). If each firm spends on R\&D, Kimberly-Clark's economic profit is $\$ 45$ million and P\&G's is $\$ 5$ million (top left). The other parts of the matrix show the economic profits for each when one spends on R\&D and the other doesn't.

Confronted with these payoffs, Kimberly-Clark sees that it gets a bigger profit if it spends on R\&D regardless of what P\&G does. P\&G reaches the same conclusion: It, too, gets a bigger profit by spending on R\&D regardless of what Kimberly-Clark does.

Because $\mathrm{R} \& \mathrm{D}$ is the best strategy for both players, it is the Nash equilibrium-a dominant-strategy Nash equilibrium.

The outcome of this game is that both firms conduct R\&D. They make less profit than they would if they could collude to achieve the cooperative outcome of no R\&D. But you get a better Kleenex or Puffs tissue.


Kleenex Versus Puffs: An R\&D Game

AnEconomic ExampleofChicken Aneconomic game of chicken can arise when research and development (R\&D) creates a new technology that cannot be kept secret or patented, so both firms benefit from the R\&D of either firm. The chicken in this case isthe firm that does the R\&D.
Suppose, for example, that either Apple or Nokia spends $\$ 9$ million developing a new touchscreen technology that both would end up being able to use regardless of which of them developed it.
Table 15.3 illustrates a payoff matrix for the game that Apple and Nokia play. Each firm has two strategies: Do the R\&D ("chicken out") or do not do theR\&D. Each entry shows the additional profit (theprofit from the new technology minus the cost of the research), given the strategies adopted.
If neither firm does the R\&D, each makes zero additional profit. If both firms conduct the R\&D, each firm makes an additional \$5 million. If one ofthe firms does the R\&D ("chickens out"), the chicken makes $\$ 1$ million and the other firm makes $\$ 10$ million. Confronted with these payoffs the two firms calculate their best strategies. Nokia is better off doing R\&D if Apple does no R\&D. Apple is better off doing R\&D if Nokia does no R\&D. There are two Nash equilibrium outcomes: Only one of them does the R\&D, but we can't predict which one.
You can see that an outcome with no firm doing R\&D isn't a Nash equilibrium because one firm would be better off doing it. Also both firms doing R\&D isn't a Nash equilibrium because one firm would be better off not doing it. To decide which firm does the R\&D, the firms might toss a coin, called a mixed strategy. (Ref: Table 15.3 Pg. 389)

## Repeated Games and Sequential Games

The games that we've studied are played just once. In contrast, many real-world games are played repeatedly. This feature of games turns out to enable real-world duopolists to cooperate, collude, and make a monopoly profit.
Another feature of the games that we've studied is that the players move simultaneously. But in many real-world situations, one player moves first and then the other moves-the play is sequential rather than simultaneous. This feature of real-world games creates a large number of possible outcomes.
We're now going to examine these two aspects of strategic decision making.

## A Repeated Duopoly Game

If two firms play a game repeatedly, one firm has the opportunity to penalize the other for previous "bad" behavior. IfGear cheats this week, perhaps Trick will cheat next week. Before Gear cheats this week, won't it consider the possibility that Trick will cheat next week? What is the equilibrium of this game?
Actually, there is more than one possibility. One is the Nash equilibrium that we have just analyzed. Both players cheat, and each makes zero economic profit forever. In such a situation, it will never pay one ofthe players to start complying unilaterally because to do so would result in a loss for that player and a profit for the other. But a cooperative equilibrium in which the players make and share the monopoly profit is possible. A cooperative equilibrium might occur if cheating is punished. There are two extremes of punishment.

The smallest penalty is called "tit for tat." A tit-for-tat strategy is one in which a player cooperates in the cur- rent period if the other player cooperated in the previous period, but cheats in the current period if the other player cheated in the previous period. The most severe form of punishment is called a trigger strategy. A triggerstrategy is one in which a player cooperates if the other player cooperates but plays the Nash equilibrium strategy forever thereafter if the other player cheats.
In the duopoly game between Gear and Trick, a tit-for-tat strategy keeps both players cooperating and making monopoly profits. Let's see why with an example Table 15.4 shows the economic profit that Trick and Gear will make over a number of periods under two alternative sequences of events: colluding and cheating with a tit-for-tat response by the other firm.
If both firms stick to the collusive agreement in period I, each makes an economic profit of $\$ 2$ mil- lion. Suppose that Trick contemplates cheating inperiod 1. The cheating produces a quick $\$ 4.5$ million economic profit and inflicts a $\$ 1$ million economic loss on Gear. But a cheat in period 1 produces a response from Gear in period 2. If Trick wants to getback into a profit-making situation, it must return to the agreement in period 2 even though it knows that Gear will punish it for cheating in period 1 . So in period 2 , Gear punishes Trick and Trick cooperates. Gear now makes an economic profit of $\$ 4.5$ million, and Trick incurs an economic loss of $\$ 1$ million.
Adding up the profits over two periods of play, Trick would have made more profit by cooperating-\$4 million compared with $\$ 3.5$ million.
What is true for Trick is also true for Gear. Because each firm makes a larger profit by sticking with the collusive agreement, both firms do so and the monopoly price, quantity, and profit prevail. In reality, whether a cartel works like a one-play game or a repeated game depends primarily on the number of players and the ease of detecting and punishing cheating. The larger the number of players, the harder it is to maintain a cartel.

Games and Price Wars A repeated duopoly game can help us understand real-world behavior and, in particular, price wars. Some price wars can be interpreted as the implementation of a tit-for-tat strategy. But the game is a bit more complicated than the one we've looked at because the players are uncertain about the demand for the product. Playing a tit-for-tat strategy, firms have an incentive to stick to the monopoly price. But fluctuations in demand lead to fluctuations in the monopolyprice, and sometimes, when the price changes, it might seem to one of the firms that the price has fallen because the other has cheated. In this case, a price war will break out. The price war will end only when each firm is satisfied that the other is ready to cooperate again. There will be cycles of price wars and the restoration of collusive agreements. Fluctuations in the world price of oil might be interpreted in this way.

TABLE 15.4 Cheating with Punishment

|  | Collude |  |  | Cheat with <br> tit-for-tat |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Period <br> of play | Trick's <br> profit <br> (millions of dollors) | Gear's <br> profit |  | Trick's <br> profit <br> (millions of dollars) | Gear's <br> profit |
| 1 | 2 | 2 |  | 4.5 | -1.0 |
| 2 | 2 | 2 |  | -1.0 | 4.5 |
| 3 | 2 | 2 |  | 2.0 | 2.0 |
| 4 | $\vdots$ | $\vdots$ |  | $\vdots$ | $\vdots$ |

If duopolists repeatedly collude, each makes a profit of $\$ 2$ million per period of play. If one player cheats in period 1, the other player plays a tit-for-tat strategy and cheats in period 2. The profit from cheating can be made for only one period and must be paid for in the next period by incurring a loss. Over two periods of play, the best that a duopolist can achieve by cheating is a profit of $\$ 3.5$ million, compared to an economic profit of $\$ 4$ million by colluding.

Some price wars arise from the entry of a small number of firms into an industry that had previously been a monopoly. Although the industry has a small number of firms, the firms are in a prisoners' dilemma and they cannot impose effective penalties for price cutting. The behavior of prices and outputs in the computer chip industry during 1995 and 1996 can be explained in this way. Until 1995, the market for Pentium chips for IBM-compatible computers was dominated by one firm, Intel Corporation, which was able to make maximum economic profit by producing the quantity of chips at which marginal cost equaled marginal revenue. The price of Intel's chips was set to ensure that the quantity demanded equaled the quantity produced. Then in 1995 and 1996, with the entry of a small number ofnew firms, the industry became an oligopoly. If the firms had maintained Intel's price and shared the market, together they could have made economic profits equal to Intel's profit. But the firms were in a prisoners' dilemma, so prices fell toward the competitive level.
Let's now study a sequential game. There are many such games, and the one we'll examine is among the simplest. It has an interesting implication and it will give you the flavor of this type of game. The sequential game that we'll study is an entry game in a con- testable market.

## Sequential Entry Game in a Contestable Market

If two firms play a sequential game, one firm makes a decision at the first stage of the game and the other makes a decision at the second stage.
We're going to study a sequential game in a contestable market-a market in which firms can enterand leave so easily that firms in the market face com-petition from potential entrants. Examples of con-testable markets are routes served by airlines and by barge companies that operate on the major water-ways. These markets are contestable because firms could enter if an opportunity for economic profitarose and could exit with no penalty if the opportunity for economic profit disappeared.
If the Herfindahl-Hirschman Index (p. 272) is used to determine the degree of competition, a con- testable market appears to be uncompetitive. But a contestable market can behave as if it were perfectly competitive. To see why, let's look at an entry game for a contestable air route.

A Contestable Air Route Agile Air is the only firm operating on a particular route. Demand and cost conditions are such that there is room for only one airline to operate. Wanabe Inc. is another airline that could offer services on the route.
We describe the structure of a sequential game by using a game tree like that in Fig. 15.6. At the first stage, Agile Air must set a price. Once the price is set and advertised, Agile can't change it. That is, once set, Agile's price is fixed and Agile can't react to Wanabe's entry decision. Agile can set its price at either the monopoly level or the competitive level. At the second stage, Wanabe must decide whether to enter or to stay out. Customers have no loyalty (there are no frequent-flyer programs) and they buy from the lowest-price firm. So if Wanabe enters, it setsa price just below Agile's and takes all the business.
Figure 15.6 shows the payoffs from the various decisions (Agile's in the red triangles and Wanabe's inthe blue triangles).
To decide on its price, Agile's CEO reasons as follows: Suppose that Agile sets the monopoly price. If Wanabe enters, it earns 90 (think of all payoff numbers as thousands of dollars). If Wanabe stays out, it earns nothing. So Wanabe will enter. In thiscase Agile will lose 50. Now suppose that Agile sets the competitive price. If Wanabe stays out, it earns nothing, and if it enters, it loses 10 , so Wanabe will stay out. Inthis case, Agile will make zero economic profit.
Agile's best strategy is to set its price at the competitive level and make zero economic profit. The option of earning 100 by setting the monopoly price with Wanabe staying out is not available to Agile. If Agile sets the monopoly price, Wanabe enters, undercuts Agile, and takes all the business. (Ref: Fig. 15.6 Pg. 393)
In this example, Agile sets its price at the competitive level and makes zero economic profit: A less costly strategy, called limit pricing, sets the price at the highest level that inflicts a loss on the entrant. Any loss is big enough to deter entry, so it is not always necessary to set the price as low as the competitive price. In the example of Agile and Wanabe, at the competitive price, Wanabe incurs a loss of 10 if it enters. A smaller loss would still keep Wanabe out.
This game is interesting because it points to the possibility of a monopoly behaving like a competitive industry and serving the social interest without regulation. But the result is
not general and depends on one crucial feature of the setup of the game: At the second stage, Agile is locked in to the price set at the first stage.
If Agile could change its price in the second stage, it would want to set the monopoly price if Wanabe stayed out-100 with the monopoly price beats zero with the competitive price. But Wanabe can figure out what Agile would do, so the price set at the first stage has no effect on Wanabe. Agile sets the monopoly price and Wanabe might either stay out or enter.
We've looked at two of the many possible repeated and sequential games, and you've seen how these types of games can provide insights into the complex forces that determine prices and profits. So far, we've studied oligopoly with unregulated market power. Firms like Trick and Gear are free to collude to maximize their profit with no concern for the consumer or the law.
But when firms collude to achieve the monopoly outcome, they also have the same effects on efficiency and the social interest as monopoly. Profit is made at the expense of consumer surplus and a deadweight loss arises. Your next task is to see how U.S. antitrust law limits market power.

## Antitrust Law

Antitrust Law is the law that regulates oligopolies and prevents them from becoming monopolies or behaving like monopolies. Two government agencies cooperate to enforce the antitrust laws: the Federal Trade Commission and the Antitrust Division of the U.S. Department of Justice.

## The Antitrust Laws

The United States has two main antitrust laws:

- The Sherman Act, 1890
- The Clayton Act, 1914

The Sherman Act The Sherman Act made it a felony to create or attempt to create a monopoly or a cartel. During the 1880 s, lawmakers and the general public were outraged and disgusted by the practices of some of the big-name leaders of American business. The actions of J.P. Morgan, John D. Rockefeller, and WH. Vanderbilt led to them being called the "robber barons." It turns out that the most lurid stories of the actions of these great American capitalists were not of their creation of monopoly power to exploit consumers but of their actions to damage each other.
Nevertheless, monopolies that damaged the consumer interest did emerge. For example, John D. Rockefeller had a virtual monopoly in the market for oil.
Table 15.5 summarizes the two main provisions of the Sherman Act. Section 1 of the act is precise: Conspiring with others to restrict competition is illegal. But Section 2 is general and imprecise. Just whatis an "attempt to monopolize"?

The Clayton Act The Clayton Act, which was passed in response to a wave of mergers that occurred at the beginning of the twentieth century, provided the answer to the question left
danglingby the Sherman Act: It defined the "attempt to monopolize." The Clayton Act supplemented the Sherman Act and strengthened and clarified the antitrust law.

## TABLE 15.5 The Sherman Act of 1890

## Section 1:

Every contract, combination in the form of trust or otherwise, or conspiracy, in restraint of trade or commerce among the several States, or with foreign nations, is hereby declared to be illegal.

## Section 2:

Every person who shall monopolize, or attempt to monopolize, or combine or conspire with any other person or persons, to monopolize any part of the trade or commerce among the several States, or with foreign nations, shall be deemed guilty of a felony.

## table 15.6 The Clayton Act and Its Amendments

Clayton Act 1914

Robinson-Patman Act 1936

Celler-Kefauver Act 1950

These acts prohibit the following practices only if they substantially lessen competition or create monopoly:

1. Price discrimination
2. Contracts that require other goods to be bought from the same firm (called tying arrangements)
3. Contracts that require a firm to buy all its requirements of a particular item from a single firm (called requirements conitracts)
4. Contracts that prevent a firm from selling competing items (called exclusive dealing)
5. Contracts that prevent a buyer from reselling a product outside a specified area (called territorial confinement)
6. Acquiring a competitor's shares or assets
7. Becoming a director of a competing firm

When Congress passed the Clayton Act, it also established the Federal Trade Commission, the federal agency charged with the task of preventing monopoly practices that damage the consumer interest.
Two amendments to the Clayton Act-the Robinson-Patman Act of 1936 and the CellerKefauver Act of 1950-oudaw specific practices and provide even greater precision to the antitrust law.Table 15.6 describes these practices and summarizes the main provisions of these threeacts.

## Price Fixing Always Illegal

Colluding with competitors to fix the price is always a violation of the antitrust law. If the Justice Department can prove the existence of a price fixing cartel, also called a horizontal pricefixing agreement, defendants can offer no acceptable excuse.
The predictions of the effects of price fixing that you saw in the previous sections of this chapter pro-vide the reasons for the unqualified attitude toward price fixing. A duopoly cartel can maximize profit and behave like a monopoly. To achieve the monopoly outcome, the cartel restricts production and fixes the price at the monopoly level. The consumer suffers because consumer surplus shrinks. And the outcome is inefficient because a deadweight loss arises.
It is for these reasons that the law declares that all price fixing is illegal. No excuse can justify the practice.
Other practices raise antitrust concerns but are more controversial and generate debate among lawyers and economists. We'll examine three of these practices.

## Three Antitrust Policy Debates

Some practices that engender antitrust policy debate are

- Resale price maintenance
- Tying arrangements
- Predatory pricing

Resale Price Maintenance Most manufacturers sell their products to the final consumer indirectly through a wholesale and retail distribution system. Resale price maintenance occurs when a distributor agrees with a manufacturer to resell a product at orabove a specified minimum price.
A resale price maintenance agreement, also called a vertical pricefixing agreement, is not illegal under the Sherman Act provided it is not anticompetitive. Nor is it illegal for a manufacturer to refuse to supply a retailer who doesn't accept guidance on what the minimum price should be.
In 2007, the Supreme Court ruled that a handbag manufacturer could impose a minimum retail price on a Dallas store, Kay's Kloset. Since that. ruling, many manufacturers have imposed minimum retail prices. The practice is judged on a case-by-case basis. Does resale price maintenance create an inefficient or efficient use of resources? Economists can be found on both sides of this question.

Inefficient Resale price Maintenance Resale price maintenance is inefficient if it enables dealers to charge themonopoly price. By setting and enforcing the resale price, the manufacturer might be able to achieve the monopoly price.

Efficient Resale Price Maintenance Resale price maintenance might be efficient if it enables a manufacturer to induce dealers to provide the efficient standard ofservice. Suppose that SilkySkin wants shops to demonstrate the use of its new unbelievable moisturizingcream in an inviting space. With resale price maintenance, SilkySkin can offer all the retailers the same incentive and compensation. Without resale price maintenance, a cut-price drug store might offer SilkySkin products at a low price. Buyers would then have an incentive to visit a highprice shop for a product demonstration and then buy from the low-price shop. The low-price shop would be a free rider (like the consumer of a public good in Chapter 16, p. 412), andan inefficient level of service would be provided.
SilkySkin could pay a fee to retailers that provide good service and leave the resale price to be determined by the competitive forces of supply and demand. But it might be too costly for SilkySkin to monitor shops and ensure that they provide the desired level of service.

A Tying arrangement is an agreement to sell one product only if the buyer agrees to buy another, different product. With tying, the onlyway the buyer can get the one product is to also buy the other product. Microsoft has been accused of tying Internet Explorer and Windows. Textbook publishers sometimes tie a Web site and a textbook and force students to buy both. (You can't buy the bookyou're now reading, new, without the Web site. But you can buy the Web site access without the book, so these products are not tied.) Could textbook publishers make more money by tying a book and access to a Web site? The answer is sometimes but not always. Suppose that you and other students are willing to pay $\$ 80$ for a book and $\$ 20$ for access to a Web site. The publisher can sell these items separately for these prices or bundled for $\$ 100$. The publisher does not gain frombundling. But now suppose that you and only half of the students are willing to pay $\$ 80$ for a book and $\$ 20$ for a Web site and the other half of the students are willing to pay $\$ 80$ for a Web site and $\$ 20$ for a book. Now if the two items are sold separately, the publisher can charge $\$ 80$ for the book and $\$ 80$ for the Web site.
Half the students buy the book but not the Web site, and the other half buy the Web site but not the book. But if the book and Web site are bundled for $\$ 100$,everyone buys the bundle and the publisher makes an extra $\$ 20$ per student. In this case, bundling has enabled the publisher to price discriminate.
There is no simple, clear-cut test of whether a firm is engaging in tying or whether, by doing so, it has increased its market power and profit and created inefficiency.

Predatory Pricing is setting a low price to drive competitors out of business with the intention of setting a monopoly price when the competition has gone. John D. Rockefeller's StandardOil Company was the first to be accused of this practice in the 1890s, and it has been claimed often in antitrustcases since then. Predatory pricing is an attempt to create a monopoly and as such it is illegal under Section 2 of the Sherman Act.

It is easy to see that predatory pricing is an idea, not a reality. Economists are skeptical that predatory pricing occurs. They point out that a firm that cuts its price below the profitmaximizing level loses during the low-price period. Even if it succeeds in driving its competitors out of business, new competitors will enter as soon as the price is increased, so any potential gain from a monopoly position is temporary. A high and certain loss is a poor exchange for a temporary and uncertain gain. No case of predatory pricing has beendefinitively found.

## ECONOMICS IN ACTION

## The United States Versus Microsoft

In 1998, the Antitrust Division of the U.S. Department of Justice along with the Departments of Justice of a number of states charged Microsoft, the world's largest producer of software for personal computers, with violations of both sections of the Sherman Act.

A 78-day trial followed that pitched two prominent MIT economics professors against each other, Franklin Fisher for the government and Richard Schmalensee for Microsoft.

The Case Against Microsoft The claims against Microsoft were that it

- Possessed monopoly power.
- Used predatory pricing and tying arrangements.
- Used other anticompetitive practices.

It was claimed that with 80 percent of the market for PC operating systems, Microsoft had excessive monopoly power. This monopoly power arose from two barriers to entry: economies of scale and network economies. Microsoft's average total cost falls as production increases (economies of scale) because the fixed cost of developing an operating system such as Windows is large while the marginal cost of producing one copy of Windows is small. Further, as the number of Windows users increases, the range of Windows applications expands (network economies), so a potential competitor would need to produce not
only a competing operating system but also an entire range of supporting applications as well.

When Microsoft entered the Web browser market with its Internet Explorer, it offered the browser for a zero price. This price was viewed as predatory pricing. Microsoft integrated Internet Explorer with Windows so that anyone who uses this operating system would not need a separate browser such as Netscape Navigator. Microsoft's competitors claimed that this practice was an illegal tying arrangement.

Microsoff's Response Microsoft challenged all these claims. It said that Windows was vulnerable to competition from other operating systems such as Linux and Apple's Mac OS and that there was a permanent threat of competition from new entrants.

Microsoft claimed that integrating Internet Explores with Windows provided a single, unified product of greater consumer value like a refrigerator with a chilled water dispenser or an automobile with a CD player.

The Outcome The court agreed that Microsoft was in violation of the Sherman Act and ordered that it be broken into two firms: an operating systems produces and an applications producer. Microsoft successfully appealed this order. In the final judgment, though, Microsoft was ordered to disclose to other software developers details of how its operating system works, so that they could compete effectively against Microsoft. In the summer of 2002, Microsoft began to comply with this order.

## Mergers and Acquisitions

Mergers, which occur when two or more firms agree to combine to create one larger firm, and acquisitions, which occur when one firm buys another firm, arecommon events. Mergers occurred when Chrysler andthe German auto producer Daimler-Benz combined toform DaimlerChrysler and when the Belgian beer producer InBev bought the U.S. brewing giant Anheuser-Busch and created a new combined company, Anheuser-Busch InBev. An acquisition occurred when Rupert Murdoch's News Corp bought Myspace.
The mergers and acquisitions that occur don't create a monopoly. But two (or more) firms might be tempted to try to merge so that they can gain market power and operate like a monopoly. If such a situation arises, the Federal Trade Commission (FTC) takes an interest in the move and stands ready to block the merger. To determine which mergers it will
examine and possibly block, the FTC uses guidelines, one of which is the HerfindahlHirschman Index (HHI) (see Chapter 10, pp. 272-273).
A market in which the HHI is less than 1,500 is regarded as competitive. An index between 1,500 and 2,500 indicates a moderately concentrated market, and a merger in this market that would increase the index by 100 points is challenged by the FTC. An index above 2,500 indicates a concentrated market, and a merger in this market that would increase the index by 200 points is generally blocked. You can see an application of these guide- lines in Economics inAction below.

## ECONOMICS IN ACTION <br> NoCellphone Service Merger

The FTC used its HHI guidelines (summarized in the figure) to block a proposed merger in the market for cellphone service. In 2011, AT\&T wanted to buy T-Mobile from the German Deutsch Telecommunication. But the market for cellphone service is highly concentrated. Verizon is the largest service provider with close to 40 percent of the market share. AT\&T has a 30 percent share and Sprint and T-Mobile about 12 percent each. Another 14 small firms share the remaining market. So the four largest firms in this market have a 94 percent market share. The HHI is around 2,800 .

The figure shows how the HHI would have changed with the merger of AT\&T and T-Mobile. Itwould have increased the HHI by around 700 points. With an HHI of 2,800 and an increase of 700 , there was a presumption that the merger would give AT\&T too much market power, so the FTC decided to block it.


Herfindahl-Hirschman Index (HHI)
Figure 1 Merger Guidelines

Oligopoly is a market structure that you often encounter in your daily life. Economics inthe News on pp. 398-399 looks at a game played in the market for cellphone service.

## ECONOMIC ANALYSIS

- The U.S. market in cellphone service is dominated by four firms: Verizon, AT\&T Mobility, Sprint, and T-Mobile.
- Figure 1 shows the shares in this market. You can see that Verizon has 35 percent of the market; AT\&T has 33 percent, Sprint has 16 percent; and T-Mabile has 14 percent. Another nine small firms share the remaining 2 percent.
- In 2013, T-Mobile and AT\&T increased the intensity of their competition.
- T-Mobile increased the value and cost of production of its plans but held prices, and so decreased its profit.
- AT\&T cut its prices, which decreased its profit.
- We can inferpret this competition as a prisoners' dilemma game.
- Table 1 shows the payoff matrix (millions of dollars of profit) for the game played by T-Mobile and AT\&T. (The numbers are hypothetical.)
- This game is a prisoners' dilemma like that on p. 383 and has a dominant-strategy Nash equilibrium.
- If T-Mobile offers an increased value of service, AT\&T avoids a larger loss by cutting its price; and if T-Mobile does nothing different (no change), AT\&T increases its profit by cutting its prices and gaining market share.
- So AT\&T's best strategy is to cut prices.
- If AT\&T cuts its prices, T-Mobile avoids a larger loss by incurring a higher cost of offering its customers greater value plans; and if AT\&T doesn't cut its price \{no changel,

T-Mobile increases profit by offering its customers greater value plans.

- So T-Mobile's best strategy is to incur the higher cost of offering its customers greater value plans.
- Because the best strategy for AT\&T is to cut price and for T-Mobile to incur the higher cost of offering its customers greater value plans, that is the equilibrium of the game.
- The firms are in a prisoners' dilemma because each would be befter off avoiding the contest on quality of service and price.
- But each firm can see that if it doesn't take its best acfion, the other firm will gain and it will lose.
- The game wéve fust described is played only once, and both firms move at the same time.
- But in the game played by T-Mobile and AT\&T, the firms move sequentially and repeatedly.
- If T-Mobile and AT\&T adopt a tit-for-tat punishment strategy, it is possible that the current war between them will end and they will return to higher prices and larger profits.
- This market is one that affects everyone, and one that will expand as cellphones become ever smarter and used for more data-intensive activities.
(Ref. to Fig. 1 \& Table 1 on page 399)


## PART FIVE: MARKET FAILURE AND GOVERNMENT

## CHAPTER 16: Public Choices, Public Goods, and Healthcare

After studying this chapter, you will be able to:

- Explain why some choices are public choices and how they are made in a political marketplace
- Explain how the free-rider problem arises and how the quantity of public goods is determined
- Explain why governments provide healthcare and how our healthcare markets work

The George Washington Bridge that links New York and New Jersey is one of 607,380 road bridges inthe United States built and maintained by governments. Healthcare along with a wide array of other goods and services are also provided by governments. But why governments? Why not private firms? Are governments efficient in their provision of bridges, healthcare, and other goods and services? These are the questions we study in this chapter. In Economics in the News at the end of the chapter, wereturn to bridges and look at the problem of keeping them safe.

## Public Choices

All economic choices are made by individuals, butsome choices are private and some are public. A private choice is a decision that has consequences for the person making it. Your decision to buy a textbook or work at McDonald's is a private choice. Apublic choice is a decision that has consequences for many people and perhaps for an entire society. The decision by members of Congress to pass the Affordable Care Act - Obamacare - and the decisionsthat created the enrollment website Healthcare.gov are examples of public choices. Some of the things we care most about are decided by people making public choices, and very large quantities of scarce resources get used as a result of these choices. Economics in Action opposite provides a snapshot of what our federal, state, and local governments buy. Why do governments allocate a large quantity of resources? The economic theory of government answers this question.

## Why Governments?

Governments perform three economic functions: They establish and maintain property rights, provide nonmarket mechanisms for allocating scarce resources, and redistribute income and wealth.
Property rights and the legal system that enforces them are the foundation of the market economy, which, in many situations, functions well and allocates scarce resources efficiently. But sometimes the market fails (see Chapter 5, pp. 152-153).
When market failure occurs, choices made in the pursuit of self-interest have not served the social interest. The market economy also delivers a distribution of income and wealth that most people regard as unfair. Equity requires some redistribution.

Replacing markets with government resource- allocation decisions is no simple matter. Just as there can be market failure, there can also be governmentfailure. Government failure is a situation in which government actions lead to inefficiency-to either underprovision or overprovision. Government failure can arise because government is made up of many individuals, each with their own economic objectives. Public choices are theoutcome of the choices made by these individuals. To analyze these choices, economists have developed a public choice theory of the political marketplace.

Public Choice and the Political Marketplace
The four groups of decision makers shown in Fig. 16.1 interact in the political marketplace. They are

- Voters
- Firms
- Politicians
- Bureaucrats


## ECONOMICS IN ACTION

## The Goods and Services Provided by Governments

The U.S. federal government spent $\$ 2$ trillion on goods and services in 2013. The state and local governments spent another $\$ 1.8$ trillion. The total of $\$ 3.8$ trillion is one quarter of all expenditure in the U.S. economy.

The figure shows what our governments buy. The biggest item is healthcare, with education not far behind. National defense comes next. The total spent on the nation's highways and bridges (transportation), police and fire services (protection), and government itself is about one fifth of government expenditure.


What Governments Buy

Voters evaluate politicians' policy proposals, benefit from public goods and services, and pay some of the taxes. They support the politicians whose policy proposals make them better off and express. their demand for public goods and services by voting, helping in political campaigns, lobbying, and making campaign contributions.

Firms also evaluate politicians' policy proposals, benefit from public goods and services, and paysome of the taxes. Although firms don't vote, they do make campaign contributions and are a major source of funds for political parties. Firms also engage in lobbying activity to persuade politicians to propose policies that benefit them.

Politicians are the elected persons in the federal, state, and local governments. The goal of a politician is to get elected and to remain in office. Votes to a politician are like profit to a firm. Politicians also direct bureaucrats.

FIGURE 16.1 The Political Marketplace


Voters and firms demand public goods and services and policies that serve their self-inferest.

Voters support the policies they like with their votes. And voters and firms support policies they like with campaign contributions and by lobbying.

Politicians and bureaucrats supply public goods and services, set taxes, and make policy proposals that serve their self-interest.

Politicians want to altract enough votes to get them elected and keep them in office. Bureaucrats want to get the largest possible budget for their departments.

A political equilibrium balances all these public choices.

Bureaucrats are the public servants who work in government departments. They administer tax collection, the delivery of public goods and services, and the administration of rules and regulations.
The self-interest of a bureaucrat is served when the budget of her or his department is maximized budget of a department, the greater is the prestige of its chief and the greater are the opportunities for promotion for people farther down he bureaucratic ladder. This economic assumption implies that in doing what they perceive to be a good Job in the public interest, they take care of their own self-interest too.

## Political Equilibrium

Voters, firms, politicians, and bureaucrats make theireconomic choices in their own selfinterest. Public choices, like private choices, are constrained by what is feasible. Each person's public choices are also constrained by the public choices of others.
The balance of forces in the political marketplace determines the outcome of all the public choices that people make. Ina political equilibrium the choices of voters, firms, politicians, and bureaucrats are all compatible and no group can see a way of improving its position by making a different choice.
Ideally, the political equilibrium achieves allocative efficiency and serves the social interest, but this out- come is not guaranteed, as you'll see in this chapter.
We make public choices because some situations just don't permit private choices. The core of the reason we can't always make private choices is that some goods and services (and some factors of production) have a public nature-they are public goods and services. Your next task is to see exactly what we mean by a public good or service.

## What is a Public Good?

To see what makes a good a public good, we distinguish between two features of all goods: the extent to which people can be excluded from consuming them and the extent to which one person's consumption rivals someone else's consumption.

Excludable A good is excludable if it is possible to prevent someone from enjoying its benefits. Brink'ssecurity services, East Point Seafood's fish, and a U2 concert are examples. People must pay to benefit from them. A good is nonexcludable if it is impossible (or extremely costly) to prevent anyone from benefiting from it. The services of the LAPD, fish in the Pacific Ocean, and a concert on network television are examples. When an LAPD cruiser enforces the speed limit, everyone on the highway benefits; anyone with a boat can fish in the ocean; and anyone with a TV can watch a network broadcast.

Rival A good is rival if one person's use of it decreases the quantity available for someone else. A Brink's truck can't deliver cash to two banks at the same time. A fish can be consumed only once.
A good is nonrival if one person's use of it does not decrease the quantity available for someone else. Theservices of the LAPD and a concert on network television are nonrival. One person's benefit doesn't lower the benefit of others.

## A Fourfold Classification

Figure 16.2 classifies goods, services, and resources into four types.

Private Goods A private good is both rival and excludable. A can of Coke and a fish on East Point Seafood's farm are examples of private goods.

Public Goods A public good is both nonrival and nonexcludable. A public good simultaneously benefits everyone, and no one can be excluded from its benefits. National defense is the best example of apublic good.

Common Resources Acommon resource is rival and nonexcludable. A unit of a: common resource can be used only once, but no one can be prevented from using what is available. Ocean fish are a common resource. They are rival because a fish taken by one person isn't available for anyone else, and they are nonexcludable because it is difficult to prevent people from catching them.

Natural Monopoly Goods Anatural monopoly good is nonrival but excludable. Potential consumers can be excluded if they don't pay but adding one more user doesn't rival other users, so marginal cost is zero. Examples of natural monopoly goods are the Internet cable television and an uncongested bridge or tunnel.
Why is a nonrival but excludable good a natural monopoly good? It is because there is a fixed cost of producing it. With a zero marginal cost, average total ; cost falls as output increases so economies of scale exist over the entire range of output for which there is a demand, and one firm can produce the good at a." lower cost than can two or more firms (see p. 336).

## FIGURE 16.2 Fourfold Classification of Goods



A private good is rival and excludable: You must pay to get it and you alone enjoy it. A public good is nonrival and nonexcludable: You and everyone else enjoy it without paying for it. A common resource is rival but nonexcludable. And a good that is nonrival but excludable is produced by a natural monopoly.

## The Things Our Governments Buy

Of the things that our governments buy, shown in Economics in Action on p. 408, national defense, protection, and constructing and maintaining the transportation infrastructure fit the definition of a public good. They are nonrival and nonexcludable. But what about healthcare and education, the two biggest items? They don't look like public goods. A person can be excluded from a hospital or college. And one person's use of a hospital bed
or place in college rivals another's. So aren't healthcare and education private goods? Why do governments provide them?

## ECONOMICS IN ACTION

## Is a Lighthouse a Public Good?

Built on Little Brewster Island in 1716 to guide ships into and out of the Boston Harbor, Boston Lighthouse was the first light station in North America.

For two centuries, economists used the lighthouse as an example of a public good. No one can be prevented from seeing its warning light-nonexcludable-and one person seeing its light doesn't prevent someone else from doing so too-nonrival.

Ronald Coase, who won the 1991 Nobel Prize for ideas he first developed when he was an undergraduate at the London School of Economics, discovered that before the nineteenth century, lighthouses in England were built and operated by private corporations that earned profits by charging tolls on ships docking at nearby ports. A ship that refused to pay the lighthouse toll was excluded from the port.

So the benefit arising from the services of a lighthouse is excludable. Because the services provided by a lighthouse are nonrival but excludable, a lighthouse is an example of a natural monopoly good and not a public good.


Healthcare is two goods: care supplied by doctors and other professionals, and insurance. Governments provide both healthcare services and health insurance because, left to the market alone; they would be inefficiently underprovided and unfairly distributed. For a majority of voters, it is unfair and unethical to allocate scarce healthcare resources only to those who can afford them-those who are willing and able to pay. Most people want the government to pro-vide a single mother who is living in poverty with access to a doctor or hospital when she needs those services.
To work well, a market needs buyers and sellers who are well informed about the item being traded. In thecase of healthcare services, doctors and healthcare professionals are better informed than patients about the service and treatment needed. And in the case of health insurance, the insured is better informed than theinsurer about the risks faced. Also, _lacking the relevant information about risk, fit and healthy people often take too short a view and don't buy enough insurance.

Education Governments provide public education because it brings benefits that spill over to others- called external benefits. Everyone benefits from living in an educated society where friends and neighbors share a common heritage.
The rest of this chapter studies the public choices that must be made to avoid the underprovision of public goods and healthcare. Chapter 17 studiesthe ways of coping with externalities, including those arising from education, and the challenge of conserving common resources.

## Providing Public Goods

Why do governments provide firefighting services? Why don't the people of California buy brush fire-fighting services from Firestorm, a private firm thatcompetes for our dollars in the marketplace in the same way that McDonald's does? The answer is thatfirefighting is a public good. It is nonexcludable and nonrival and it has a free-rider problem.

## The free-Rider Problem

A free rider enjoys the benefits of a good or service without paying for it. Because a public good is pro- vided for everyone to use and no one can be excluded from its benefits, no one has an incentive to pay his or her share of the cost. Everyone has an incentive tofree ride. The free-rider problem is that theeconomy would provide an inefficiently small quantity of a public good. Marginal social benefit from the publicgood would exceed its marginal social cost and a deadweight loss would arise. Let's look at the marginal social benefit and marginal social cost of a public good.

## Marginal Social Benefit from a Public Good

Lisa and Max (the only people in a society) value fire-fighting airplanes. Figures 16.3(a) and 16.3(b) graph their marginal benefits from the airplanes as MBL for Lisa and MBM for Max. The marginal benefit from a public good (like that from a private good) diminishes as the quantity of the good increases.
Figure 16.3(c) shows the marginal social benefit curve, MSB. Because everyone gets the same quantity of a public good, its marginal social benefit is the sum of the marginal benefits of all the individuals at each quantity. So the marginal social benefit curve is the vertical sum of the individual marginal benefit curves. The curve MSB is the marginal social benefit curve for the economy made up of Lisa and Max.
For each airplane, Lisa's marginal benefit is added to Max's marginal benefit. Contrast the MSB curve for a public good with that of a private good. To obtain the economy's MSB curve for a private good, we sum thequantities demanded by all the individuals at each price-we sum the individual marginal benefit curves horizon- tally (see Chapter 5, pp. 146147).

## Marginal Social Cost of a Public Good

The marginal social cost of a public good is determined in exactly the same way as that of a private good-see Chapter 5, p. 148. The principle of increasing marginal cost applies to the marginal cost of a public good, so the marginal social cost increases as the quantity of the public good increases.

FIGURE 16.3 Benefits of a Public Good

(a) Lisa's marginal benefit

(b) Max's marginal benefit

(c) Economy's marginal social benefit

The marginal social benefit at each quantity of the public good is the sum of the marginal benefits of all individuals. The marginal benefit curves are $M B_{L}$ for Lisa and $M B_{M}$ for Max. The economy's marginal social curve is MSB.

## Efficient Quantity of a Public Good

To determine the efficient quantity of a public good, we use the principles that you learned in Chapter 5. The efficient quantity is that at which marginal social benefit equals marginal social cost.
Figure 16.4 shows the marginal social benefit curve, $M S B$, and the marginal social cost curve, MSC, for firefighting airplanes. (We'll now think of society as consisting of Lisa and Max and the other 39 million Californians.)
If marginal social benefit exceeds marginal social cost, as it does with 2 airplanes, resources can be used more efficiently by increasing the number of airplanes. The extra benefit exceeds the extra cost. If marginal social cost exceeds marginal social benefit, as it does with 4 airplanes, resources can be used more efficiently by decreasing the number of airplanes. The cost saving exceeds the loss of benefit.
If marginal social benefit equals marginal social cost, as it does with 3 airplanes, resources are allocated efficiently. Resources cannot be used more efficiently because to provide more than 3 airplanes increases cost by more than the extra benefit, and to provide fewer airplanes lowers the benefit by more than the cost saving.

## Inefficient Private Provision

Could a private firm-Firestorm-deliver the efficient quantity offirefighting airplanes? Most likely it couldn't, because no one would have an incentive to buy his or her share of the airplanes. Everyone would reason as follows:The number of airplanes provided by Firestorm is not affected by my decision to pay my share or not. But my own private consumption will be greater if I free ride and do not pay my share of the cost of the airplanes. If I don't pay, I enjoy the same level of fire protection and I can buy more private goods. I will spend my money on private goods and free ride on fire protection. Such reasoning is the free-rider problem. If everyone reasons the same way, Firestorm has no revenue and so provides no airplanes. Because the efficient number of airplanes is 3, private provision is inefficient.
Contemplating this outcome, the Fears realize that they are too fearful to get elected. They figure that, if they scale back to 3 airplanes, they will win the election if the Hopes stick with 2. The Hopes reason in a similar way and figure that, if they increase the number of airplanes to 3 , they can win the election if the Fears propose 4.
So they both propose 3 airplanes. The voters are indifferent between the parties, and each party receives 50 percent of the vote. But regardless of which party wins the election, 3 airplanes are pro- vided and this quantity is efficient. Competition in the political marketplace results in the efficient provision of a public good.

The Principle of Minimum Differentiation The principle of minimum differentiation is the tendency for competitors (including political parties) to make themselves similar to appeal to the maximum number of clients or voters. This principle describes thebehavior of political parties. It also explains why fast- food restaurants cluster in the same block. For example, if Domino's opens a new pizza outlet, it is likelythat Pizza Hut will soon open nearby.

## Efficient Public Provision

The outcome of the political process might be efficient or inefficient. We look first at an efficient out-come. There are two political parties: Fears and Hopes. They agree on all issues except the number of firefighting airplanes: The Fears want 4, and the Hopes want 2. Both parties want to get elected, sothey run a voter survey and discover the marginal social benefit curve of Fig. 16.5. They also consultwith airplane producers to establish the marginal cost curve. The parties then do a "what-if" analysis. If the Fears propose 4 airplanes and the Hopes propose 2 , the voters will be equally unhappy with both parties. Compared to the efficient quantity, the Hopes want an underprovision of 1 airplane and the Fears want anoverprovision of 1 airplane. The deadweight losses areequal and the election would be too close to call.

FIGURE 16.4 The Efficient Quantity of a Public Good


With fewer than 3 airplanes, marginal social benefit, MSB, exceeds marginal social cost, MSC. With more than 3 airplanes, MSC exceeds MSB. Only with 3 airplanes is MSC equal to MSB and the number of airplanes is efficient.
Contemplating this outcome, the Fears realize that they are too fearful to get elected. They figure that, if they scale back to 3 airplanes, they will win the election if the Hopes stick with 2. The Hopes reason in a similar way and figure that, if they increase the number of airplanes to 3, they can win the election if the Fears propose 4.
So they both propose 3 airplanes. The voters are indifferent between the parties, and each party receives 50 percent of the vote. But regardless of which party wins the election, 3 airplanes are pro- vided and this quantity is efficient. Competition in the political marketplace results in the efficient provision of a public good.

For the political process to deliver the efficient outcome, voters must be well informed, evaluate thealternatives, and vote in the election. Political parties must be well informed about voter preferences. As the next section shows, we can't expect to achieve this outcome.

## Inefficient Public Overprovision

If competition between two political parties is to deliver the efficient quantity of a public good, bureaucrats must cooperate and help to achieve this outcome. But bureaucrats might have a different idea and end up frustrating rather than facilitating an efficient outcome. Their actions might bring government failure.

FIGURE 16.5 An Efficient Political Outcome


The Hopes would like to provide 2 airplanes and the Fears would like to provide 4 airplanes. The political outcome is 3 airplanes because unless each party proposes 3 airplanes, the other party will beat it in an election.
Objective of Bureaucrats Bureaucrats want to maximize their department's budget because a bigger budget brings greater status and more power. So the Emergency Services Department's objective is to maximize the budget for firefighting airplanes. Figure 16.6 shows the outcome if the bureaucrats are successful in the pursuit of their goal. They mighttry to persuade the politicians that 3 airplanes cost more than the originally budgeted amount; or they might press their position more strongly and argue for more than 3 airplanes. In Fig. 16.6, the Emergency Services Department persuades the politicians to provide 4 airplanes.
Why don't the politicians block the bureaucrats? Won't overproviding airplanes cost future votes? It will if voters are well informed and know what is best for them. But voters might not be well informed, and well-informed interest groups might enable the bureaucrats to achieve their objective and overcome the objections of the politicians.

Rational Ignorance A principle of the economic analysis of public choices is that it is rational for avoter to be ignorant about an issue unless that issuehas a perceptible effect on the voter's economic welfare. Each voter knows that he or she can make virtually no difference to the fire protection policy of thegovernment of California and that it would take an enormous amount of time and effort to become even moderately well informed about alternative fireprotection technologies. Rationally uninformed voters enable bureaucrats and special interest groups to overprovide public goods.

FIGURE 16.6 Bureaucratic Overprovision


Well-informed bureaucrats want to maximize their budget and rationally ignorant voters enable the bureaucrats to go some way toward achieving their goal. A public good might be inefficiently overprovided with a deadweight loss.

## The Economics of Healthcare

Governments spend more on healthcare than on any other item. (see Economics inAction on p. 408). The key reason why governments play a large role in healthcare is that it would be underprovided and unfairly distributed if left to the market alone.
You are now going to see how governments influence the provision of healthcare, how healthcare markets in the United States compare with those in some other countries, and how it might be possible to improve on our current healthcare programs.
We begin by seeing why the market would underprovide healthcare-why there would be ahealthcare market failure.

## ECONOMICS IN ACTION

## Fighting Colorado's Wildfires

The 2012 wildfire season was extreme. Colorado had its worst season in a decade with major fires that devastated communities in and around Colorado Springs. And across the Western states, close to half a million acres ( 780 square miles) of land were burned.

Wildfires are natural and vital for the ecosystem, but some fires are started by human action, and some both human-made and naturally occurring fires, like the 2012 fires in Colorado, burn close to where people live. So protection against wildfires is a vital public good.

But not all fire protection services are produced by government. Private wildfire service contractors supply around 40 percent of wildfire suppression services. So fighting wildfires is an example of a public good that is provided by government and paid for with tax revenues, but in part produced by private firms.

Firestorm Wildfire Suppression Inc. is one such firm. Operating from Chico, CA, Firestorm hires and trains firefighters and produces firefighting services to maximize its profit. To achieve this goal, the firm must produce firefighting services at the lowest possible cost.

But if Firestorm (and its competitors) tried to sell their services to each individual homeowner in the wildfire regions, they wouldn't get enough revenue to remain in business. There would be a free-rider problem. The free-rider problem is avoided because governments buy the services of Firestorm.


## Healthcare Market Failure

Healthcare consists of healthcare services (the services of physicians, specialists, nurses, other professionals, and hospitals) and health insurance. Both would be underprovided and unfairly distributed without some government action.
The social benefit of healthcare exceeds the benefit perceived by its consumers, which means that the marginal social benefit of healthcare exceeds the willingness and ability to pay for it. Three problems make healthcare a special good: Consumers and potential consumers of healthcare

- Underestimate its benefit
- Underestimate their future needs
- Can't afford the care they need

Underestimate Benefit People don't have enough in- formation to value the benefit of healthcare correctly.Most people lack the medical knowledge to determinetheir treatment needs. And many (especially healthy and young people) optimistically underestimate the health risks that they face. So they undervalue the insurance policies that can help them pay for healthcare; and they undervalue the healthcare resources that stand ready to help them when needed.

Underestimate Future Needs People take too short a view of the benefits of healthcare. The young and healthy know that they will become old and less healthy and will likely become big consumers of healthcare. But the time horizon over which theyplan doesn't stretch that far into the future. The end result is that many people perceive too small a marginal benefit from health insurance, and are not willing to pay what it is actually worth to them.

Can't Afford For many people, the price of health insurance is beyond their ability to pay for it. Two groups of people are unable to afford adequate health insurance: those with a longterm health problem and the aged. But these are the people with the greatest need for healthcare.
Most people want to live in a society that treats these less healthy, older, and poorer people with com-passion and ensure that they are provided with accessto affordable healthcare. So there is an additional social benefit from healthcare for less healthy, older, and poorer people. Because the marginal social benefit of healthcare exceeds the marginal benefit perceived by its consumers, a competitive market in healthcare would underprovide it. And this underprovision has two dimensions. It is

- Inefficient, and
- Unfair

Inefficient Figure 16.7 illustrates the inefficient underprovision of a competitive market in healthcare.
The curve $D=M B$ shows the demand for healthcare, which is determined by the willingness and ability to pay for it, which in turn is determined by its perceived marginal benefit, $M B$.
The curve $S=$ MSC shows the supply of health-
care. The supply curve is also the marginal cost curve, which we will assume correctly measures the marginal social cost, MSC.
The equilibrium quantity of healthcare is at the intersection of the demand curve, $D=M B$, and the supply curve, $S=M S C$. In Fig. 16.7, that quantity is 0.3 billion patients per year. The curve MSB shows the marginal social benefit of healthcare, which, for the reasons explained above, exceeds the willingness and ability to pay.
The efficient quantity of healthcare is that at which MSB $=$ MSC. In Fig. 16.7, that quantity is 0.7 billion patients per year.

Because the efficient quantity exceeds the market quantity, the market underprovides healthcare. The marginal social benefit that patients who fail to get care would generate exceeds the marginal social cost of providing care, so the underprovision creates the deadweight loss shown by the gray triangle.

FIGURE 16.7 Healthcare Market Failure


The demand curve, $D=M B$, shows the marginal benefit and willingness to pay for healthcare, and the supply curve, $S=$ MSC, shows the marginal social cost of providing healthcare. A competitive market would provide care for 0.3 billion patients. The curve MSB shows the marginal social benefit of healthcare. The efficient quantity is that at which MSB = MSC and is 0.7 billion patients. The gray triangle shows the deadweight loss from underprovision.
Unfair Not only is there deadweight loss, but also there is an unfair distribution of the loss. The peoplewho receive healthcare are the ones who are willingand able to pay for it. They are people who are normally fit and healthy and able to earn an average or better than average income. The people who don't receive healthcare and who bear the deadweight loss are the long-term sick and the aged who can't afford the cost of healthcare.

## ECONOMICS IN ACTION

## U.S. Healthcare Expenditures in Global Perspective

The best U.S. healthcare is the best in the world, but it is costly. Americans spend 18 percent of income-almost $\$ 9,000$ per person per year-on healthcare. This amount is more than double the average spent in other rich countries. And this expenditure is projected to rise as the U.S. population ages and the "baby boom" generation retires.

The figure below compares U.S. healthcare costs with those of 13 other countries: some rich and somenot so rich.

The figure shows that even rich countries such as Canada, Germany, and France spend barely a half as much, per person, on healthcare as do Americans. And health outcomes measured by life expectancy and quality of health are as good as or better in these other rich countries than those in the United States.

Another feature of U.S. healthcare expenditure is its very large public component (the blue bars in the figure). Government expenditure per person on healthcare in the United States is the second highest in the world, exceeded only by that of the Netherlands. It exceeds that in Canada, where selling private health insurance is illegal, and in Germany and France, where there is a greater acceptance of high taxes and big government than in the United States.


Because the market would deliver an inefficient and unfair outcome, healthcare is provided by the public choices of governments. But there arealternative methods and scales of public provision, so governments must consider the effects of the alternatives in making their choices.

## Alternative Public Choice Solutions

We're now going to examine the public choices that deliver healthcare services. You can see in Economics in Action on this page that there is a wide range of levels of public funding of healthcare. Public choices pay 83 percent of the healthcare bills in the United Kingdom, 70 percent in Canada, and 46 percent in the United States. But among these three countries, public expenditure per person is highest in the United States.
Different public choice solutions lead to these differences in healthcare expenditures. And the different solutions have implications for both the efficiency and the cost of healthcare. We examine three approaches to supplementing or replacing the market: one used in Canada and the United Kingdom and two used in the United States. The three approaches are

- Universal coverage, single payer
- Private and government insurance
- Subsidized private insurance: Obamacare

Universal Coverage, Single Payer In Canada and the United Kingdom, healthcare is provided by a sys- tem with two key features: universal coverage and asingle payer.
Universal coverage means that everyone is covered by health insurance, with no exceptions and no excluded preconditions.
Single payer means that the government alone pays the healthcare bills. The government pays the doctors, nurses, and hospitals, and it pays for prescription drugs.
In the United Kingdom, most of the doctors are employed by the government and most of the hospitals are government owned. In Canada, the doc- tors and hospitals are independent private agents, but they may not sell their services other than to the government. Because the government is the sole buyer of healthcare services, it chooses the quantity of careto supply. A public choice, not market equilibrium, determines the quantity of healthcare service.
Patients access healthcare services at a zero (or low) price, so the quantity demanded is that at which the marginal benefit is zero (or low). But the demand forhealthcare has no direct effect on the quantity actually available, which is determined by the government's supply decision. The quantity demanded exceeds the quantity sup- plied and in the absence of a market price to allocatethe scarce resources, services are allocated on a first- come, first-served basis (see Chapter 5, pp. 144-145). The result is a long wait time for treatment.
Figure 16.8 shows how this system works. The $D=M B, M S B$, and $S=M S C$ curves are the same as in Fig. 16.7. The quantity supplied is fixed at a level that exceeds the market equilibrium but is less than the efficient quantity, so a deadweight loss arises.

Defenders of this system say that inefficiencies are small and worth accepting because the outcome is fair as everyone has equal access to services. But it isn't exactly true that everyone
has equal access. Some people are better at playing the system than others and are able to jump the line.

Private and Government Insurance In the United States, most healthcare services are produced by private doctors and hospitals that receive their in-comes from three sources: private health insurance, governments, and patients. Private health insurancepays 41 percent of the bills, government Medicare,Medicaid, and other programs pay 46 percent, and the remaining 13 percent comes from patients.

FIGURE 16.8 Public Production with Waiting


The government produces a fixed quantity of healthcare that exceeds the market quantity at the intersection of the $D=M B$ and $S=M S C$ curves but is less than the efficient quantity at the intersection of the MSB and $S=$ MSC curves. Patients face a zero price, so the quantity of healthcare demanded exceeds the quantity supplied. The available healthcare resources are allocated by doctors and patients wait for their treatment.
The area of the gray triangle shows the smallest possible amount of deadweight loss, which would be the loss if patients were treated in the order of their willingness to pay. But some people who are waitingin a long line are willing to pay more than the cost of the attention their health problem needs-their marginal benefit exceeds the marginal cost. Others get service even if their marginal benefit is less than the marginal cost. So the full deadweight loss exceeds the area of the gray triangle.

Patients' out-of-pocket payments arise because some people are uninsured, and those who are insured face deductibles in their private insurance policies or co-payments for services provided under Medicare andMedicaid.
The scale of out-of-pocket costs combined with conditions set out in insurance plans and the demand for healthcare service determine the quantity of healthcare service provided. Measured by visits to a physician's office, that quantity is 1 billion patients per year. (The average number of physician visits is 3 per person per year.)

Inefficient Overproduction? Wedon't know whether private insurance together with Medicare and Medicaid provides the efficient quantity of healthcare. But relative to what is spent in other richcountries, the scale of expenditure on these programssuggests that they do overprovide. That is what we will assume in examining how the market works.
Figure 16.9 shows the market for healthcare services. Again, the $D=M B, M S B$, and $S=M S C$ curves are the same as those in Fig. 16.7.
The quantity provided is the quantity demanded by patients and supplied by doctors at the out-of-pocket cost to patients.
Doctors and hospitals negotiate fees with the government that equal marginal cost, which also equals marginal social cost. Marginal social cost of the quantity provided exceeds marginal benefit, so the over- provision creates the deadweight loss shown by the area of the gray triangle in Fig. 16.9(a).

Uncontrolled Expenditure Figure 16.9(b) illustrates the sources of expenditure: patients' out-of-pocket expenditure, expenditure by private insurers, and government expenditure. Government expenditure is determined by the quantity of care demanded, not by a fixed budget. And without changes in the Medicare and Medicaid programs, this expenditure will grow as the aged population grows. At Issue below suggests a solution to this problem.

Obamacare The Patient Protection and Affordable Care Act, 2010 (Obamacare) has created a Health Insurance Marketplace to provide subsidized insurance.
On the supply side of the new marketplace are private insurers. On the demand side are the uninsured and those who want to find a better plan than their current one.
To qualify for subsidies, plans offered through the marketplace must cover pre-existing conditions, preventive services, and 10 essential health benefits.
The premium paid depends on family size and in- come. An example: A couple both aged 31 with two children and earning the median family income of $\$ 51,371$ receive a subsidy (as a tax credit) of $\$ 5,534$ leaving them to pay $\$ 3,562$ for a $\$ 9,096$ policy.
Families with no health insurance in 2014 are required to pay a monthly fee of $\$ 95$ per adult and $\$ 47.50$ per child or 1 percent of income, whichever is higher, with a maximum family fee of $\$ 285$ (\$3,420 a year). Figure 16.10 illustrates how the Obamacare subsidy works. The demand and marginal benefit curve $D=M B$ and the supply and marginal social cost curve $S=$ MSC determine the market quantity of 2 million insured at a price of $\$ 6,000$ per family. A subsidy of $\$ 5,000$ puts a gap between the price paid of $\$ 4,000$ and the price received by the insurer, $\$ 9,000$. The government pays the subsidy and the total payment, the rectangle, is the subsidy per family multiplied by the number of families receiving the
subsidy.The numbers in Fig. 16.10 are calibrated approximately to the actual premiums and subsidies in Obamacare. The median household receives a subsidyof about $\$ 5,000$, and (in 2014) 8 million families signed up through the Obamacare marketplace. The total subsidy paid by the government is around $\$ 40$ billion ( $\$ 5,000 \times 8$ million $=\$ 40$ billion).
To determine whether the outcome of the Obamacare subsidy achieves an efficient outcome, we would need to have an estimate of the extent to which the marginal social benefit of health insurance for the affected families exceeds their ability and willingness to pay. The subsidy might be too large, too small, or just right.

## Vouchers a Better Solution?

The power of the market leads economists to seek a market solution to all resource allocation problems, and healthcare is no exception.
When a market failure arises because marginal social benefit exceeds the ability and willingness topay, economists say that vouchers should be used. A voucher is a token that can be used to buy only the item that the voucher specifies. So a healthcare voucher could be used only to buy health insurance.
Laurence Kodikoff (see At Issue on the previous page) suggests that every American be given a health care voucher, the value of which is tied to each individual's health profile. The voucher program would replace Medicare, Medicaid, Obamacare, and some other smaller government programs. The total value of vouchers would be limited, and the government, not the consumer of healthcare services, would be in control of the healthcare budget. The markets for health insurance and healthcare services would be free to work like any other competitive market to seek an efficient outcome.

FIGURE 16.10 Subsidized Health Insurance


In the market for health insurance, the demand and marginal private benefit curve, $D=M B$, and the supply and marginal social cost curve, $S=$ MSC, would deliver an equilibrium with too many people uninsured. A subsidy puts a gap between what the insured pays and the insurer receives and increases the number of families covered. The outcome is efficient if the subsidy brings marginal social benefit into equality with marginal social cost.

## ECONOMIC ANALYSIS

- The road transportation infrastructure of the United States consists of 4 million miles of roads, 47,000 miles of interstate highways, and 607,000 bridges.
- As the news article explains, the gas tax is the main source of funds for maintaining this vast stock of transportation capital, and declining gas use is decreasing these funds.
- The result is too little expenditure on maintaining the infrastructure.
- Bridge collapse is the most visible and worrying consequence of inadequate maintenance.
- The American Society of Civil Engineers says that one in nine (that's 67,487 ) bridges are structurally deficient.
- A spending increase of $\$ 8$ billion annually is needed to bring the nation's bridges to a safe standard by 2028.
- You can explain the problem of bridge maintenance by using the tools you've learned in this chapter.
- In Fig. 1, the $x$-axis measures the number of bridges repaired per year and the $y$-axis measures the marginal benefit and cost of repairing a bridge.
- The MSC curve shows the marginal social cost of repairing a bridge and the $M S B$ curve shows the marginal social benefit.
- The efficient use of resources occurs when 6,000 bridges per year are repaired at a cost of $\$ 3$ million per bridge, with a total expenditure of $\$ 18$ billion per year.
- Restricted funds block this efficient outcome. The actual number of bridges repaired is 4,000 per year at a cost of $\$ 2.5$ million per bridge, with a total expenditure of $\$ 10$ billion per year. (The numbers are in the region of the actual U.S. data but are assumptions.)
- Because the number of bridges repaired per year is less than the efficient quantity, there is a deadweight loss.
- If Fig. 1 is a correct description of the problem, a political party can propose a bridge repair and tax program that achieves an efficient outcome.
- The problem is that taxes are collected now and bridges are repaired later, so the political party must be able to credibly commit to doing the work after it has collected the funds.


Figure 1 Underprovision of Infrastructure


The collapsed freeway bridge over the Mississippi in Minneapolis in August 2007.

## Chapter 17: Externalities

After studying this chapter, you will be able to:

- Explain how externalities arise
- Explain why external costs bring market failure and overproduction and how property rights and public choices might achieve an efficient outcome
- Explain the tragedy of the commons and its possible solutions
- Explain why external benefits bring market failure and underproduction and public choices might achieve an efficient outcome

How can we use less coal to generate electricity and reduce carbon emissions that bring climate change? What can we do to conserve the ocean's fish stocks and save them from extinction? How can we ensure that we spend enough on our schools and colleges These are the questions we study in this chapter. They arise because some of our choices impose costs on or bring benefits to others that we don't think about when we make those choices. In Economics in the News at the end of the chapter we look at the Obama administration's ideas on how to counter carbon emissions and climate change.

## Externalities in Our Lives

An externality is a cost of or a benefit from an action that falls on someone other than the person orfirm choosing the action. We call an externality that imposes a cost a negative externality; and we call an
externality that provides a benefit a positive externality.
We identify externalities as four types:

- Negative production externalities
- Positive production externalities
- Negative consumption externalities
- Positive consumption externalities


## Negative Production Externalities

Burning coal to generate electricity emits carbon dioxide that is warming the planet. Logging and the clearing of forests is destroying the habitat of wildlifeand also adding carbon dioxide to the atmosphere. These activities are negative production externalities, the costs of which are borne by everyone, and even by future generations.
Noise is another negative production externality.
When the U.S. Open tennis tournament is beingplayed at Flushing Meadows, players, spectators, and television viewers around the world share a cost that New Yorkers experience every day: the noise of airplanes taking off from LaGuardia Airport. Aircraft noise imposes a cost on millions of people who live under the flight paths to airports in every major city.

## Positive Production Externalities

To produce orange blossom honey, Honey Run Honey of Chico, California, locates beehives next toan orange orchard. The honeybees collect pollen andnectar from the orange blossoms to make the honey.At the same time, they transfer pollen between the blossoms, which helps to fertilize the blossoms. Two positive production externalities are present in this example. Honey Run Honey gets a positive production externality from the owner of the orange orchard; and the orange grower gets a positive production externality from Honey Run.

## Negative Consumption Externalities

Negative consumption externalities are a source of irritation for most of us. Smoking tobacco in a confined space creates fumes that many people find unpleasant and that pose a health risk. So smoking in restaurants and on airplanes generates a negative externality. To avoid this negative externality, many restaurants and all airlines ban smoking. But while a smoking ban avoids a negative consumption externality for most people, it imposes a negative external cost on smokers who would prefer to enjoy the consumption of tobacco while dining or taking a plane.

Noisy parties and outdoor rock concerts are other examples of negative consumption externalities. They;:are also examples of the fact that a simple ban on an activity is not a solution. Banning noisy parties avoids.the external cost on sleep-seeking neighbors, but it results in the sleepers imposing an external cost on the fun-seeking partygoers. Permitting dandelions to grow in lawns, not picking up leaves in the fall, allowing a dog to bark loud or to foul a neighbor's lawn, and letting a cellphone i'ring in class are other examples of negative consumption externalities.

## Positive Consumption Externalities

When you get a flu vaccination, you lower your risk; of being infected. If you avoid the flu, your neighbor, who didn't get vaccinated, has a better chance of remaining healthy. Flu vaccinations generate positive consumption externalities.
When the owner of a historic building restores it, everyone who sees the building gets pleasure from it.
Similarly, when someone erects a spectacular home such as those built by Frank Lloyd Wright during the 1920s and 1930s-or other exciting building- such as the Chrysler and Empire State Buildings in New York or the Walt Disney Concert Hall in Los Angeles - an external consumption benefit flows to everyone who has an opportunity to view it. Education, which we examine in more detail in this chapter, is a major example of this type of externality.

## ECONOMICS IN ACTION

## Opposing Trends: Success and failure

The trends in local U.S. air quality and global green- house gas concentrations are starkly opposing. The concentrations of air pollutants in U.S. cities is decreasing, as it has done so
for the past 32 years. In contrast, the concentration of greenhouse gases (mainly carbon dioxide) in the global atmosphere is increasing and posing an ever more urgent problem. Air Pollution Trends Figure 1 shows the trends inthe concentrations of the six main pollutants of the air in the United States between 1980 and 2012. The concentrations of all these pollutants decreased.
The Clean Air Act has brought regulations that almost eliminated lead and cut emissions of carbon monoxide, sulfur dioxide, and nitrogen dioxide to below 40 percent and ozone to 80 percent of their 1980 levels, and particulate matter to 64 percent of its 2000 level. And economic actions that you will learn about in this chapter have almost eliminated lead from highway vehicles and industrial processes.
These reductions in air pollution are even more impressive seen against the trends in economic activity. Between 1980 and 2012, total production in the United States increased by 145 percent, vehicle miles traveled increased by 95 percent, and the population increased by 40 percent.

Global C02 and Temperature Trends Figure 2 show the global trends in carbon dioxide ( CO 2 ) concentration and temperature. Both trends are starkly upward. CO 2 concentration has increased by almost 40 percent since 1850, and global temperature has been rising for more than 100 years. Scientists agree that the scale on which we burn fossil fuels is the major source of the rising $\mathrm{CO}_{2}$ trend. There is more uncertainty about the effect of the increase in CO2 on global temperature, but the consensus is that the effect is significant. Stopping the rising $\mathrm{CO}_{2}$ trend requires joint action by the governments of every nation. But a binding agreement among nations to reduce green- house gas emissions, the Kyoto Protocol, excluded the major developing countries and the United States refused to ratify it. You will see in this chapter whyglobal warming is a much harder problem to solve than reducing air pollution in the United States.


Figure 1 U.S. Air Pollution Trends


Figure 2 Global Warming Trends
Sources of data: Temperature: Met Office Hadley Centre |combined land and oceans); $\mathrm{CO}_{2}$ : Scripps Institution of Oceanography, Mauna Loa Observatory, Hawaii, data since 1960 and ice-core estimates before 1960.

## Negative Externality: Pollution

To see the effects and possible remedies for a negative production externality, we'll look at the example of production activities that pollute. We begin by distinguishing among three costs: private, external, and social.

## Private, External, and Social Cost

A private cost of production is a cost that is borne by the producer of a good or service. Marginal cost is the cost of producing an additional unit of a good or service. So marginal private cost ( $M C$ ) is the cost of producing an additional unit of a good or service that is borne by its producer.
An external cost is a cost of producing a good or service that is not borne by the producer but borne by other people. A marginal external cost is the cost of producing an additional unit of a good or service that falls on people other than the producer.
Marginal social cost (MSC) is the marginal cost incurred by the producer and by everyone else on whom the cost falls by society. It is the sum of marginal private cost and marginal external cost. That is,

$$
M S C=M C+\text { Marginal external cost. }
$$

We express costs in dollars, but we mustalways remember that a cost is an opportunity cost-some-thing real, such as dean air or a dean river, is given up to get something.

Valuing External Cost Economists use market prices to put a dollar value on the external cost of pollution. For example, suppose that there are two similar rivers, one polluted and the other dean. Ten identical homes are built along the side of each river. The homes on the
dean river rent for $\$ 2,000$ per month, and those on the polluted river rent for $\$ 1,500$ per month. If the pollution is the only detectable difference between the two rivers and the two locations, the rent difference of $\$ 500$ per monthis the pollution cost per home. With 10 homes onthe side of a polluted river, the external cost of pollution is $\$ 5,000$ per month.

External Cost and Output Figure 17.1 shows an example of the relationship between output and cost in a paint industry that pollutes rivers. The marginal cost curve, MC, describes the marginal private cost borne by the paint producers, which increases as the quantity of paint produced increases.
figure 17.1 An External Cost


The MC curve shows the marginal private cost borne by the factories that produce paint. The MSC curve shows the sum of marginal private cost and marginal external cost. When output is 3 million gallons of paint per month, marginal private cost is $\$ 1.00$ per gallon, marginal external cost is $75 \$$ per gallon, and marginal social cost is $\$ 1.75$ per gallon.

If a firm pollutes a river, it imposes an external cost borne by other users of the river. Pollution and its marginal external cost increase with the amount of paint produced.
The marginal social cost curve, MSC, is found by adding marginal external cost to marginal private cost. So a point on the MSC curve shows the sum ofthe marginal private cost and marginal external costat a given level of output.
For example, if 3 million gallons of paint per month are produced, marginal private cost is $\$ 1.00$ per gallon, marginal external cost is $75 c t$ per gallon, and marginal social cost is $\$ 1.75$ per gallon.
Let's now see how much paint gets produced and how much pollution gets created.

Equilibrium and Amount of Pollution Equilibrium in the market for paint determines the amount of pollution. Figure 17.2 has the same MC and MSC curves as Fig. 17.1 and also has a market demand and marginal social benefit curve, $D=M S B$. Equilibrium $\cdot$ occurs at a price of $\$ 1.00$ per gallon and 3 million gallons per month. This equilibrium is one with inefficient overproduction (Chapter 5, p. 152) because marginal social cost at $\$ 1.75$ per gallon exceeds marginal social benefit at $\$ 1.00$ per gallon.

FIGURE 17.2 Inefficiency with an External Cost


The factories' marginal private cost curve is the market supply curve, $S=M C$. The market demand curve is the marginal social benefit curve, $D=$ MSB. The market equilibrium occurs at a price of $\$ 1.00$ per gallon and 3 million gallons per month. This outcome is inefficient because marginal social cost exceeds marginal social benefit. The efficient quantity of paint is 2 million gallons per month. The gray triangle shows the deadweight loss created by the pollution.
The efficient equilibrium occurs where marginal social benefit equals marginal social cost at 2 million gallons of paint per month. Too much paint is produced, too much pollution is created, and the area of the deadweight loss triangle measures the society's loss. The deadweight loss arises because the paint factories only take their private cost into account when making their production decision. If some method can be found to get paint factories to create less pollution and eliminate the deadweight loss, everyone - the owners of paint factories and the residents of the riverside homes - can gain. So, what can be done to fix the inefficiency that arises from an external cost? Three approaches are available and we will examine each of them. They are

- Establish property rights
- Mandate clean technology
- Tax or price pollution


## Establish Property Rights

Propertyrights arelegallyestablished titles to the owner- ship, use, and disposal of factors of production and goods and services that are enforceable in the courts. Property rights are a foundation stone of the market economy. But they don't apply to all property.
Establishing property rights can confront producers with the costs of their actions and provide the incentives that allocate resources efficiently.
To see how property rights work, suppose that the paint producers have property rights on a river and thehomes alongside it-they own the river and the homes. The rental income that the paint producers are able tomake on the homes depends on the amount of pollution they create. Using the earlier example, people are willing to pay a rent of $\$ 2,000$ a month to live alongside a pollution-free river but only $\$ 1,500$ a month to live with the pollution created by producing 3 million gallons of paint per month.
The forgone rental income from homes alongside a polluted river is an opportunity cost of producing paint. The paint producers must now decide how to respond tothis cost. There are two possible responses:

- Use an abatement technology
- Produce less and pollute less

Use an Abatement Technology Anabatementtechnology is a production technology that reduces or prevents pollution. The catalytic converter in every U.S. car is an example of an abatement technology. Its widespreadadoption (with lead-free gasoline) has dramaticallyreduced pollution from highway vehicles and helped to achieve the trends in U.S. air quality shown on p . 431.

Abatement technologies exist to eliminate or reduce pollution from electricity generation and many industrial processes, including the manufacture of paint.

Produce Less and Pollute Less An alternative to incurring the cost of using an abatement technology isto use the polluting technology but cut production, re-duce pollution, and get a higher income from renting homes by the river. The decision turns on cost: Firmswill choose the least-cost alternative.

Efficient Market Equilibrium Figure 17.3illustrates the efficient market outcome. With property rights inplace, the paint producers face the pollution costs orthe abatement costs, whichever is lower. The MSC curve includes the cost of producing paint plus either the cost of abatement or the cost of pollution (forgone rent), whichever is lower. This curve, labeled $S=$ $M C=M S C$, is now the market supply curve.
Market equilibrium occurs at a price of $\$ 1.25$ per gallon and 2 million gallons of paint per month. This outcome is efficient.
If the forgone rent is less than the abatement cost, the factories will still create some pollution, but it will be the efficient quantity. If the abatement cost is lower than the
forgone rent, the factories will stop polluting. But they will produce the efficient quantity because marginal cost includes the abatement cost.

FIGURE 17.3 Property Rights Achieve an Efficient Outcome


With property rights, the marginal cost curve that excludes pollution and abatement costs shows only part of the producers' marginal cost. The marginal cost of producing paint now includes the cost of pollution-the external cost-or the cost of abatement. So the market supply curve is $S=$ MC $=$ MSC. Market equilibrium occurs at a price of $\$ 1.25$ per gallon and 2 million gallons of paint per month. Marginal social cost equals marginal social benefit, so the outcome is efficient.

The Coase Theorem Does it matter whether the polluter or the victim of the pollution owns the resource that might be polluted? Until 1960, everyone thought that it did matter. But in 1960, Ronald Coase (see p. 455) had a remarkable insight that wenow call the Coase theorem. The Coase theorem is the proposition that if property rights exist and the transactions costs of enforcing them are low, then private transactions are efficient and it doesn't matter who has the property rights.

Application of the Coase Theorem Suppose that instead of the paint factories owning the homes, the residents own their homes and the river. Now the factories must pay a fee to the homeowners for the right 'to dump their waste. The greater the quantity of waste dumped into the river; the more the factories must pay. So again, the factories face the opportunity cost of the pollution they create. The quantity of paint produced and the
amount of waste dumped are the same whoever owns the homes and the river. If the factories own them, they bear the cost of pollution because they receive a lower income from home rents. If the residents own the homes and the river, the factories bear the cost of pollution because they must pay a fee to the homeowners. In both cases, the factories bear the cost of their pollution and dump the efficient amount of waste into the river.
The Coase solution works only when transactions costs are low. Transactions costs are the opportunity costs of conducting a transaction. For example, when you buy a house, you pay an agent to help you find the best place and a lawyer to run checks that assure you that the seller owns the property and that after you've paid for it, the ownership has been properly transferred to you. These costs are transactions costs.
Inthe example of the homes alongside a river, the transactions costs that are incurred by a small number of paint factories and a few homeowners might be low enough to enable them to negotiate the deals that produce an efficient outcome. But in many situations, transactions costs are so high that it would be inefficient to incur them. In these situations, the Coase solution is not available.

## Mandate Clean Technology

When property rights are too difficult to define and enforce, public choices are made. Regulation is a government's most likely response. Most countries regulate what may be dumped in rivers and lakes and emitted into the atmosphere. The environmental resources of the United States are heavily regulated.
An example of environmental regulation is the Clean Air Act of 1970 and its later amendments, which give the Environmental Protection Agency (EPA) the authority to issue regulations that limit emissions and achieve defined air quality standards.
The EPA has issued thousands of regulations that require chemical plants, utilities, and steel mills to adopt best-practice pollution abatement technologies and limit their emissions of specified air pollutants. Other regulations have been issued that govern road vehicle emission limits, which must be met by the vehicle manufacturers.
In 2007, the Supreme Court ruled that the E P A has authority to regulate greenhouse gas emissions.
Although direct regulation can and has reduced emissions and improved air quality, economists are generally skeptical about this approach. Abatement is not always the leastcost solution. Also, government agencies are not well placed to find the cost-minimizing solution to a pollution problem. Individual firms seeking to minimize cost and maximize profit and responding to price signals are more likely to achieve an efficient outcome. We'll now examine these other approaches to pollution.

## Tax or Cap and Price Pollution

Governments use two main methods of confronting polluters with the costs of their decisions:

- Taxes
- Cap-and-trade

Taxes Governments can use taxes as an incentive for producers to cut back the pollution they create. Taxes used in this way are called Pigovian taxes, in honor of Arthur Cecil Pigou, the British economist who first worked out this method of dealing with externalcosts during the 1920s.
By setting the tax equal to the marginal external cost (or marginal abatement cost if it is lower), firms can be made to behave in the same way as they would if they bore the cost of the externality directly.
To see how government actions can change the outcome in a market with external costs, let's return to the example of paint factories and the river.
Assume that the government has assessed the marginal external cost of pollution accurately and imposes a tax on the factories that exactly equals this cost. The producers are now confronted with the social cost of theiractions. The market equilibrium is one in which price equals marginal social cost-an efficient outcome.
Figure 17.4 illustrates the effects of a Pigovian tax on paint factory pollution. The curve $D=$ $M S B$ is the market demand and the marginal social benefit curve. The curve MC is the marginal cost curve. The tax equals the marginal external cost of the pollution. We add this tax to the marginal private cost to find the market supply curve, the curve labeled $S=M C+$ tax=MSC. This curve is the market supply curve because it tells us the quantity supplied at each price, given the factories' marginal cost and the tax they must pay. This curve is also the marginal social cost curve.because the pollution tax has been set equal to the marginal externalcost.
Demand and supply now determine the market equilibrium price at $\$ 1.25$ per gallon and a quantity of 2 million gallons of paint a month. At this quantity of paint production, the marginal social cost is
$\$ 1.25$ and the marginal social benefit is $\$ 1.25$, so the market outcome is efficient. The factories incur a marginal private cost of $75 ¢$ per gallon and pay a pollution tax of $50 ¢$ per gallon. The government collects tax revenue of \$1 million per month.

Cap-and-Trade Acap is an upper limit. You met the idea of a government imposing an upper limit when you learned about production quotas (Chapter 6, p. 177) and import quotas (Chapter 7, p 198). A cap is a quota - a pollution quota.
A government that uses this method must first estimate the efficient quantity of pollution and set the overall cap at that level.
Just like a production or import quota, a pollution quota or cap must somehow be allocated to individual firms (and possibly even households). In an efficient allocation of pollution quotas, each firm has the same marginal social cost. So to make an efficient allocation of the cap across firms, the government would need to know each firm's marginal production cost and marginal abatement cost.
A Pigovian tax achieves an efficient allocation of pollution across firms because each firm chooses how. much to produce and pollute taking the tax intoaccount, and produces the quantity at which marginal social cost equals price. Because all firms face the same ' market price, they also incur the same marginal social cost.

The government solves the allocation problem by making an initial distribution of the cap across firms and allowing them to trade in a market for pollution permits. Firms that have a low marginal abatement cost sell permits and make big cuts in pollution. Firms that have a high marginal abatement cost buy permits and make smaller cuts or perhaps even no cuts in pollution.
The market in permits determines the equilibrium price of pollution and each firm, confronted with that price, maximizes profit by setting its marginal pollution cost or marginal abatement cost, whichever is lower, equal to the market price of a permit By confronting polluters with a price of pollution trade in pollution permits can achieve the same efficient outcome as a Pigovian tax.

## FIGURE 17.4 A Pollution Tax to Achieve an

 Efficient Outcome

When the government imposes a pollution tax equal to themarginal external cost of pollution, the market supply .curve becomes the marginal private cost curve, $M C$, plus the tax-the curve $S=M C+$ tax. Market equilibrium occurs at a price of $\$ 1.25$ per gallon and a quantity of 2 million gallons of paint per month. This equilibrium is efficient because marginal social cost equals marginal social benefit. The gray rectangle shows the government's tax revenue.

## ECONOMICS IN ACTION

## Taxing Carbon Emissions

British Columbia, Ireland, and the United Kingdom are making their carbon footprints smaller.

British Columbia's Carbon Tax Introduced in 2008 at $\$ 10$ per metric ton of carbon emitted, British Columbia's tax increased each year to its final rate of $\$ 30$ per metric ton in 2012. The tax applies to all forms of carbon emission from coal, oil, and natural gas. The tax is revenue-neutral, which means that other taxes, personal and corporate income taxes, are cut by the amount raised by the carbon tax. Between 2008 and 2012, carbon emissions fell by 17 percent.

Ireland's Carbon Tax Since 2010, Ireland has taxed kerosene, gas oil, liquid petroleum gas, fuel oil, natural gas, and solid fuels. The tax rate in 2013 was 10 euros ( $\$ 13$ ) per metric ton of $\mathrm{CO}_{2}$ emitted and 20 euros (\$26) per ton in 2014. Emissions have fallen since the tax was introduced, but recession as well as the carbon tax brought this fall.
U.K. Tax on Gasoline The United Kingdom doesn't call its gasoline tax a carbon tax, but it has the same effect on drivers. The figure shows the U.K. price of gasoline compared with that in three other countries. The enormous differences arise almost entirely from tax differences. An effect of these price differences is that cars in the United Kingdom get an average of 38 miles per gallon while in the United States, the average is 23 miles per gallon. A high gas tax cuts carbon emissions by inducing people to drive smaller cars and to drive less.


Demand and supply now determine the market equilibrium price at $\$ 1.25$ per gallon and a quantity of 2 million gallons of paint a month. At this quantity of paint production, the marginal social cost is $\$ 1.25$ and the marginal social benefit is $\$ 1.25$, so the market outcome is efficient. The factories incur a marginal private cost of $75 \$$ per gallon and pay a pollution tax of $50 \$$ per gallon. The government collects tax revenue of $\$ 1$ million per month.

Cap-and-Trade A cap is an upper limit. You met the idea of a government imposing an upper limit when you learned about production quotas (Chapter 6, p. 177) and import quotas (Chapter 7, p. 198). A cap is a quota-a pollution quota.

A government that uses this method must first estimate the efficient quantity of pollution and set the overall cap at that level.

Just like a production or import quota, a pollution quota or cap must somehow be allocated to individual firms (and possibly even households). In an efficient allocation of pollution quotas, each firm has the same marginal social cost. So to make an efficient allocation of the cap across firms, the government would need to know each firm's marginal production cost and marginal abatement cost.

A Pigovian tax achieves an efficient allocation of pollution across firms because each firm chooses how. much to produce and pollute taking the tax into account, and produces the quantity at which marginal social cost equals price. Because all firms face the same market price, they also incur the same marginal social cost.

The government solves the allocation problem by making an initial distribution of the cap across firms and allowing them to trade in a market for pollution permits. Firms that have a low marginal abatement cost sell permits and make big cuts in pollution. Firms that have a high marginal abatement cost buy permits and make smaller cuts or perhaps even no cuts in pollution.

The market in permits determines the equilibrium price of pollution and each firm, confronted with that price, maximizes profit by setting its marginal pollution cost or marginal abatement cost, whichever is lower, equal to the market price of a permit.

By confronting polluters with a price of pollution, trade in pollution permits can achieve the same efficient outcome as a Pigovian tax.

## Coping with Global Externalities

The United States has cut the emissions of local air pollutants and made its own air cleaner by adopting the measures you've just seen. But one country, even a big country such as the United States that accounts for 20 percent of the value of global production and 14 percent of carbon emissions, can't solve the problem of global warming and climate change alone.
Coping with this problem requires public choices at a globallevel, choices by all governments, which are much harder to make and coordinate.
A lower $\mathrm{CO}_{2}$ concentration in the world's atmosphere is a global public good. And like all public goods, it brings a free-riderproblem (see Chapter 16, Pg. 412). Without a mechanism to ensure participation in a global carbon reduction program, countries are in a prisoners 'dilemma (see Economics in Action above).

## Negative Externality: The Tragedy of the Commons

Overgrazing the pastures around a village in Middle Ages England, and overfishing the Southern Bluefin tuna, Pacific Yellowfin tuna, Atlantic cod, and Minke whales during the recent past are tragedies of the commons. Other current tragedies are the destruction of tropical rainforests in Africa and the Amazon basin of SouthAmerica.
The tragedy of the commons is the overuse of a common resource that arises when its users have no incentive to conserve it and use it sustainably and efficiently.
To study the tragedy of the commons and its possible remedies, we'll focus on one of the recent and current tragedies - overfishing and depleting the stock of tuna. You're about to discover that there are two problems that give rise to the tragedy of the commons:

- Unsustainable use of a common resource
- Inefficient use of a common resource

Unsustainable Use of a Common Resource Many common resources are renewable-they replenish themselves by the birth and growth of new members of the population. Fish, trees, and the fertile soil is all examples of this type of resource. At anygiven time, there is a stock of the resource and a rate at which it is being used.
A common resource is being used unsustainably if its rate of use persistently decreases its stock. A common resource is being used sustainably if its rate of use is less than or equal to its rate of renewal so that the stock available either grows or remains constant. Focusing on the example of fish, a species is being used unsustainably if the catch persistently decreases the stock and it is being used sustainably if the catch is less than or equal to the rate of renewal of the fish population.
The sustainable catch depends on the stock of fish and in the way illustrated by the sustainable catch curve, SCC, in Fig. 17.5.
Along the SCC curve, with a small stock of fish the quantity of new fish born is also small, so the sustainable catch is small.

With a large stock of fish many fish are born but they must compete with each other for food, so only a few survive to reproduce and to grow large enough to catch, and again the sustainable catch is small.

FIGURE 17.5 Sustainable Catch


As the fish stock increases (on the $x$-axis), the sustainable catch (on the $y$-axis) increases to a maximum. As the stock increases further, the fish must compete for food and the sustainable catch falls. If the catch exceeds the sustainable catch, such as ot point $A$, the fish stock diminishes. If the catch is less than the sustainable catch, such as at point $Z$, the fish stock increases.

Between a small and a large stock is a stock of fish that maximizes the sustainable catch. In Fig. 17.5, this stock is 20 million tons and the maximum sustainable catch is 300,000 tons per year.
The maximum sustainable catch arises from a balancing of the birth of new fish from the stock and the availability of food to sustain the fish population. If the quantity of fish caught is less than the sustainable catch, at a point such as $Z$, the fish stock grows. If the quantity caught equals the sustainable catch, at any point on the sec, the fish stock remains constant and is available for future generations of fishers in the same quantity that is available today.
But if the quantity caught exceeds the sustainable catch, at a point such as $A$, the fish stock shrinks and, unchecked, eventually falls to zero.
You now understand the problem of using a common renewable natural resource sustainably. But another problem is using it efficiently.

## Inefficient Use of a Common Resource

In an unregulated market, even if the catch is sustain- able, it will be bigger than the efficient catch: over- fishing occurs. And most likely, the catch will not be sustainable. Why does overfishing occur? The answer is that fishers face only their own private cost and don't face the cost they impose on others-the external cost. The social cost of fishing combines the private cost and the external cost. Let's examine these costs.

Marginal Private Cost The marginal private cost of catching fl.sh is the cost incurred by keeping a boat and crew at sea for long enough to increase the catch by one ton. Keeping a fishing boat at sea eventually runsinto diminishing marginal returns (see Chapter 11,
p. 289). As the crew gets tired, the storage facilities get overfull, and the boat's speed is cut to conserve fuel,the catch per hour decreases. The cost of keeping theboat at sea for an additional hour is constant, so the marginal cost of catching fl.sh increases as the quantity caught increases.
You've just seen that the principle of increasing marginal cost applies to catching fl.sh just as it applies to other production activities: Marginal private cost increases as the quantity of fish caught increases.
The marginal private cost of catching fish determines an individual fisher's supply of fish. A profl.t- maximizing fisher is willing to supply the quantity atwhich the market price of fl.sh covers the marginalprivate cost. And the market supply is the sum of thequantities supplied by each individual fisher.

Marginal External Cost The marginal external cost of catching fl.sh is the cost per additional ton that one fisher's production imposes on all other fishers. This additional cost arises because one fisher's catch decreases the remaining stock, which in turn decreases the renewal rate of the stock and makes it harder for others to find and catch fish. Marginal external cost also increases as the quantity of fish caught increases. If the quantity of fl.sh caught is so large that it drives the species to near extinction, the marginal external cost becomes infinitelylarge.

Marginal Social Cost The marginal social cost of catching fish is the marginal private cost plus the marginal external cost. Because both components of marginal social cost increase as the quantity caught increases, marginal social cost also increases with the quantity of fish caught.

MarginalSocial Benefitand Demand Themarginal social benefit from fl.sh is the price that consumers are willing to pay for an additional pound of fl.sh. Marginal social benefit decreases as the quantity of fl.sh consumed ' increases, so the market demand curve, which is also the marginal social benefit curve, slopes downward.

Overfishing Equilibrium Figure 17.6illustrates over- fishing and how it arises. The market demand curve for fish is the marginal social benefit curve, MSB. The market supply curve is the marginal private cost curve, MC Market equilibrium occurs at the inter- section point of these two curves. The equilibrium quantity is 800,000 tons of fl.sh per year and the equilibrium price is $\$ 10$ per pound. At this market equilibrium, overfishing occurs.

FIGURE 17.6 Why Overfishing Occurs


The market supply curve is the marginal private cost curve, MC. The market demand curve is the marginal social bene fit curve, MSB. Market equilibrium occurs at a quantity of 800,000 tons per year and a price of $\$ 10$ per pound.

The marginal social cost curve is MSC and at the market equilibrium there is overfishing-marginal social cost exceeds marginal social benefit.

The quantity at which MSC equals MSB is the efficient quantity, 300,000 tons per year. The area of the gray triangle measures the deadweight loss from overfishing.
Figure 17.6 illustrates why overfishing occurs. At the market equilibrium quantity, marginal social benefit (and willingness to pay) is $\$ 10$ per pound, but the marginal social cost ( $\$ 37$ per pound) exceeds this amount. The marginal external cost is the cost of running down the fish stock.

Efficient Equilibrium What is the efficient use of a common resource? It is the use of the resource thatmakes the marginal social benefit from the resource equal to the marginal social cost of using it.
In Fig. 17.6, the efficient quantity of fish is 300,000 tons per year-the quantity that makes marginal social cost (on the MSC curve) equal to marginal social benefit (on the MSB curve). At this quantity, the marginal catch of each individual fisher costs society what people are willing to pay for it.

Deadweight Loss from Overfishing Deadweight loss measures the cost of overfishing. The area of thegray triangle in Fig. 17.6 illustrates this loss. It is themarginal social cost minus the marginal social benefit from all the fish caught in excess of the efficientquantity.

## ECONOMICS IN ACTION

## The Original Tragedy of the Commons

The term "tragedy of the commons" comes from 14thcentury England, where areas of rough grassland surrounding villages were overgrazed and the quantity of cows and sheep that they could feed kept falling.

During the 16th century, the price of wool increased, sheep farming became profitable, and sheep owners wanted to control the land they used. So the commons were gradually privatized and land use became more efficient.


## One of Today's Tragedies of the Commons

Before 1970, Atlantic cod was abundant, despite the fact that it had been fished for many centuries and was a major food source for the first European settlers in North America. By 1620, there were more than 1,000 fishing boats catching large quantities of cod in the northwest Atlantic off the coast of what is now New England and Newfoundland. Most of the fishing during these years was done using lines and productivity was
low. But low productivity limited the catch and enabled cod to be caught sustainably over hundreds of years.

The situation changed dramatically during the 1960 s with the introduction of huge nets, sonar technology to find fish concentrations, and large ships with efficient processing and storage facilities. These technological advances brought soaring cod harvests and cod landings almost tripled in a decade.

This volume of cod could not be taken without a serious collapse in the remaining stock and by the 1980 s it became vital to regulate cod fishing. But regulation was of limited success and stocks continued to decline. In 1992, a total ban on cod fishing in the North Atlantic stabilized the population but at a very low level. Two decades of ban have enabled the species to repopulate, and it is now hoped that one day cod fishing will return but at a low and sustainable rate.


## Achieving an Efficient Outcome

Defining the conditions under which a common resource is used efficiently is easier than delivering those conditions. To use a common resource efficiently, it is necessary to design an incentive mechanism that confronts the users of the resource with themarginal social consequences of their actions. Thesame principles apply to common resources as those that you met earlier in this chapter when you studiedthe external cost of pollution.
The three main methods that might be used to achieve the efficient use of a common resource are:

- Property rights
- Production quotas
- Individual transferable quotas (ITQs)

Property Rights A common resource that no one owns and that anyone is free to use contrasts with private property, which is a resource that someone owns and has an incentive to
use in the way that maximizes its value. One way of overcoming the tragedy of the commons is to convert a common resource to private property. By assigning private property rights to what was previously a common resource, its owner faces the same conditions as society faces. It doesn't matter who owns the resource.
The users of the resource will be confronted with the full cost of using it because they either own it or pay a fee to the owner for permission to use it.
When private property rights over a resource are established and enforced, the MSC curve becomes the marginal private cost curve, and the use of theresource is efficient.
Figure 17;3, which illustrates an efficientoutcome with property rights when confronted with pollution, also applies to the over-fishing problem. The supply curve $S=M C=M S C$ and the demand curve $D=M S B$ determine the equilibrium price and quantity. The price equals both marginal social benefit and marginal social cost and the fish catch is efficient. The private property solution to the tragedy of the commons is available in some cases. It was the solution to the original tragedy of the commons in England's Middle Ages. It is also a solution that has been used to prevent overuse of the airwaves that carry cellphone services. The right to use this space (called the frequency spectrum) has been auctioned by governments to the highest bidders. The owner ofeach part of the spectrum is the only one permitted to use it (or to license someone else to use it).
But assigning private property rights is not always feasible. It would be difficult to assign private property rights to the oceans because the cost ofenforcing them would be too high. In the absence of property rights, some form of government intervention is used. One such intervention is a production quota.

Production Quotas Aproduction quota is an upper limit to the quantity of a good that may be produced in a specified period. Each individual producer is allocated a quota.
You studied the effects of a production quota in Chapter 6 (pp. 177-178) and learned that a quota candrive a wedge between marginal social benefit and marginal social cost and create deadweight loss. In that earlier example, the market was efficient without a quota. But in the case of common resources, the market overuses the resource and produces an inefficient quantity. A production quota in this market brings a move toward a more efficient outcome. But implementing a production quota has two problems.
First, it is in every fisher's self-interest to catch more fish than the quantity permitted under the quota. The reason is that price exceeds marginal private cost, so by catching more fish, a fisher gets ahigher income. If enough fishers break the quota, overfishing and the tragedy of the commons remain.
Second, marginal cost is not, in general, the same for all producers-as we're assuming here. Efficiency requires that the quota be allocated to the producers with the lowest marginal cost. But bureaucrats who allocate quotas do not have information about the marginal cost of individual producers. Even if they tried to get this information, producers would have an incentive to lie about their costs so as to get a bigger quota.
So where producers are difficult, or very costly, to monitor or where marginal cost varies across producers, a production quota cannot achieve an efficient outcome.

Individual Transferable Quotas Where producers are difficult to monitor or where marginal cost varies across producers, a more sophisticated quota system can be effective. Itis an individual transferable quota (ITQ), which is a production limit that is assigned to an individual who is then free to transfer (sell) the quota to someone else. A market in ITQs emerges and ITQs are traded at their market price. (Cap-and- trade to limit pollution is an ITQ for pollution).
The market price of an ITQ is the highest price that someone is willing to pay for one. That price ismarginal social benefit minus marginal cost. The price of an ITQ will rise to this level because fisherswho don't have a quota would be willing to pay thisamount to get one. A fisher with an ITQ could sell it for the market price, so by not selling the ITQ the fisher incurs an opportunity cost. The marginal cost of fishing, which now includes the opportunity cost of the ITQ, equals the marginal social benefit from the efficient quantity. Figure 17.7 illustrates how ITQs work. Each fisher receives an allocation of ITQs and the total catch per- mitted by the ITQs is 300,000 tons per year. Fishers trade ITQs:Those with low marginal cost buy ITQsfrom those with high marginal cost, and the market price of an ITQ settles at $\$ 10$ per pound of fish. The marginal private cost offishing now becomes the original marginal private cost, $M C$, plus the cost of the ITQ. The marginal private cost curve shifts upward from MC to MC +price ofITQ and each fisher is con- fronted with the marginal social cost of fishing. No one has an incentive to exceed the quota because to do so would send marginal cost above price and result in a loss on the marginal catch. The outcome is efficient.

## ECONNOMICS IN ACTION

## ITQs Work

Iceland introduced the first ITQs in 1984 to conserve its stocks of lobster. In 1986, New Zealand and a bit later Australia introduced ITQsto conserve fish stocks in the South Pacific and Southern Oceans. Andthis system was introduced in Alaska to manage halibut stocks during the 1990s.
The evidence from these cases suggests that ITQs work well when they are properly enforced.
ITQs help maintain fish stocks, but they also reduce the size of the fishing industry. This consequence of ITQs puts them against the self-interest of fishers. In all countries, the fishing industry opposes restrictions on its activities, but in Iceland, Australia, New Zealand, and Alaska the opposition is not strong enough to block ITQs.
In the United States, other than Alaska, the opposition has been harder to overcome, and in 1996 Congress passed the Sustainable Fishing Act that put a moratorium on ITQs. This moratorium was lifted. in 2004 and since then, ITQs have been applied to 28 fisheries from the Gulf of Alaska to the Gulf of Mexico.
Economists have studied the effects of ITQs extensively and agree that in most cases they work and offer an effective tool for achieving an efficient use of $t$ he stock of ocean fish. But in some situations, ITQs have been unsuccessful.
The main reasons for failure of an IT Q system are thatthe total permitted catch might be set too large; some fish species migrate and get overfished outside the regions covered by the ITQs; and sometimes the monitoring and enforcement of quotas is inadequate.

## FIGURE 17.7 ITQs to Use a Common Resource Efficiently



ITQs are issued on a scale that keeps output at the efficient level. The market price of an ITQ equals the marginal social benefit minus marginal cost. Because each user of the common resource faces the opportunity cost of using the resource, selfinterest achieves the social interest.

## Positive Externality: Knowledge

Knowledge comes from education and research. And both bring external benefits. To keep the explanation of the problems that arise from external benefits and the possible solutions to those problems as clear as possible, we will focus on just one aspect of the production of knowledge, the provision of college education.
We begin by distinguishing between private benefits and social benefits.

## Private Benefits and Social Benefits

A private benefit is a benefit that the consumer of a good or service receives. For example, expanded job opportunities and a higher income are private benefits of a college education.
Marginal benefit is the benefit from an additional unit of a good or service. So marginal private benefit $(M B)$ is the benefit that the consumer of a good orservice receives from an additional unit of it. Whenone additional student attends college, the benefit that student receives is the marginal private benefitfrom college education.
The external benefit from a good or service is the benefit that someone other than the consumer of the good or service receives. College graduates generate many external benefits. On average, they are better citizens, have lower crime rates, and are more
tolerant of the views of others. They enable the success of high-quality newspapers and television broad- casts, music, theater, and other organized social activities that bring benefits to many other people.
A marginal external benefit is the benefit from an additional unit of a good or service that people other than its consumer enjoy. The benefit that your friends and neighbors get from your college education is the marginal external benefit of your college education.
Marginal Social Benefit (MSB) is the marginal benefit enjoyed by society - by the consumer of a good or service (marginal private benefit) and by others (the marginal external benefit). That is,

$$
M S B=M B+\text { Marginal external benefit. }
$$

Figure 17.8 shows an example of the relationship between marginal private benefit, marginal external benefit, and marginal social benefit. The marginal benefit curve, $M B$, describes the marginal private benefit enjoyed by the people who receive a college education. Marginal private benefit decreases as the number of students enrolled in college increases.
figure 17.8 An External Benefit


The $M B$ curve shows the marginal private benefit enjoyed by the people who receive a college education. The MSB curve shows the sum of marginal private benefit and marginal external benefit. When 15 million students attend college, the marginal private benefit is $\$ 10,000$ per student, the marginal external benefit is $\$ 15,000$ per student, and the marginal social benefit is $\$ 25,000$ per student.
In the example in Fig. 17.8, when 15 million students enroll in college, the marginal external benefit is $\$ 15,000$ per student per year. The marginal social benefit curve, $M S B$, is the sum of marginal private benefit and marginal external benefit at each number of students. For example, when 15 million students a year enroll in college, the
marginal private benefit is $\$ 10,000$ per student and the marginal external benefit is $\$ 15,000$ per student, so the marginal social benefit is $\$ 25,000$ per student.
When people make schooling decisions, they ignore its external benefits and consider only its private benefits. So if education were provided by private schools that charged full-cost tuition, there would be too few collegegraduates.
Figure 17.9 illustrates this private underprovision. The supply curve is the marginal social cost curve, $\mathrm{S}=\mathrm{MSC}$. The demand curve is the marginal private benefit curve, $D=M B$. Market equilibrium occurs at tuition of $\$ 15,000$ per student per year and7.5 million students per year. At this equilibrium, the marginal social benefit of $\$ 38,000$ per student exceeds the marginal social cost by $\$ 23,000$ per student. Too few students are enrolled in college. The efficient number is 15 million per year, where marginal social benefit equals marginal social cost. The gray triangle shows the deadweight loss created.
To get closer to producing the efficient quantity of a good with an external benefit, we make public choices, through governments, to modify the market outcome.

## Government Actions in the Market with External Benefits

To encourage more students to enroll in college-to achieve an efficient quantity of college education-students must be confronted with a lower market price and the taxpayer must somehow pay for the costs not covered by what the student pays.

Three devices that governments can use to achieve a more efficient allocation of resources in the presence of external benefits are

- Public production
- Private subsidies
- Vouchers

Public Production With public production, a good or service. is produced by a public authority that receivesits revenue from the government. The education services produced by state universities and colleges andpublic schools are examples of public production.

Private Subsidies A subsidy is a payment that the government makes to private producers. By making the subsidy depend on the level of output, the government can induce private decision makers to con- sider external benefits when they make their choices.

Vouchers are tokens that the government provides to households, which they can use to buy specified goods or services. Food stamps are examples of vouchers. The vouchers (food

## FIGURE 17.9 Inefficiency with an

 External Benefit

The market demand curve is the marginal private benefit curve, $D=M B$. The supply curve is the marginal social cost curve, $S=$ MSC. Market equilibrium at a tuition of $\$ 15,000$ a year and 7.5 million students is inefficient because marginal social benefit exceeds marginal social cost. The efficient quantity is 15 million students. A deadweight loss arises (gray triangle) because too few students enroll in college.
stamps) can be spent only on food and are designed to improve the diet and health of extremely poorfamilies.
School vouchers have been advocated as a means of improving the quality of education and are used in Washington, D.C. A school voucher allows parents to choose the school their children will attend and to use the voucher to pay part of the cost. The school cashes the vouchers to pay its bills. A voucher couldbe provided to a college student in a similar way, andalthough technically not a voucher, a federal Pell Grant has a similar effect.
Because vouchers can be spent only on a specified item, they increase the willingness to pay for that itemand so increase the demand for it.

## Illustrating an Efficient Outcome

Figure 17.10 illustrates an efficient outcome. With marginal social cost curve MSC and marginal socialbenefit curve MSB, the efficient number of college students is 15,000 . The marginal private benefit curve $M B$ tells us that 15,000 students will enroll only if the tuition is $\$ 10,000$ per year. But the marginal social cost of 15,000 students is $\$ 25,000$ per year. To enable the marginal social cost to be paid, taxpayers must pay the balance of $\$ 15,000$ per student per year.

To achieve the efficient outcome with public production, public colleges receive funds from government equal to $\$ 15,000$ per student per year, charge tuition of $\$ 10,000$ per student per year, and enroll 15 million students.
To achieve the efficient outcome with private provision, private colleges receive a government subsidy of $\$ 15,000$ perstudent peryear. This subsidy reduces the colleges' costs and would make their marginal cost equal to $\$ 10,000$ per student at the efficient quantity. Tuition of $\$ 10,000$ covers this cost, and the subsidy of $\$ 15,000$ per student covers the balance of the cost.
To achieve an efficient outcome with vouchers, government provides a voucher to each student with a value equal to the marginal external benefit at theefficient number of students. In Fig. 17.10 , the efficient number of students is 15 million and the voucher is valued at $\$ 15,000$ per student. Each student pays $\$ 10,000$ tuition and gives the college a $\$ 15,000$ voucher. The colleges receive $\$ 25,000$ per student, which equals their marginal cost.

## Bureaucratic Inefficiency and Government Failure

You've seen three government actions that achieve an efficient provision of a good with an external benefit. In each case, if the government estimates the marginal external benefit correctly and makes marginal social benefit equal to marginal social cost, the out- come is efficient.
Does the comparison that we've just made mean that pubic provision, subsidized private provision, and vouchers are equivalent? It does not. And the reason lies in something that you've already encountered in your study of public goods in Chapter 16 - the behavior of bureaucrats combined with rational ignorance that leads to government failure.

Problems with Public Production Public colleges and schools are operated by a bureaucracy and are subject to the same problems as the provision of public goods. Bureaucrats seek to maximize their budgets, which brings inefficient overspending.
Inthe case of colleges and schools, overspending doesn't mean overprovision. Just the opposite: People complain about underprovision. The overspending is budget padding and waste.
Education bureaucrats incur costs that exceed the minimum efficient cost. They might hire more assistants than the number needed to do their work efficiently; give themselves sumptuous offices; get generous expense allowances; build schools in the wrong places where land costs are too high.

Economists have compared the costs of private and public schools and have found that the costs per student of public schools are on the order of three times the costs of comparable private schools (see Talking with Carolyn Hoxby on p. 456.)
But bureaucracy can be limited and public production made more efficient. Charter schools (see Economics in Action on p . 447) are showing one way of cutting costs even while improving quality.

## FIGURE 17.10 An Efficient Outcome with an

 External Benefit

The efficient number of college students is 15 million, where marginal social benefit equals marginal social cost. With the demand and marginal private benefit curve, $D=M B$, the price at which the efficient number will enroll is $\$ 10,000$ per year. If students pay this price, the taxpayer must somehow pay the rest, which equals the marginal external benefit at the efficient quantity $-\$ 15,000$ per student per year.

Problems with Private Subsidies Subsidizing private producers might overcome some of the problems created by public production. A private producer has anincentive to produce at minimum cost and avoid thebudget padding of a bureaucratic producer. But two problems arise with private subsidies.
First, the subsidy budget must be allocated by a bureau. A national, state, or local department of education must lobby for its own budget and allocate
this budget between school subsidies and its own administration costs. To the extent that the bureaucrats succeed in maximizing their own administration budget, they siphon off resources from schools and a problem similar to that of public production arises. Second, it is in the self-interest of subsidized producers to maximize their subsidy. These producers might even spend some of the subsidy they receive lobbying for an even bigger one.
So subsidized private provision is unlikely to achieve an efficient allocation of resources. What about the third method: vouchers?

Are Vouchers the Solution? Vouchers have four advantages over the other two approaches:

1. Vouchers can be used with public production, private provision, or competition between the two.
2. Governments can set the value of vouchers and the total voucher budget to overcome bureaucratic overprovision and budget padding.
3. Vouchers spread the public contribution thinly across millions of consumers, so no one consumer has an interest in wasting part of the value received in lobbying for overprovision.
4. By giving the buying power to the final consumer, producers must compete for business and provide a high standard of service at the lowest attainable cost.

For these four reasons, vouchers are popular with economists. But they are controversial and opposed by most education administrators and teachers.
InTheEconomics ofSchoolChoice, a book edited by Caroline M. Hoxby, economists study the effect of school choice on student achievement and school productivity and show how vouchers can be designed to achieve their goals while avoiding their potential pitfalls. Caroline Hoxby is confident that she can design a voucher that best achieves any educational and school performanceobjective.
Economics in the News on pp. 448-449 looks at the economics of President Obama's proposals for lowering carbon emissions.
The next two chapters examine the third big question of economics: For whom are goods and services produced? We examine the markets for factors of production and discover how factor incomes and the distribution of income are determined.

## ECONOMICS IN ACTION

## EDUCATION Quality and Cost: Charter Schools

A charter school is a public school, so it is funded like a regular public school but is free to make its own education policy. Around 4,000 charter schools in 40 states, teaching more than 1 million students, are in operation today. When the demand for places in a charterschool exceeds the supply, students are chosen by lottery. This method of selection provides rich data fortesting the performance of charter schools.
Are charter schools succeeding? Success has two dimensions: educational standards attained and cost per student. Charter schools perform well on both dimensions. They achieve higher standards and cost less. Charter school students achieve better test scores in math and reading than equivalent students who apply to but randomly don't get into a charter school.
Charter schools also achieve this higher standard at lower cost. For example, in Detroit, cost per student in charter schools is 25 percent less than in regular publicschools in equivalent areas.

## ECONOMIC ANALYSIS

- Because coal used to generate electricity creates twice as much $\mathrm{CO}_{2}$ as gas, by converting from coal to gas, $\mathrm{CO}_{2}$ emissions can be cut without lowering the amount of electricity consumed.
- A switch from coal to gas can be achieved by mandating power utilities or by confronting them with a price for carbon emissions that makes a switch in their self-interest.
- Regardless of the method used, a switch from coal to gas decreases the demand for coal and increases the demand for gas. The price of coal falls and the price of gas rises.
- The inefficiency arising from carbon emissions is measured by the deadweight loss it creates. Switching from coal to gas decreases the deadweight loss from using coal, increases the deadweight loss from using gas, and decreases the overall deadweight loss.
- But deadweight loss remains unless the consumption of electricity is cut to the point at which its marginal benefit and price equals its marginal social cost.
- The figures illustrate what will happen if the EPA plan is achieved.
- In Fig. 1, the demand for and marginal social benefit of electricity generated using coal is $D=$ MSB. The supply of coal and the marginal private cost curve is $S=M C$, and the marginal social benefit curve is MSC. The efficient quantity of coal-generated electricity is 1.2 terawatt-hours per year. But the equilibrium quantity is 1.6 terawatt-hours per year, so overproduction brings a deadweight loss as shown by the area of the gray triangle.
- The EPA plan decreases the use of coal. In Fig. 1, the EPA target for coal generators is the efficient quantity, so the deadweight loss is eliminated.
- Figure 2 illustrates the effects of the switch from coal to gas. Gas generators are a bit more costly to operate than coal generators, so the marginal cost of producing electricity increases from $M C_{0}$ to $M C_{1}$ when gas generators replace coal generators.
- But gas is cleaner than coal, so emissions decrease and the marginal external cost falls. The marginal social cost falls from $M S C_{0}$ to $M S C_{1}$.
- Because the marginal cost of generating electricity increases, supply of electricity decreases, the price rises, and the equilibrium quantity decreases.
- The deadweight loss from generating electricity decreases from the area of the light gray triangle to the area of the dark gray triangle.


Figure 1 Making Coal Generation Efficient


Figure 2 More Efficient but Remains Inefficient

- The success of this plan to lower carbon emissions will depend on the ways in which incentives are changed. The cost of using coal must rise to induce power utilities to find the most efficient substitutes for coal.
- The planned changes in the quantity and price of electricity leave a large amount of carbon emissions in place.


# PART SIX: FACTOR MARKETS, INEQUALITY, AND UNCERTAINTY 

## CHAPTER 18: MARKETS FOR FACTORS OF PRODUCTION

After studying this chapter, you will be able to:

- Describe the anatomy of factor markets
- Explain how the value of marginal product determines the demand for a factor of production
- Explain how wage rates and employment aredeter- mined and how labor unions influence labor markets
- Explain how capital and land rental rates and natural resource prices are determined

Ateacher in a publicschool earns much morethanan equally good teacher in a charter school, and your college basketball coach earns much more than your economics professor. Why? What determines the wages that people earn? Wages are important, but finding a job is important too. Why are so many manufacturing jobs disappearing and what new jobs are being created? In this chapter, we study labor markets as well as markets for capital and natural resources. In Economics in the News at the end of the chapter we answer the question about teachers' pay.

## The Anatomy of Factor Markets

The four factors of production are:

- Labor
- Capital
- Land (natural resources)
- Entrepreneurship

Let's take a brief look at the anatomy of the markets in which these factors of production are traded.

## Markets for Labor services

Labor services are the physical and mental work effort that people supply to produce goods and services. A labor market is a collection of people and firms who trade labor services. The price of labor services is thewage rate.
Some labor services are traded day by day. These services are called casual labor. People who pick fruit and vegetables often just show up at a farm and takewhatever work is available that day. But most laborservices are traded on a job contract.
Most labor markets have many buyers and many sellers and are competitive. In these markets, supply and demand determine the wage rate and quantity of labor employed. Jobs expand when demand increases and jobs disappear when demand decreases. In some labor markets, a labor union operates like a monopoly on the supply-side of the labor market. In this type of labor market, a bargaining process between the union and the employer determines the wage rate.

We'll study both competitive labor markets and labor unions in this chapter.

## Markets for Capital Services

Capital consists of the tools, instruments, machines, buildings, and other constructions that have been produced in the past and that businesses now use to produce goods and services. These physical objects are themselves goods - capital goods. Capital goodsare traded in goods markets, just as bottled water and toothpaste are. The price of a dump truck, a capital good, is determined by supply and demand in the market for dump trucks. This market is not a market for capital services.
A market for capital services is a rental market-a market in which the services of capital are hired. An example of a market for capital services is the vehicle rental market in which Avis, Budget, Hertz U-Haul, and many other firms offer automobiles and, trucks for hire. The price in a capital services market is a rental rate. Most capital services are not traded in a market. Instead, a firm buys capital and uses it itself. The services of the capital that a firm owns and operates have an implicit price that arises from depreciationand interest costs (see Chapter 10, pp. 262-263).
You can think of this price as the implicit rental rate of capital. Firms that buy capital and use it themselves are implicitly renting the capital to themselves.

## Markets for Land Services and Natural Resources

Land consists of all the gifts of nature-natural resources. The market for land as a factor of production is the market for the services ofland-the use of land. The price of the services of land is a rental rate.
Most natural resources, such as farm land, can be used repeatedly. But a few natural resources are non-renewable. Nonrenewable natural resources are resources that can be used only once. Examples are oil, natural gas, and coal. The prices of nonrenewable 'natural resources are determined in global commodity markets and are called commodity prices.

## Entrepreneurship

Entrepreneurial services are not traded in markets. : Entrepreneurs receive the profit or bear the loss that. '.results from their business decisions.

## The Demand for a Factor of Production

The demand for a factor of production is a derived demand-it is derived from the demand for the goods and services that the labor produces. You've seen, in Chapters 10 through 15, how a firm determines its profit-maximizing output. The quantities of factors of production demanded are a consequence of the firm's output decision. A firm hires the quantities of factors of production that produce the firm's profit- maximizing output.
To decide the quantity of a factor of production to hire, a firm compares the cost of hiring an additional unit of the factor with its value to the firm. The cost of hiring an additional unit of a factor of production is the factor price. The value to the firm of hiring one more unit of a factor of production is called the factor's value of marginal product. We calculate the value of marginal product as the price of a unit of output multiplied by the marginal product of the factor of production.

To study the demand for a factor of production, we'll use labor as the example. But what you learn here about the demand for labor applies to the demand for all factors of production.

## Value of a Marginal Product

Table 18.1 shows you how to calculate the value of marginal product of labor at Angelo's Bakery The first two columns show Angelo's total product schedule-the number of loaves per hour that each quantity of labor can produce. The third column shows the marginal product of labor-the change intotal product that results from a one-unit increase in the quantity of labor employed. (See Chapter 11, pp. 287-290 for a refresher on product schedules.)
Angelo can sell bread at the going market price of $\$ 2$ a loaf Given this information, we can calculate the value of marginal product (fourth column). It equals price multiplied by marginal product. For example, the marginal product of hiring the second worker is 6 loaves. Each loaf sold brings in $\$ 2$, so the value of marginal product of the second worker is \$12 (6 loaves at \$2 each).

## A Firm's Demand of Labor

The value of marginal product of labor tells us what an additional worker is worth to a firm. It tells us the revenue that the firm earns by hiring one more worker. The wage rate tells us what an additional worker costs a firm.
The value of marginal product of labor and the wage rate together determine the quantity of labor demanded by a firm. Because the value of marginal product decreases as the quantity of labor employed increases, there is a simple rule for maximizing profit: Hire the quantity of labor at which the value of marginal product equals the wage rate.
If the value of marginal product of labor exceeds the wage rate, a firm can increase its profit by hiring one more worker. If the wage rate exceeds the value of marginal product of labor, a firm can increase its profit by firing one worker. But if the wage rate equals the value of marginal product of labor, the firm cannot increase its profit by changing the number of workers it employs. The firm is making the maximum possible profit. So The quantity of labor demanded by a firm is the quantity at which the value of marginal product of labor equals the wage rate.

## A firm's Demand for Labor Curve

A firm's demand for labor curve is derived from itsvalue of marginal product curve. Figure 18.1 showsthese two curves. Figure 18.1(a) shows the value of marginal product curve at Angelo's Bakery. The blue bars graph the numbers in Table 18.1. The curve labeled VMP is Angelo's value of marginal product curve.
If the wage rate falls and other things remain the same, a firm hires more workers. Figure 18.1(b) shows Angelo's demand for labor curve.

TABLE 18.1 Value of Marginal Product at Angelo's Bakery

|  | Quantity of labor <br> (L) <br> (workers) | Total product (TP) (lloaves per hour) | Marginal product $(M P=\Delta T P / \Delta L)$ (looves per worker) | Value of marginal product $(V M P=M P \times P)$ (dollars per worker) | The value of marginal product of labor equals the price of the product multiplied by marginal product of labor. If Angelo's hires 2 workers, the |
| :---: | :---: | :---: | :---: | :---: | :---: |
| A | 0 | 0 | . . 7 | 14 | marginal product of the second worker is 6 loaves (in the third |
| B | 1 | 7 | 6 | 12 | column). The price of a loaf is |
| $C$ | 2 | 13 | $\cdots 5$ | 10 | $\$ 2$, so the value of marginal product of the second worker is |
| D | 3 | 18 | 4 | 8 | $\$ 2$ a loaf multiplied by 6 loaves, which is $\$ 12$ (in fourth column). |
| E | 4 | 22 | . . . 3 | 6 |  |
| F | 5 | 25 |  |  |  |

Suppose the wage rate is $\$ 10$ an hour. You cansee in Fig.18.1(a) that if Angelo hires 2 workers, the value of the marginal product of labor is $\$ 12$ an hour. At a wage rate of $\$ 10$ an hour, Angelo makes a profitof $\$ 2$ an hour on the second worker. If Angelo hires athird worker, the value of marginal product of that worker is $\$ 10$ an hour. So on this third worker, Angelo breaks even.
If Angelo hired 4 workers, his profit would fall. The fourth worker generates a value of marginal product of only $\$ 8$ an hour but costs $\$ 10$ an hour, so Angelo does not hire the fourth worker. When the wage rate is $\$ 10$ an hour, the quantity of labor demanded by Angelo is 3 workers.
Figure 18.l(b) shows Angelo's demand for labor curve, D. At $\$ 10$ an hour, the quantity of labor demanded by Angelo is 3 workers. If the wage rate increased to $\$ 12$ an hour, Angelo would decrease the quantity of labor demanded to 2 workers. If the wage rate decreased to $\$ 8$ an hour, Angelo would increase the quantity of labor demanded to 4 workers. A change in the wage rate brings a change in the quantity of labor demanded and a movement along the demand for labor curve.
A change in any other influence on a firm's labor- hiring plans changes the demand for labor and shifts the demand for labor curve.

## Changes in a Firm's Demand for Labor

A firm's demand for labor depends on

- The price of the firm's output
- The prices of other factors of production
- Technology

FIGURE 18.1 The Demand for Labor at Angelo's Bakery

(a) Value of marginal product


## (b) Demand for Iabor

Angelo's Bakery can sell any quantity of bread at $\$ 2$ a loaf. The blue bars in part (a) represent the firm's value of marginal product of labor (based on Table 18.1). The line labeled VMP is the firm's value of marginal product curve. Part (b) shows Angelo's demand for labor curve. Angelo hires the quantity of labor that makes the value of marginal product equal to the wage rate. The demand for labor curve slopes downward because the value of marginal product diminishes as the quantity of labor employed increases.

The Price of the Firm's Output The higher the price of a firm's output, the greater is the firm's demand for labor. The price of output affects the demand for labor through its influence on the value of marginal product of labor. A higher price for the firm's outputincreases the value of marginal product of labor.
A change in the price of a firm's output leads to a shift in the firm's demand for labor curve. If the price of the firm's output increases, the demand for labor increases and the demand for labor curve shifts rightward.
For example, if the price of bread increased to $\$ 3$ a loaf, the value of marginal product of Angelo's fourth worker would increase from $\$ 8$ an hour to $\$ 12$ an hour. At a wage rate of $\$ 10$ an hour, Angelo would now hire 4 workers instead of 3.

The Prices of Other Factors of Production If the price of using capital decreases relative to the wage rate, a firm substitutes capital for labor and increases the quantity of capital it uses. Usually, the demand for labor will decrease when the price of using capital falls. For example, if the price of a bread-making machine falls, Angelo might decide to install one machine and lay off a worker. But the demand for labor could increase if the lower price of capital led to a sufficiently large increase in the scale of production. For example, with cheaper machines available, Angelo might install a machine and hire more labor to operate it. This type of factor substitution occurs in the long run when the firm can change the size of its plant.

Technology New technologies decrease the demand for some types of labor and increase the demand for other types. For example, if a new automated bread- making machine becomes available, Angelo might install one of these machines and fire most of his workforce-a decrease in the demand for bakery workers. But the firms that manufacture and service automated bread-making machines hire more labor, so there is an increase in the demand for this type of labor. An event similar to this one occurred during the 1990s when the introduction of electronic telephone exchanges decreased the demand for telephone operators and increased the demand for computer programmers and electronics engineers.
Table 18.2 summarizes the influences on a firm's demand for labor.

## Labor Markets

Labor services are traded in many different labormarkets. Examples are markets for bakery workers, van drivers, crane operators, computer support specialists, air traffic controllers, surgeons, and economists. Some of these markets, such as the market for bakery workers, are local. They operate in a given neighborhood or town. Some labor markets, such as the market for air traffic controllers, are national. Firms and workers search across the nation for the right match of worker and job. And some labor markets are global, such as the market for superstarhockey, basketball, baseball, and soccer players.
We'll look at a local market for bakery workers as an example. First, we'll look at a competitive labormarket. Then, we'll see how monopoly elements can influence a labor market.

## TABLE 18.2 A Firm's Demand for Labor

## The Law of Demand

(Movements along the demand curve for labor)
The quantily of labor demanded by a firm

| Decreases if: | Increases if: |
| :--- | :--- |
| - The wage rate | - The wage rate |
| increases | decreases |

Changes in Demand
(Shifts in the demand curve for labor)
A firm's demand for labor
Decreases if: Increases if:

- The price of the firm's output decreases
- The price of a substitute for labor falls
- The price of a complement of labor rises
- A new technology or new capital decreases the marginal product of labor
- The price of the firm's output increases
- The price of a substitute for labor rises
- The price of a complement of labor falls
- A new technology or new capital increases the marginal product of labor


## A Competitive Labor Market

A competitive labor market is one in which many firms demand labor and many households supply labor.

Market Demand for Labor Earlier in the chapter, you saw how an individual firm decides how much laborto hire. The market demand for labor is derived fromthe demand for labor by individual firms. We determine the market demand for labor by adding together the quantities of labor demanded by all the firms in the market at each wage rate. (The market demand for a good or service is derived in a similar way-see Chapter 5, pp. 146-147.) Because each firm's demand for labor curve slopes downward, the market demand for labor curve also slopes downward.

The Market Supply of Labor The market supply of labor is derived from the supply of labor decisionsmade by individual households.
Individual's LaborSupply Decision People can allocate their time to two broad activities: labor supply and leisure. (Leisure is a catch-all term. It includes all activities other than supplying
labor.) For most people, leisure is more fun than work so to induce them to work they must be offered a wage. Think about the labor supply decision of Jill, one of the workers at Angelo's Bakery. Let's see how the wage rate influences the quantity of labor she is willing to supply.

Reservation Wage Rate Jill enjoys her leisure time, and she would be pleased if she didn't have to spendher time working at Angelo's Bakery. But Jill wantsto earn an income, and as long as she can earn a wage rate of at least $\$ 5$ an hour, she's willing to work. This wage is called her reservation wage. At any wage rate above her reservation wage, Jill sup-plies some labor. The wage rate at Angelo's is $\$ 10$ an hour, and at that wage rate, Jill chooses to work 30 hours a week. At a wage rate of $\$ 10$ an hour, Jill regards this use of her time as the best available. Figure18.2 illustrates.

Backward-Bending Labor Supply Curve If Jill were offered a wage rate between $\$ 5$ and $\$ 10$ an hour, shewould want to work fewer hours. If she were offered a wage rate above $\$ 10$ an hour, she would want to work more hours, but only up to a point. If Jill could earn $\$ 25$ an hour, she would be willing to work 40 hours a week (and earn $\$ 1,000$ a week). But at a wage rate above $\$ 25$ an hour, with the goods and services that Jill can buy for $\$ 1,000$, her priority would be a bit more leisure time. So if the wage rate increased above $\$ 25$ an hour, Jill would cut back on her work hours and take more leisure. Jill's labor supply curve eventually bends backward. Jill's labor supply decisions are influenced by a substitution effect and an income effect.

Substitution Effect At wage rates below $\$ 25$ an hour, the higher the wage rate Jill is offered, the greater isthe quantity of labor that she supplies. Jill's wage rateis her opportunity cost of leisure. If she quits work anhour early to catch a movie, the cost of that extra hour of leisure is the wage rate that Jill forgoes. The higher the wage rate, the less willing Jill is to forgothe income and take the extra leisure time. This tendency for a higher wage rate to induce Jill to work longer hours is a substitution effect.

Income Effect The higher Jill's wage rate, the higher is her income. A higher income, other things remaining the same, induces Jill to increase her demand for most goods and services. Leisure is one of those goods. Because an increase in income creates an increase in the demand for leisure, it also creates a decrease in the quantity of labor supplied.

Market Supply Curve The market supply curve shows the quantity of labor supplied by all households in a particular job market. It is found by adding together the quantities of labor supplied by all households to a given job market at each wage rate, so the greater the number of households (the greater is the working-age population), the greater is the market supply of labor.
Despite the fact that an individual's labor supply curve eventually bends backward, the market supply curve of labor slopes upward. The higher the wage rate for bakery workers, the greater is the quantity of bakery workers supplied in that labor market.

One reason why the market supply curve doesn't bend backward is that different households have different reservation wage rates and different wage rates at which their labor supply curves bend backward.
Also, along a supply curve in a particular job market, the wage rates available in other job markets remain the same. For example, along the supply curve of bakers, the wage rates of salespeople and all other types of labor are constant. Let's now look at labor market equilibrium.

FIGURE 18.2 Jill's Labor Supply Curve


Jill's labor supply curve is $S$. Jill supplies no labor at wage rates below her reservation wage of $\$ 5$ an hour. As the wage rate rises above $\$ 5$ an hour, the quantity of labor that Jill supplies increases to a maximum of 40 hours a week at a wage rate of $\$ 25$ an hour. As the wage rate rises above $\$ 25$ an hour, Jill supplies a decreasing quantity of labor: her labor supply curve bends backward. The income effect on the demand for leisure dominates the substitution effect.

Competitive Labor Market Equilibrium Labor market equilibrium determines the wage rate and employment. In Fig. 18.3, the market demand curve for bakery workers is $D$ and the market supply curve of bakery workers is 5 . The equilibrium wage rate is $\$ 10$ an hour, and the equilibrium quantity is 300 bakery workers. If the wage rate exceeded $\$ 10$ an hour, there would be a surplus of bakery workers. More people would be looking for jobs in bakeries than firms were willing to hire. In such a situation, the wage rate would fall as firms found it easy to hire people at a lower wage rate. If the wage rate were less than $\$ 10$ an hour, there would be a shortage of bakery workers. Firms would not be able to fill all the positions they had available. In this situation, the wage
rate would rise as firms found it necessary to offer higher wages to attract labor. Only at a wage rate of $\$ 10$ an hour are there no forces operating to change the wage rate.
$\begin{aligned} \text { FIGURE 18.3 } & \text { The Market for } \\ & \text { Bakery Workers }\end{aligned}$


A competitive labor market coordinates firms' and households' plans. The market is in equilibrium-the quantity of labor demanded equals the quantity supplied at a wage rate of $\$ 10$ an hour when 300 workers are employed. If the wage rate exceeds $\$ 10$ an hour, the quantity supplied exceeds the quantity demanded and the wage rate will fall. If the wage rate is below $\$ 10$ an hour, the quantity demanded exceeds the quantity supplied and the wage rate will rise.

## Differences and Trends in Wage Rates

You can use what you've learned about labor markets to explain some of the differences in wage rates across occupations and the trends in wage rates. Wage rates are unequal, and Economics in Action on this page shows a sample of the inequality in wages in 2013. The differences in wage rates across occupations-including college professors and basketball coaches-are driven by differences in demand 'and supply in labor markets. The highest wage rates are earned in occupations where the value of marginal product is highest and where few people have the ability and training to perform the job.

Rising Wage Rates Wage rates increase over timeand; trend upward. The reason is that the value of marginal product of labor also increases over time.Technological change and the new types of capitalthat it brings make workers more productive. Withgreater labor productivity, the demand for labor increases, which increases the wage rate. Even jobs in which physical productivity doesn't increase experience increases in the value of marginal product. Child
care is an example. A child-care worker can't care for an increasing number of children, but an increasing \{ number of parents who earn high wages are willing to hire child-care workers. The value of marginal product of these workers increases, so the demand for their services increases, and so does their wage rate.

Increased Wage Inequality In recent years wage inequality has increased. High wage rates have increased more rapidly than the low ones, and some low wage rates have stagnated or even fallen. The reasons are complex and not fully understood, but thebest explanation is that there is an interaction between technology and education. The new information technologies of the 1990s and 2000s made well-educated, skilled workers more productive, so it raised their wage rates. For example, the computer created the jobs and increased the wage rates of computer programmers and electronic engineers. These same technologies destroyed some low-skilled jobs. For example, the ATM took the jobs and lowered the wage rate of bank tellers, and automatic telephones took the jobs of telephone operators. Another reason for increased inequality is that globalization has brought increased competition for low-skilled workers and at the same time opened globalmarkets for highskilled workers.

## A Labor Market with a Union

A labor union is an organized group of workers that aims to increase the wage rate and influence other job conditions. Let's see what happens when a union enters a competitive labor market.

Influences on LaborSupply One way of raising the wage rate is to decrease the supply of labor. In somelabor markets, a union can restrict supply by controlling entry into apprenticeship programs or by influ-encing job qualification standards. Markets for skilled workers, doctors, dentists, and lawyers are the easiest ones to control in this way.
If there is an abundant supply of nonunion labor, a union can't decrease supply. For example, in the market for farm labor in southern California, the flow of nonunion labor from Mexico makes it difficult for a union to control the supply.
On the demand side of the labor market, the union faces a tradeoff: The demand for labor curve slopes downward, so restricting supply to raise the wage rate costs jobs. For this reason, unions also try to influence the demand for union labor.

Influences on Labor Demand A union tries to increase the demand for the labor of its members in four main ways:

1. Increasing the value of marginal product of its members by organizing and sponsoring training schemes and apprenticeship programs, and by professional certification.
2. Lobbying for import restrictions and encouraging people to buy goods made by unionized workers.
3. Supporting minimum wage laws, which increase the cost of employing low-skilled labor and leadfirms to substitute high-skilled union labor forlow-skilled nonunion labor.
4. Lobbying for restrictive immigration laws to decrease the supply of foreign workers.

Labor Market Equilibrium with a Union Figure 18.4 illustrates what happens to the wage rate and employment when a union successfully enters a competitive labor market. With no union, the demand curve is $D e$, the supply curve is $S c$, the wage rate is $\$ 10$ anhour, and 300 workers have jobs.
Now a union enters this labor market. First, look at what happens if the union has sufficient control over the supply of labor to be able to restrict supply below its competitive level-to Su. If that is all the union is able to do, employment falls to 200 workers and the wage rate rises to $\$ 15$ an hour.

FIGURE 18.6 Minimum Wage Law in Monopsony


In a monopsony labor market, the wage rate is $\$ 10$ an hour and 100 workers are hired. If a minimum wage law increases the wage rate to $\$ 15$ an hour, the wage rate rises to this level and employment increases to 150 workers.
Suppose now that the union is also able to in- crease the demand for labor to Du. The union can get an even bigger increase in the wage rate and with a smaller fall in employment. By maintaining the restricted labor supply at Su , the union increases the wage rate to $\$ 20$ an hour and achieves an employment level of 250 workers.
Because a union restricts the supply of labor in the market in which it operates, the union's actions spill over into nonunion markets. Workers who can't get union jobs must look elsewhere for work. This action increases the supply of labor in nonunion markets and lowers the wage rate in those markets. This spillover effect further widens the gap between union and nonunion wages.

Monopsony in the Labor Market Not all labor markets in which unions operate are competitive. Rather, some are labor markets in which the employer possesses market power and the union enters to try to counteract that power.
A market in which there is a single buyer is called a monopsony. A monopsony labor market has one employer. In some parts of the country, managed healthcare organizations are the major employer of healthcare professionals. In some communities, Wal-Mart is the main employer of salespeople. These firms have monopsony power.
A monopsony acts on the buying side of a market in a similar way to a monopoly on the selling side. The firm maximizes profit by hiring the quantity of labor that makes the marginal cost of labor equal to the value of marginal product of labor and by paying the lowest wage rate at which it can attract this quantity of labor.
Figure 18.5 illustrates a monopsony labor market. Like all firms, a monopsony faces a downward-sloping value of marginal product curve, VMP, which is its demand for labor curve, D-the curve labeled $V M P=D$ in the figure.
What is special about monopsony is the marginal cost of labor. For a firm in a competitive labor market, the marginal cost of labor is the wage rate. For a monopsony, the marginal cost of labor exceeds the wage rate. The reason is that being the only buyer in the market, the firm faces an upward-sloping supply of labor curve - the curve $S$ in the figure. To attract one more worker, the monopsony must offer a higher wage rate. But it must pay this higher wage rate to all its workers, so the marginal cost of a worker is the wage rate plus the increased wage bill that arises from paying all the workers the higher wage rate. The supply curve is now the average cost of labor curve and the relationship between the supply curve and the marginal cost of labor curve, MCL, is similar to that between a monopoly's demand curve and marginal revenue curve (see p. 338). The relationship between the supply curve and the MCL curve is also similar to that between a firm's average cost curve and marginal cost curve (see pp. 296-297).
To find the profit-maximizing quantity of labor to hire, the monopsony sets the marginal cost of labor equal to the value of marginal product of labor. In Fig. 18.5, this outcome occurs when the firm employs 100 workers.
To hire 100 workers, the firm must pay $\$ 10$ an hour (on the supply of labor curve). Each worker is paid $\$ 10$ an hour, but the value of marginal product of labor is $\$ 20$ an hour, so the firm makes an economic profit of $\$ 10$ an hour on the marginal worker. If the labor market in Fig. 18.5 were competitive, the equilibrium wage rate and employment would be determined by the demand and supply curves. The wage rate would be $\$ 15$ an hour, and 150 workers would be employed. So compared with a competitive labor market, a monopsony pays a lower wage rate and employs fewer workers.

A Union and a Monopsony A union is like a monopoly. If the union (monopoly seller) faces a monopsonybuyer, the situation is called bilateral monopoly. An example of bilateral monopoly is the Writers Guild of America that represents film, television, and radio writers, and an employers' alliance that represents CBS, MGM, NBC, and other entertainment companies. Every three years, the Writers Guild and the employers' alliance negotiate a pay deal. In bilateral monopoly, the outcome is determined by bargaining, which depends on the costs that each party can inflict on the other. The firm can shut down temporarily and lock
out its workers, and the workers can shut down the firm by striking. Each party estimates the other's strength and what it will lose if it does not agree to the other's demands. Usually, an agreement is reached without a strike or a lockout. The threat is usually enough to bring the bargaining parties to an agreement. When a strike or lockout does occur, it is because one party has mis- judged the costs each party can inflict on the other.
Such an event occurred in November 2007 when the writers and entertainment producers failed to agree on a compensation deal. A 100-day strike followed that ended up costing the entertainment industry anestimated \$2billion.
FIGURE 18.7 A Rental Market for Capital


The value of marginal product of tower cranes, VMP, determines the demand, $D$, for tower crane rentals. With the supply curve $S$, the equilibrium rental rate is $\$ 1,000$ a day and 100 cranes are rented.
In the example in Fig. 18.5, if the union and employer are equally strong, and each party knows the strength of the other, they will agree to split the gap between $\$ 10$ (the wage rate on the supply curve) and $\$ 20$ (the wage rate on the demand curve) and agree to a wage rate of $\$ 15$ an hour.
You've now seen that in a monopsony, a union can bargain for a higher wage rate without sacrificing jobs. A similar outcome can arise in a monopsony labor market when a minimum wage law is enforced. Let's look at the effect of a minimum wage.

Monopsony and the Minimum Wage Ina competitive labor market, a minimum wage that exceeds the equilibrium wage decreases employment (see Chapter 6, pp. 169-170). Ina monopsony labor market, a minimum wage can increase both the wage rate and employment. Let's see how.

Figure 18.6 shows a monopsony labor market without a union. The wage rate is $\$ 10$ an hour and 100 workers are employed.
A minimum wage law is passed that requires employers to pay at least $\$ 15$ an hour. The monopsony now faces a perfectly elastic supply of labor at $\$ 15$ an hour up to 150 workers (along the mini-mum wage line). To hire more than 150 workers, a wage rate above $\$ 15$ an hour must be paid (along the supply curve). Because the wage rate is $\$ 15$ an hour up to 150 workers, so is the marginal cost of labor $\$ 15$ an hour up to 150 workers. To maximize profit, the monopsony sets the marginal cost of labor equal to the value of marginal product of labor (on the demand curve). That is, the monopsony hires 150 workers and pays $\$ 15$ an hour. Theminimum wage law has succeeded in raising the wage rate and increasing the number of workers employed.

## Capital and Natural Resource Markets

The markets for capital and land can be understood by using the same basic ideas that you've seen when studying a competitive labor market. But markets for nonrenewable natural resources are different. We'll now examine three groups of factor markets:

- Capital rental markets
- Land rental markets
- Nonrenewable natural resource markets


## Capital Rental Markets

The demand for capital is derived from the value ofmarginal product of capital. Profitmaximizing firms hire the quantity of capital services that makes the value of marginal product of capital equal to the rental rate of capital. The lower the rental rate of capital, other things remaining the same, the greater is the quantity of capital demanded. The supply of capital responds in the opposite way to the rental rate. The higher the rental rate, other things remaining the same, the greater is the quantity of capital supplied. The equilibrium rental rate makes the quantity of capital demanded equal to the quantity supplied.
Figure 18.7 illustrates the rental market for tower cranes-capital used to construct highrise buildings. The value of marginal product and the demand curve is $V M P=D$. The supply curve is $S$. The equilibrium rental rate is $\$ 1,000$ per day and 100 tower cranes are rented.

Rent-Versus-Buy Decision Some capital services are obtained in a rental market like the market for tower cranes. And as with tower cranes, many of the world's large airlines rent their airplanes. But not all capitalservices are obtained in a rental market. Instead, firms buy the capital equipment that they use. Yousaw in Chapter 10 (pp. 262-263) that the cost of theservices of the capital that a firm owns and operatesitself is an implicit rental rate that arises from depreciation and interest costs. Firms that buy capital implicitly rent the capital to themselves.
The decision to obtain capital services in a rental market rather than buy capital and rent it implicitly is made to minimize cost. The firm compares the cost of explicitly renting the capital and the cost of buying and implicitly renting it. This decision is the same as the one that a household makes in deciding whether to rent or buy a home.

To make a rent-versus-buy decision, a firm must compare a cost incurred in the present with a stream of rental costs incurred over some future period. TheMathematical Note (pp. 476-477) explains how tomake this comparison by calculating the present value of a future amount of money. If the present value of the future rental payments of an item of capital equipment exceeds the cost of buying the capital, thefirm will buy the equipment. If the present value of the future rental payments of an item of capital equipment is less than the cost of buying the capital,the firm will rent (or lease) the equipment.

## Land Rental Markets

The demand for land is based on the same factors as the demand for labor and the demand for capital- the value ofmarginal product ofland Profit-maximizing firms rent the quantity of land at which the value of marginal product of land is equal to the rental rate of land. The lower the rental rat, other thins remaining the same, the greater 1sthe quantity of land demanded. But the supply of land is special: Its quantity is fixed, so the quantity supplied cannot be changed by people's decisions. The supply of each particular block of land is perfectly inelastic. The equilibrium rental rate makes the quantity of land demanded equal to the quantity available.
Figure 18.8 illustrates the market for a 10-acre block of land on 42nd Street in New York City. The quantity supplied is fixed and the supply curve is $S$. The value of marginal product and the demand curve is $V M P=D$. The equilibrium rental rate is $\$ 1,000$ anacre per day. The rental rate of land is high in New York because the willingness to pay for the services produced by that land is high, which in turn makes the VMP of land high. A Big Mac costs more at McDonald's on 42nd Street, New York, than at McDonald's on Jefferson Avenue, St. Louis, but not because the rental rate of land is higher in New York. The rental rate of land is higher in New York because of the greater willingness to pay for a Big Mac (and other goods and services) in New York.

## Nonrenewable Natural Resources Market

The nonrenewable natural resources are oil, gas, and coal. Burning one of these fuels converts it to energy and other by-products, and the used resource cannot be re-used. The natural resources that we use to make metals are also nonrenewable, but they can be used again, at some cost, by recycling them.
Oil, gas, and coal are traded in global commodity markets. The price of a given grade of crude oil is the same in New York, London, and Singapore. Traders, linked by telephone and the Internet, operate these markets around the clock every day of the year. Demand and supply determine the prices and the quantities traded in these commodity markets. We'll look at the influences on demand and supply by considering the global market for crude oil.

The Demand for Oil The two key influences on the demand for oil are

1. The value of marginal product of oil
2. The expected future price of oil

FIGURE 18.8 A Rental Market for Land


The value of marginal product of a 10 -acre block, VMP, determines the rental demand, $D$, for this land. With the supply curve $S$, this block rents for $\$ 1,000$ per acre per day.

The value of marginal product of oil is thefundamental influence on demand. It works in exactly the same way for a nonrenewable resource as it does for any other factor of production. The greater the quantity of oil used, the smaller is the value of marginal product of oil. Diminishing value of marginal product makes the demand curve slope downward. The lower the price, the greater is the quantity demanded.
The higher the expected future price of oil, the greater is the present demand for oil. The expected future price is a speculative influence on demand.
Oil in the ground and oil in storage tanks are inventories that can be held or sold. A trader might plan to buy oil to hold now and to sell it later for a
profit. Instead of buying oil to hold and sell later, the trader could buy a bond and earn interest. The interest forgone is the opportunity cost of holdingthe oil. If the price of oil is expected to rise by a bigger percentage than the interest rate, a trader will hold oil and incur the opportunity cost. In this case, the return from holding oil exceeds the return from holding bonds.

The Supply of Oil The three key influences on the supply of oil are

1. The known oil reserves
2. The scale of current oil production facilities
3. The expected future price of oil

Known oil reserves are the oil that has been discovered and can be extracted with today's technology. This quantity increases over time because advances in technology enable ever-less
accessible sources to be discovered. The greater the size of known reserves, the greater is the supply of oil. But this influence on supply is small and indirect. It operates by changing the expected distant future price of oil. Even a major new discovery of oil would have a negligible effect on current supply of oil.
The scale of current oil production facilities is the fundamental influence on the supply of oil. Producing oil is like any production activity: It is subject to increasing marginal cost. The increasing marginal cost of extracting oil means that the sup- ply curve of oil slopes upward. The higher the price of oil, the greater is the quantity supplied. When new oil wells are sunk or when new faster pumps are installed, the supply of oil increases. When existing wells run dry, the supply of oil decreases. Over time, the factors that increase supply are more powerful than those that decrease supply, so changes in the scale of current oil production facilities increase the supply of oil.
Speculative forces based on expectations about the future price also influence the supply of oil. The higher the expected future price of oil, the smaller is the present supply of oil. A trader with an oil inventory might plan to sell now or to hold and sell later. You've seen that interest forgone is the opportunity cost of holding the oil. If the price of oil is expected to rise by a bigger percentage than the interest rate, it is profitable to incur the opportunity cost of holdingoil rather than selling it immediately.

The Equilibrium Price of Oil The demand for oil and the supply of oil determine the equilibrium price andquantity traded. Figure 18.9 illustrates the market equilibrium.
The value of marginal product of oil, VMP, is the fundamental determinant of demand, and the marginal cost of extraction, MC, is the fundamental determinant of supply. Together, they determine the market fundamentals price.
If expectations about the future price are also based on fundamentals, the equilibrium price is the market fundamentals price. But if expectations about the future price of oil depart from what the marketfundamentals imply, speculation can drive a wedge between the equilibrium price and the market fundamentals price.

The Hotelling Principle Harold Hotelling, an economist at Columbia University, had an incredible idea.Traders expect the price of a nonrenewable natural, resource to rise at a rate equal to the interest rate. We call this idea the Hotelling principle. Let's see why it is correct. You've seen that the interest rate is the opportunity cost of holding an oil inventory. If the price of oil is expected to rise at a rate that exceeds the interest rate, it is profitable to hold a bigger inventory.
Demand increases, supply decreases, and the price rises. If the interest rate exceeds the rate at which the price of oil is expected to rise, it is not profitable to hold an oil inventory. Demand decreases, supply increases, and the price falls. But if the price of oil is expected to rise at a rate equal to the interest rateholding an inventory of oil is just as good as holding bonds. Demand and supply don't change and the price does not change. Only when the price of oil is, expected to rise at a rate equal to the interest rate is the price at its equilibrium.

## FIGURE 18.9 A Nonrenewable Natural Resource Market



The value of marginal product of a natural resource, VMPand the marginal cost of extraction, MC, determine the market fundamentals price. Demand, D, and supply, S, which determine the equilibrium price, are influenced by the expected future price. Speculation can bring a gap between the market fundamentals price and the equilibrium price.

## ECONOMICS IN ACTION

## The World and U.S. Markets for Oil

The world produced 85 million barrels of oil per day in2013 and the price was steady at around $\$ 110$ a barrel. Although the United States imports oil from other countries, most of it comes from close to home and domestic production has increased in recent years.
Figure 1on page 473 provides the details: Only 12 percent of the U.S. oil supply comes from the Middle East and one third comes from Canada, Mexico, and other Western Hemisphere countries.
Even if the United States produced all its own oil, it would still face a fluctuating global price. U.S. producers would not willingly sell to U.S. buyers for a price below the world price. So energy independence doesn't mean an independent oil price. The Hotelling principle tells us that we must expect the price of oil to rise at a rate equal to the interest rate. But expecting the price to rise at a rate equal to the interest rate doesn't mean that the price will rise at this rate. As you can see in Fig. 2, the price of oil over the past 50 or so years has not followed the path predicted by the Hotelling principle.
The forces that influence expectations are not well understood. The expected future price of oil depends on its expected future rate of use and the rate of discovery of new sources of supply. One person's expectation about a future price also depends on guesses about other people's expectations. These guesses can change abruptly and become self-
reinforcing. When the expected future price of oil changes for whatever reason, demand and supply change, and so does the price. Prices in speculative markets are always volatile.


Figure 2 The Price of Oil and Its Hotelling Path

## ECONOMIC ANALYSIS

- Teachers in charter schools are paid much lower salaries (wage rates), on average, than teachers in traditional public schools for three main reasons:

1. Charter school teachers are less experienced.
2. Charter schools have tighter budgets than traditional public schools.
3. Charter school teachers are not (usually) unionized.

- Charter school teachers are, on average, less experienced than public school teachers, so their marginal product is less than that of traditional public school teachers, which means that their VMP is lower and the demand for their services is reduced.
- Because charter schools have tighter budgets than public schools, the market values the services of charter school teachers at a lower (implicit) price, which further lowers their VMP and reduces the charter schools' demand for teachers.
- Some teachers say that they value the freedom and autonomy that a charter school gives them. This preference for working in a charter school increases the supply of charter school teachers (relative to what it would be without the autonomy.)
- The combination of a reduced demand and greater supply lowers the equilibrium wage rate of charter school teachers. Figure 1 illustrates.
- If teachers in charter schools had the same experience as those in traditional public schools and if charter school budgets were as big as traditional public school budgets, the charter schools' demand curve for teachers would be $D_{p}$.
- If charter school teachers didn't have the autonomy they enjoy, the supply curve of charter school teachers would be $S_{\rho}$. The equilibrium wage rate would be $\$ 70,000$ per year. (This wage rate is an assumption.)
- The factors discussed above decrease demand to $D_{C}$, increase supply to $S_{C}$, and lower the equilibrium wage rate to $\$ 60,000$ per year.
- The fact that traditional public school teachers are unionized and most charter school teachers are not makes the wage gap even bigger.
- In Fig. 2, the teachers' union at traditional public schools boosts the demand for teachers by lobbying for bigger school budgets and the union restricts the supply of teachers by raising qualification requirements.
- The demand for public school teachers increases from its competitive level, $D_{P}$, to $D_{U}$ and the supply decreases


Figure 1 The Market for Charter School Teachers


Figure 2 The Market for Unionized Public School Teachers
from its competitive level, $S_{P}$, to $S_{U}$. The equilibrium wage rate rises to $\$ 80,000$ per year.

- When a traditional public school is replaced by a charter school, the teachers who remain in the charter school take a very large wage cut.
- The wage cut explains the higher teacher turnover rate of charter schools. The wage cut is too big for some teachers, so they search for an alternative better paying job and quit teaching.


## MATHEMATICAL NOTE

## Present Value and Discounting

## Rent-Versus-Buy Decision

To decide whether to rent an item of capital equipment or to buy the capital and implicitly rent it, a firm must compare the present expenditure on the capital with the future rental cost of the capital.

## Comparing Current and Future Dollars

To compare a present expenditure with a future expenditure, we convert the future expenditure to its "present value."

The present value of a future amount of money is the amount that, if invested today, will grow to be as large as that future amount when the interest that it will earn is taken into account.

So the present value of a future amount of money is smaller than the future amount. The calculation that we use to convert a future amount of money to its present value is called discounting.

The easiest way to understand discounting and present value is to first consider its opposite: How a present value grows to a future amount of money because of compound interest.

## Compound Interest

Compound interest is the interest on an initial investment plus the interest on the interest that the investment has previously earned. Because of compound interest, a present amount of money (a present value) grows into a larger future amount. The future amount is equal to the present amount (present value) plus the interest it will earn in the future. That is,

Future amount $=$ Present value + Interest income.
The interest in the first year is equal to the present value multiplied by the interest rate, $r$, so

Amount after 1 year $=$ Present value + ( $r \times$ Present value).

## or

Amount after 1 year $=$ Present value $\times(1+r)$.

If you invest $\$ 100$ today and the interest rate is 10 percent a year ( $r=0.1$ ), 1 year from today you will have $\$ 110$-the original $\$ 100$ plus $\$ 10$ interest.

Check that the above formula delivers that answer:

$$
\$ 100 \times 1.1=\$ 110
$$

If you leave this $\$ 110$ invested to earn 10 percent during a second year, at the end of that year you will have

Amount after 2 years $=$ Present value $\times(1+r)^{2}$.
With the numbers of the previous example, you invest $\$ 100$ today at an interest rate of 10 percent a year $(r=0.1)$. After 1 year, you will have $\$ 110$-the original $\$ 100$ plus $\$ 10$ interest. And after 2 years, you will have $\$ 121$. In the second year, you earned $\$ 10$ on your initial $\$ 100$ plus $\$ 1$ on the $\$ 10$ interest that you earned in the first year.

Check that the above formula delivers that answer:

$$
\$ 100 \times(1.1)^{2}=\$ 100 \times 1.21=\$ 121
$$

If you leave your $\$ 100$ invested for $n$ years, it will grow to

Amount after $n$ years $=$ Present value $\times(1+r)^{n}$.
With an interest rate of 10 percent a year, your $\$ 100$ will grow to $\$ 195$ after 7 years $(n=7)$-almost double the present value of $\$ 100$.

## Discounting a Future Amount

We have just calculated future amounts 1 year, 2 years, and $n$ years in the future, knowing the present value and the interest rate. To calculate the present value of these future amounts, we just work backward.

To find the present value of an amount 1 year in the future, we divide the future amount by $(1+r)$.

That is,

$$
\text { Present value }=\frac{\begin{array}{c}
\text { Amount of money } \\
\text { one year in future }
\end{array}}{(1+r)} .
$$

Let's check that we can use the present value formula by calculating the present value of $\$ 1101$ year from now when the interest rate is 10 percent a year.

You'll be able to guess that the answer is $\$ 100$ because we just calculated that $\$ 100$ invested today at 10 percent a year becomes $\$ 110$ in one year. So the present value of $\$ 110$ to be received 1 year from today is $\$ 100$. But let's use the formula. Putting the numbers into the above formula, we have

$$
\begin{aligned}
\text { Present value } & =\frac{\$ 110}{(1+0.1)} \\
& =\frac{\$ 110}{1.1}=\$ 100
\end{aligned}
$$

To calculate the present value of an amount of money 2 years in the future, we use the formula:

$$
\text { Present value }=\frac{\begin{array}{l}
\text { Amount of money } \\
\text { two years in future }
\end{array}}{(1+r)^{2}}
$$

Use this formula to calculate the present value of $\$ 121$ to be received 2 years from now at an interest rate of 10 percent a year. With these numbers, the formula gives

$$
\begin{aligned}
\text { Present value } & =\frac{\$ 121}{(1+0.1)^{2}} \\
& =\frac{\$ 121}{(1.1)^{2}} \\
& =\frac{\$ 121}{1.21} \\
& =\$ 100
\end{aligned}
$$

We can calculate the present value of an amount of money $n$ years in the future by using the general formula

$$
\text { Present value }=\frac{\begin{array}{c}
\text { Amount of money } \\
n \text { years in future }
\end{array}}{(1+r)^{n}}
$$

For example, if the interest rate is 10 percent a year, $\$ 100$ to be received 10 years from now has a present value of $\$ 38.55$. That is, if $\$ 38.55$ is invested today at 10 percent a year it will accumulate to $\$ 100$ in 10 years.

## Present Value of a Sequence of Future Amounts

You've seen how to calculate the present value of an amount of money to be received 1 year, 2 years, and $n$ years in the future. Most practical applications of present value calculate the present value of a sequence of future amounts of money that are spread over several years. An airline's payment of rent for the lease of airplanes is an example.

To calculate the present value of a sequence of amounts over several years, we use the formula you have learned and apply it to each year. We then sum the present values for all the years to find the present value of the sequence of amounts.

For example, suppose that a firm expects to pay $\$ 100$ a year for each of the next 5 years and the interest rate is 10 percent a year $(r=0.1)$. The present value ( $P V$ ) of these five payments of $\$ 100$ each is calculated by using the following formula:

$$
P V=\frac{\$ 100}{1.1}+\frac{\$ 100}{1.1^{2}}+\frac{\$ 100}{1.1^{3}}+\frac{\$ 100}{1.1^{4}}+\frac{\$ 100}{1.1^{5}}
$$

which equals

$$
\begin{aligned}
P V & =\$ 90.91+\$ 82.64+\$ 75.13+\$ 68.30+\$ 62.09 \\
& =\$ 379.07 .
\end{aligned}
$$

You can see that the firm pays $\$ 500$ over 5 years. But because the money is paid in the future, it is not worth $\$ 500$ today. Its present value is only $\$ 379.07$. And the farther in the future the money is paid, the smaller is its present value. The $\$ 100$ paid 1 year in the future is worth $\$ 90.91$ today, but the $\$ 100$ paid 5 years in the future is worth only $\$ 62.09$ today.

## The Decision

If this firm could lease a machine for 5 years at $\$ 100$ a year or buy the machine for $\$ 500$, it would jump at leasing. Only if the firm could buy the machine for less than $\$ 379.07$ would it want to buy.

Many personal and business decisions turn on calculations like the one you've just made. A decision to buy or rent an apartment, to buy or lease a car, and to pay off a student loan or let the loan run another year can all be made using the above calculation.

## CHAPTER 19: Economic Inequality

After studying this chapter, you will be able:

- Describe the distributions of income and wealth and the trends in economic inequality in the United States
- Describe the distribution of income and the trends in inequality in selected countries and the world
- Explain the sources of economic inequality and its trends
- Describe the scale of government income redistribution in the United States

When Apple hired Angela Ahrendts to manage its retail stores, it paid her a signing bonus worth $\$ 67$ million. Other top-earning women include Marissa Mayer at Yahoo! and Sheryl Sandberg at Facebook. At the other end of the pay scale are the thousands of fast-food workers calling for a rise in the minimum wage. Why do a few people get enormous incomes, while most struggle to get by on a few dollars an hour? In this chapter, we study economic inequality its extent, its sources, and the things governments do to make it less extreme. In Economics in the News at the end of the chapter, we return to Angela Ahrendts and see why Apple thinks she is worth so much.

## Economic Inequality in the United States

The most commonly used measure of economic inequality is the distribution of annual income. The Census Bureau defines income as money income, which equals market income plus cash payments to households by government. Market income equals wages, interest, rent, and profit earned in factor markets, before paying incometaxes.

## The Distribution of Income

Figure 19.1 shows the distribution of annual income across the 122 million households in the United States in 2012. Note that the $x$-axis measures household money income and the $y$-axis measures the percentage of households.
The most common household income is called the mode income. In 2012, almost 6 percent of the households had incomes of between $\$ 15,000$ and $\$ 20,000$. The value of $\$ 15,000$ marked on the figure is an estimate.
The middle level of household income in 2012, called the median income, was $\$ 51,017$. Fifty percent of households have an income that exceeds the median and fifty percent have an income below the median.
The average household money income in 2012, called the mean income, was $\$ 71,274$. This number equals total household income, about $\$ 8.7$ trillion, divided by the 122 million households. You can see in Fig. 19.1 that the mode is less than the median and that the median is less than the mean. This feature of the distribution of income tells us that there are more households with low incomes than with high incomes. It also tells us that some of the high incomes are veryhigh.
The income distribution in Fig. 19.1 is called a positively skewed distribution, which means that it has a long tail of high values. This distribution contrasts with the bell that describes the
distribution of people's heights. In a bell-shaped distribution, the mean, median, and mode are all equal.
Another way of looking at the distribution of income is to measure the percentage of total income received by each given percentage ofhouseholds.
Data are reported for five groups-called quintiles or fifth shares-each consisting of 20 percent of households.

FIGURE 19.1 The Distribution of Income in the United States in 2012


The distribution of income is positively skewed. The mode (most common) income is less than the median (middle) income, which in turn is less than the mean (average) income. The distribution shown here ends at $\$ 250,000$ because data above that level are not available, but the distribution goes up to several million dollars a year.
Figure 19.2 shows the distribution based on these shares in 2012. The poorest 20 percent of households; received 3.2 percent of total income; the second • poorest 20 percent received 8.3 percent of total income; the middle 20 percent received 14.4 percent of total income; the next highest 20 percent received 23.0 percent of total income; and the highest 20 percent received 51.1 percent of total income.
The distribution of income in Fig. 19.1 and the quintile shares in Fig. 19.2 tell us that income is distributed unequally. But we need a way of comparing the distribution of income in different periods and using different measures. A clever graphical tool called the Lorenz curve enables us to make such comparisons.

FIGURE 19.3 The Income Lorenz Curve in 2012


The cumulative percentage of income is graphed against the cumulative percentage of households. Points $A$ through $E$ on the Lorenz curve correspond to the rows of the table. If incomes were distributed equally, each quintile of households would receive 20 percent of total income and the Lorenz curve would fall along the line of equality. The Lorenz curve shows that income is unequally distributed.
The Income Lorenz Curve
The income Lorenz curve graphs the cumulative percentage of income against the cumulative percentage of households. Figure 19.3 shows the income Lorenz curve using the quintile shares from Fig. 19.2. The table shows the percentage of income of each quintile group. For example, row $A$ tells us that the lowest quintile of households receives 3.2
percent of total income. The table also shows the cumulative percent ages of households and income. For example, row $B$ tells us that the lowest two quintiles (lowest 40 per cent) of households receive 11.5 percent of total income ( 3.2 percent for the lowest quintile plus 8.3 percent for the next lowest).

The Lorenz curve provides a direct visual clue about the degree of income inequality by comparing it with the line of equality. This line, identified inFig. 19.3, shows what the Lorenz curve would be if everyone had the same level of income.
If income were distributed equally across all the households, each quintile would receive 20 percent of total income and the cumulative percentages of income received would equal the cumulative percentages of households, so the Lorenz curve would be the straight line labeled "Line of equality."
The actual distribution of income shown bythe curve labeled "Income Lorenz curve" can be com pared with the line of equality. The closerthe Lorenz curve is to the line of equality, the more equal is the distribution of income.

## The Distribution of Wealth

The distribution of wealth provides another way of measuring economic inequality. A household's wealth is the value of the things that it owns at a point in time. In contrast, income is the amount that the household receives over a given period of time. Figure 19.4 shows the Lorenz curve for wealth in the United States in 2010 (the most recent year for which we have wealth distribution data). The median household wealth in 2010 was $\$ 77,300$. Wealth is extremely unequally distributed, and for this reason, the data are grouped by five unequal groups of house holds. The poorest 25 percent of households have no wealth and the next poorest 25 percent own only 1.1percent of total wealth (row $B^{\prime}$ in the table in Fig. 19.4). Even the third richest 25 percent own only 8.5 percent of total wealth. The richest 25 percent of households own 90.4 percent of wealth, and within this wealthiest group the richest 10 percent of house holds own 74.5 percent of total wealth. Figure 19.4 shows the income Lorenz curve (from Fig. 19.3) alongside the wealth Lorenz curve. You can see that the Lorenz curve for wealth is much farther away from the line of equality than is the Lorenz curve for income, which means that the distribution of wealth is much more unequal than the distribution of income.

## Wealth or Income?

We've seen that wealth is much more unequally distributed than is income. Which distribution provides the better description of the degree of inequality? To answer this question, we need to think about the connection between wealth and income. Wealth is a stock of assets, and income is the flow of earnings that results from the stock of wealth. Suppose that a person owns assets worth $\$ 1$ million-has a wealth of $\$ 1$ million. If the rate of return on assets is 5 percent a year, then this person receives an income of $\$ 50,000$ a year from those assets. We can describe this person's economic condition by using either the wealth of $\$ 1$ million or the income of $\$ 50,000$. When the rate of return is 5 percent a year, $\$ 1$ million of wealth equals $\$ 50,000$ of income in perpetuity. Wealth and income are just different ways of looking at the same thing.

But in Fig. 19.4, the distribution of wealth is more unequal than the distribution of income. Why? It is because the wealth data do not include the value of human capital, while the income data measure income from all wealth, including human capital.
Think about Lee and Peter, two people with equal income and equal wealth. Lee's wealth is human capital and his entire income is from employment. Peter'swealth is in the form of investments in stocks and bonds and his entire income is from these investments.
When a Census Bureau agent interviews Lee and Peter in a national income and wealth survey, their incomes are recorded as being equal, but Lee's wealth is recorded as zero, while Peter's wealth is recorded as the value of his investments. Peter looks vastly more wealthy than Lee in the survey data.
Because the national survey of wealth excludes human capital, the income distribution is a more accurate measure of economic inequality than the wealth distribution.

## Annual or lifetime Income and Wealth?

A typical household's income changes over its life cycle. Income starts out low, grows to a peak when the household's workers reach retirement age, and then falls after retirement. Also, a typical household's wealth changes over time. Like income, it starts out low, grows to a peak at the point of retirement, and falls after retirement.
Think about three households with identical life time incomes, one young, one middleaged, and one retired. The middle-aged household has the highest income and wealth, the retired household has the lowest, and the young household falls in the middle. The distributions of annual income and wealth in agiven year are unequal, but the distributions of life time income and wealth are equal.
The data on inequality share the bias that you've just seen. Inequality in annual income and wealth data overstates lifetime inequality because households are at different stages in their life cycles

## Trends in Inequality

To see trends in the income distribution, we use a measure called the Gini ratio. The Gini ratio equals the area between the Lorenz curve and the line of equality divided by the entire area beneath the line of equality. The more the Lorenz curve bows away from the line of equality, the larger is the Gini ratio. And the larger the Gini ratio, the greater is the degree of income inequality. If income is equally distributed, the Lorenz curve is the same as the line of equality, so the Gini ratio is zero. If most people have an income close to zero and a few people have very large incomes, the Gini ratio is close to 1.
Figure 19.5 shows the U.S. Gini ratio from 1972 to 2012. You can see that the Gini ratio has steadilyincreased, which means that incomes have becomeless equal.

FIGURE 19.4 Lorenz Curves for Income and Wealth


| Households |  |  | Wealth |  |
| :---: | :---: | :---: | :---: | :---: |
|  | Percentage | Cumulative percentage | Percentage | Cumulative percentage |
|  | Lowest 25 | 25 | 0.0 | 0.0 |
| $B^{\prime}$ | Next 25 | 50 | 1.1 | 1.1 |
|  | Next 25 | 75 | 8.5 | 9.6 |
| D | Next 15 | 90 | 15.9 | 25.5 |
| $E^{\prime}$ | Highest 10 | 100 | 74.5 | 100.0 |
| The cumulative percentage of wealth is graphed against the cumulative percentage of households. Points $A^{\prime}$ through $E^{\prime}$ on the Lorenz curve for wealth correspond to the rows of the table. By comparing the Lorenz curves for income and wealth, we can see that wealth is distributed much more unequally than is income. |  |  |  |  |

FIGURE 19.5 The U.S. Gini Ratio:


Measured by the Gini ratio, the distribution of income in the United States became more unequal from 1972 to 2012. The percentage of income earned by the richest households increased through these years. Changes in definitions make the numbers before and after 1992 and before and after 2000 not comparable. Despite the breaks in the data, the trends are still visible.
Education A postgraduate education is the main source of a high income. A person with a professional degree (such as a medical or law degree) earns, on average, $\$ 121,000$-more than double the median income. Just completing high school raises a person's income by more than $\$ 12,000$ a year; and getting a bachelor'sdegree adds another $\$ 30,000$ a year. The average income of people who have not completed 9th grade is $\$ 22,000$ - less than half the median income.

Type of Household Married couples earn more, on average, than people who live alone. A married couple earns about $\$ 76,000$. In contrast, men who live alone earn about $\$ 37,000$, and women who live alone earn only $\$ 26,000$.

Age of Householder Households with the oldest and youngest householders have lower incomes than do those with middle-aged householders. When the house holder is aged between 45 and 54 , household income averages $\$ 66,000$. And when the householder is aged between 35 and 44 , household income averages $\$ 64,000$. When the householder is aged between 15 and 24 , average household income is $\$ 31,000$. And forhouseholders over 65, the average household income is $\$ 34,000$.

## ECONOMICS IN ACTION

## The Rich Get Richer, but School Still Pays

American incomes have been getting steadily more unequal, and 67 percent of Americans told the Gallup poll that they think incomes in 2012 are too unequal. Twenty-one years earlier, in 1991, only 21 percent of Americans thought there were too many people who were too rich.

A key feature of rising inequality is the trend in the incomes of the super rich. Emmanuel Saez of the University of California, Berkeley, used tax returns data to get the numbers graphed in Fig. 1.

After decades of a falling share, starting in 1981, the share of income received by the richest one percent began a steady climb. By 2012 (the latest year in the database), the richest one percent were earning 19.3 percent of the nation's income.

The bottom quintile gets 3.2 percent of total income, so an average household in the top one percent receives more than 120 times the average income of a household in the lowest quintile.

Movie stars, sports stars, and the CEOs of large corporations are among the super rich. People who scratch out a living doing seasonal work on farms earn the lowest incomes. Aside from these extremes, what are the characteristics of the people who earn high


Figure 1 The Income Shares of the Top One Percent
Source of dafa: Alvaredo, Facundo, Anthony B. Alkinson, Thomas Piketty, and Emmanuel Saez, The World Top Incomes Database, htp://g-mond.parisschoolofeconomics.eu/topincomes, 12/08/2012.
incomes and the people who earn low incomes? Figure 2 below answers this question. (The data are for 2012, but the patterns are persistent).


Race and Ethnicity White Americans have an average income of $\$ 57,000$, while black Americans have an average income of $\$ 33,000$. People of Hispanic origin are a bit better off, with an average income of $\$ 39,000$.Asians are best off with an average income of $\$ 69,000$.

Region People who live in the West and Northeastearn more, on average, than people who live in the Midwest and the South. While the region does makea difference, its magnitude is small compared to the dominating effect of education. The bottom line: School pays and so does marriage.

## Poverty

Households at the bottom of the income distribution are so poor that they are considered to be living in poverty. Poverty is a situation in which ahousehold's income is too low to be able to buy thequantities of food, shelter, and clothing that are deemed necessary. This concept of poverty is relative: People are poor compared to what is regarded as normal or average, and so poor that their situation is regarded as a problem that needs attention.
A different poverty concept is absolute poverty - poverty so extreme that it challenges survival. For millions of people living in Africa and Asia, poverty is absolute. The poorest struggle to survive on incomes of less than $\$ 400$ a year.

Official U.S. Poverty Measure In the United States, the poverty level is calculated each year by the Social Security Administration. In 2012, the poverty level for a four-person household was an income of $\$ 23,492$. Other income levels apply to households of different sizes.

Amount of Poverty About 46 million Americans- 15 percent of the population- live in households that have incomes below the poverty level. The incidence of poverty varies by race, age, work experience, physical ability, and household status. We'll look at each of these influences on poverty.

Race The poverty rate among white Americans is 13 percent compared to 26 percent for Hispanic origin Americans and 27 percent of black Americans.
But the absolute number of white Americans in poverty, 31 million, is greater than the number of minority households- 11 million black and 14 million Hispanic.
Age Poverty rates are 22 percent for children, 14 per cent for people aged 18 to 64 , and 9 percent for seniors aged 65 and over.

Work Experience Americans with a job, either full time or part time, are much less likely to be living inpoverty than those without a job. The poverty rate for those with jobs is 7 percent and for those withoutjobs is 33 percent.

Physical Ability The poverty rate among people with disabilities is very high-29 percent. These people often have no jobs and no easy access to care.

Household Status More than 28 percent of households in which the householder is a female with no husband present have incomes below the povertylevel.

The overall poverty rate and the averages for the groups we've just examined have increased during thepast few years following a major economic recession. But aside from this recent upturn in the poverty rate, it has been remarkably constant.

## Inequality in the World Economy

Which countries have the greatest economic inequality and which have the least and the greatest equality?Where does the United States rank? Is it one of the most equal or most unequal, or is it somewhere in
the middle? And how much inequality is there in the world as a whole when we consider the entire world as a single global economy?
We'll answer these questions by first looking at the income distribution in a selection of countries and then by examining features of the global distribution of income.

## Income Distribution in Selected Countries

By inspecting the income distribution data for every country, we can compare the degree of income inequality and identify the countries with the most inequality and those with the least inequality.
Figure 19.6 summarizes some extremes and shows where the United States lies in the range of degrees of income inequality.
Look first at the numbers in the table. They tell us that in Brazil and South Africa, the lowest quintile of households receive only 2 percent of total incomewhile the highest quintile receives 65 percent of totalincome. An average person in the highest quintile receives 32.5 times the income of an average person inthe lowest quintile.
Contrast these numbers with those for Finland and Sweden. In these countries, the lowest quintile receives 8 percent of total income and the highest quintile receives 35 percent. So an average person in the highest quintile receives 4.4 times the income of an average person in the lowest quintile. The numbers for the United States lie between these extremes with an average person in the highest quintile receiving just under 10 times the amount received by an average person in the lowest quintile.
Brazil and South Africa are extremes not matched in any other major country or region. Inequality is large in these countries because they have a relatively small but rich European population and a large and relatively poor indigenous population. Finland and Sweden are extremes, but they are not unusual. Income distributions similar to these are found in many European countries in which governments pursue aggressive income redistribution policies. We look next at the global income distribution.

## Global Inequality and Its trends

The global distribution of income is much more unequal than the distribution within any one country. The reason is that many countries, especially inAfrica and Asia, are in a preindustrial stage of economic development and are poor, while industrialcountries such as the United States are rich. Whenwe look at the distribution of income across theentire world population that goes from the low income of the poorest African to the high income of the richest American, we observe a very large degree of inequality.

## FIGURE 19.6 Lorenz Curves Compared



|  | Percentage of total income ${ }^{1}$ |  |  |
| :--- | :---: | :---: | :---: |
| Households <br> (quintile) | Brazil and <br> South Africa | United <br> States | Finland and <br> Sweden |
| Lowest | 2 | 5 | 8 |
| Second | 5 | 11 | 14 |
| Third | 10 | 16 | 20 |
| Fourth | 18 | 24 | 23 |
| Highest | 65 | 44 | 35 |

The table shows the percentages of total income received by each quintile. The figure shows the cumulative percentage of income graphed against the cumulative percentage of households. The data and the Lorenz curves show that income is distributed most unequally in Brazil and South Africa and least unequally in Finland and Sweden. The degree of income inequality in the United States lies between these extremes.
To put some raw numbers on this inequality, start with the poorest. Measured in the value of the U.S. dollar in 2005, a total of 3 billion people or 50 per cent of the world population
live on $\$ 2.50$ a day or less. Another 2 billion people or 30 percent of the world population live on more than $\$ 2.50$ but less the world population live on $\$ 10$ a day or less. In contrast, in the rich United States, the average person has an income of $\$ 115$ per day and an average person in the highest income quintile has an income of $\$ 460$ a day. So the average American earns 46 times the income of one of the world's 3 billion poorest people and more than 11.5 times the income of 80 percent of the people who live in developing economies. An American with the average income in the highest quintile earns 184 times that of the world's poorest people but only 16 times that of an average bottom quintile American.

World Gini Ratio Wecan compare world inequality with U.S. inequality by comparing Gini ratios. You saw that the U.S. Gini ratio in 2012 was about 0.47 . The world Gini ratio is about 0.61. Recalling the interpretation of the Gini ratio in terms of the Lorenz curve, the world Lorenz curve lies much farther from the line of equality than does the U.S. Lorenz curve.

World Trend You saw (in Fig. 19.5 on p. 487) that incomes have become more unequal in the United States-the Gini ratio has increased. The same trendsare found in most economies. Increased income inequality is a big issue in two of the world'slargest and poor nations, China and India. In these two economies, urban middle classes are getting richer at a faster pace than the rural farmers.
Despite greater inequality within countries, the world is becoming less unequal. Figure 19.7 shows this trend toward less inequality as measured by the world Gini ratio. How can the world income distribution become less unequal while individual countries become more unequal? The answer is that average incomes in poorer countries are rising much faster than average incomes in rich countries. While the gap between rich and poor is widening within countries, it is narrowing across countries.

## The Sources of Economic Inequality

We've described some key facts about economic inequality and its trends and our task now is to explain those facts. We began this task in Chapter 18 by learning about the forces that influence demand and supply in the markets for labor, capital, and land. We're now going to deepen our understanding of these forces.
Inequality arises from unequal labor market outcomes and from unequal ownership of capital. We'll begin by looking at labor markets and three features of them that contribute to differences in income:

- Human capital
- Discrimination
- Contests among superstars
figure 19.7 The World Gini Ratio:
1970-2005


Measured by the Gini ratio, the distribution of income in the entire world became more equal between 1970 and 2005.

## Human Capital

A clerk in a law firm earns less than a tenth of the amount earned by the attorney he assists. An operating room assistant earns less than a tenth of the amount earned by the surgeon with whom she works. A bank teller earns less than a tenth of the amount earned by the bank's CEO. Some of the differences in these earnings arise from differences in human capital. To see the influence of human capital on labor incomes, consider the example of a law clerk and the attorney he assists. (The same reasoning can be applied to an operating room assistant and surgeon, or a bank teller and bank CEO.)

Demand, Supply, and Wage Rates An attorney performs many tasks that a law clerk cannot perform. Imagine an untrained law clerk cross-examining a witness in a complicated trial. The tasks that the attorney performs are valued highly by her clients who willingly pay for her services. Using a term thatyou learned in Chapter 18, an attorney has a high value of marginal product, and a higher value of marginal product than her law clerk. But you also learned in Chapter 18 that the value of marginal product of labor determines (is the same as) the demand for labor. So, because an attorney has a higher value of marginal product, the demand for attorney services is greater than the demand for law-clerkservices.
To become an attorney, a person must acquire human capital. But human capital is costly to acquire. This cost-an opportunity cost includes expenditures on tuition and textbooks. It also includes forgone earnings during the years spent incollege and law
school. It might also include lowearnings doing on-the-job training in a law officeduring the summer.
Because the human capital needed to supply attorney services is costly to acquire, a person's willingnessto supply these services reflects this cost. The supply of attorney services is smaller than the supply of law-clerk services.
The demand for and supply of each type of labor determine the wage rate that each type of labor earns. Attorneys earn a higher wage rate than law clerks because the demand for attorneys is greater and the supply of attorneys is smaller. The gap between the wage rates reflects the higher value of marginal product of an attorney (demand) and the cost of acquiring human capital (supply).

Trends in Wage Inequality You've seen that high income households have earned an increasing share of total income while low-income households have earned a decreasing share: The distribution of income in the United States has become more unequal. Technological change and globalization are two possible sources of this increasedinequality.

Technological Change Information technologies such as computers and laser scanners are substitutes in production for low-skilled labor: They perform tasks that previously were performed by low-skilled labor. The introduction of these technologies has lowered the marginalproduct and the demand for low-skilled labor. These same technologies require highskilled labor to design, program, and run them. High-skilled labor and the information technologies are complements in production. So these new technologies have increased the marginal product and demand for high-skilled labor.
Figure 19.8 illustrates the effects of information technologies on wages and employment. The supply of low-skilled labor in part (a) and that of high-skilled labor in part (b) are $S$, and initially, the demand in each market is $D 0$. The low-skill wage rate is $\$ 5$ an hour, and the high-skill wage rate is $\$ 10$ an hour. The demand for low-skilled labor decreases to Dr in part (a) and the demand for high-skilled labor increases to $D r$ in part (b). The low-skill wage rate falls to $\$ 4$ anhour and the high-skill wage rate rises to $\$ 15$ an hour.

Globalization The entry of China and other developing countries into the global economy has lowered the prices of many manufactured goods. Lower prices for the firm's output lowers the value of marginal product of the firm's workers and decrease the demand for their labor. A situation like that in Fig. 19.S(a) occurs. The wage rate falls, and employment shrinks. At the same time, the growing global economy increases the demand for services that employ high skilled workers, and the value of marginal product and the demand for highskilled labor increases. A situation like that in Fig. 19.S(b) occurs. The wage rate rises, and employment opportunities for high skilled workers expand.

## Discrimination

Human capital differences can explain some of the economic inequality that we observe. Discrimination;', is another possible source of inequality. Suppose that black females and white males have identical abilities as investment advisors. If everyone is free of race and sex prejudice, the market determines the same wage rate for both groups. But ifcustomers
are prejudiced against women and minorities/ this prejudice is reflected in the wage rates-whitemen earn more than black women.

Counteracting Forces Economists disagree about whether prejudice actually causes wage differentials, and one line of reasoning implies that it does not. In the above example, customers who buy from white men pay a higher service charge for investment advice than do the customers who buy from black women. This price difference acts as an incentive to encourage people who are prejudiced to buy from the people against whom they are prejudiced. This force could be strong enough to eliminate the effects of discrimination altogether. Suppose, as is true in manufacturing, that a firm's customers never meet its workers. If such a firm discriminates against women or minorities, it can't compete with :, firms who hire these groups because its costs are higher than those of the nonprejudiced firms. Only firms that do not discriminate survive in a competitive industry.
Whether because of discrimination or from some other source, women and visible minorities do earn lower incomes than white males. Another possible source of lower wage rates of women arises from differences in the relative degree of specialization of women and men.

Differences in the Degree of Specialization Couples must choose how to allocate their time between working for a wage and doing jobs in the home, suchas cooking, cleaning, shopping, organizing vacations, and, most important, bearing and raising children.
Let's look at the choices of Bob and Sue. Bob might specialize in earning an income and Sue in taking care of the home. Or Sue might specialize in earning an income and Bob in taking care of the home. Or both of them might earn an income and share home production jobs.
The allocation they choose depends on their preferences and on their earning potential. The choice ofan increasing number of households is for each person to diversify between earning an income and doing some household chores. But in most households, Bob will specialize in earning an income andSue will both earn an income and bear a larger share of the task of running the home. With this allocation, Bob will probably earn more than Sue. IfSue devotes time and effort to ensuring Bob's mental andphysical well-being, the quality of Bob's market laborwill be higher than it would be if he were diversified.If the roles were reversed, Sue would be able to supply market labor that earns more than Bob's.
Totest whether the degree of specialization accounts for earnings differences between the sexes,economists have compared the incomes of never married men and women. They have found that, onthe average, with equal amounts of human capital, the wages of these two groups are the same.

FIGURE 19.8 Explaining the Trend in Income Distribution

(a) A decrease in demand for low-skilled labor

(b) An increase in demand for high-skilled labor

Low-skilled labor and information technologies are substitutes in production. In part (a), advances in information technology decrease the demand for low-skilled labor and lower its wage rate. High-skilled labor and information technologies are complements in production. In part (b), advances in information technology increase the demand for high-skilled labor and raise its wage rate.

## Contests Among Superstars

The differences in income that arise from differences in human capital are important and affect a large proportion of the population. But human capital differences can't account for some of the really large income differences.
The super-rich - those in the top one percent of the income distribution whose income share has been rising-earn vastly more than can be explained by human capital differences. What makes a person super rich?
A clue to the answer is provided by thinking about the super-rich in tennis and golf. What makes tennis players and golfers special is that their earnings depend on where they finish in a tournament.
When Rafael Nadal won the U.S. Open Tennis Championship in 2013, he received $\$ 2,600,000$. The runner-up in this event, Novak Djokovic received $\$ 1,300,000$. So Rafa earned double the amount earned by Novak. And he earned 81 times the amount received by the players who lost in the first round of the tournament.
It is true that Rafael Nadal has a lot of human capital. He practices hard and long and is a remark able athlete. But anyone who is good enough to get into a tennis Grand Slam tournament is similarlywell equipped with human capital and has spent a similar number of long hours in training and practice. It isn't human capital that explains the differences in earnings. It is the tournament and the prize differences that accounts for the large differences in earnings.
Three questions jump out: First, why do we reward superstar tennis players (and golfers) with prizes for winning a contest? Second, why are the prizes so different? And third, do the principles that apply on the tennis court (and golf course) apply more generally to corporations?

Why Prizes for a Contest? The answer to this question (which was noted in Chapter 5, p. 144) is that contests with prizes do a good job of allocating scarce resources efficiently when the efforts of the participants are hard to monitor and reward directly. There is only one winner, but many people work hard in an attempt to be that person. So a great deal of diligent effort is induced by a contest.

Why Are Prizes S Different? The prizes need to be substantially different to induce enough effort. If thewinner received 10 percent more than the runner up, the gain from being the winner would be insufficient to encourage anyone to work hard enough. Someone would win but no one would put in much effort.
Tennis matches would be boring, golf scores wouldbe high, and no one would be willing to pay to see these sports. Big differences are necessary to induce a big enough effort to generate the quality of performance that people are willing to pay to see.

Does the Principle Apply More Generally? Winner takes-all isn't confined to tennis and golf. Movie stars; superstars in baseball, basketball, football, and ice hockey; and top corporate executives can all be viewed as participants in contests that decide the winners. The prize for the winner is an income at least double that of the runner up and many multiples of the incomes of those who drop out earlier in the tournament.

Do Contests Among Superstars Explain the Trend? Contests among superstars can explain large differences in incomes. But can contests explain the trendtoward greater inequality with an increasing share of total income going to the super rich as shown on page 488? An idea first suggested by University of Chicago economist Sherwin Rosen suggests that a winner takes-all contest can explain the trend. The key is that globalization has increased the market reach of the winner and increased the spread between the winner and the runners-up.
Global television audiences now watch all the world's major sporting events and the total revenue generated by advertising spots during these events has increased. Competition among networks and cable and satellite television distributors has increased the fees that event organizers receive. And to attract the top star performers, prize money has increased and the winner gets the biggest share of the prize pot.
So the prizes in sports have become bigger and the share of income going to the "winner" has increased.
A similar story can be told about superstars and the super-rich in business. As the cost of doing business on a global scale has fallen, more and more corporations have become global in their reach. Not only are large multinational corporations sourcing their inputs from far afield and selling in every country, they are also recruiting their top executives from a global talent pool. With a larger source of talent, and larger total revenue, firms must make the "prize" - the reward for the top job- more attractive to compete for the best managers.
We've examined some sources of inequality in the labor market. Let's now look at the way inequality arises from unequal ownership of capital.

## Unequal Wealth

You've seen that wealth inequality-excluding human capital-is much greater than income inequality. This greater wealth inequality arises from two sources: life-cycle saving patterns and transfers of wealth from one generation to the next.

Life-Cycle Saving Patterns Over a family's life cycle, wealth starts out at zero or perhaps less than zero. Astudent who has financed education all the way through graduate school might have lots of human capital and an outstanding student loan of $\$ 60,000$. This person has negative wealth. Gradually loans get paid off and a retirement fund is accumulated. At the point of retiring from full-time work, the family has maximum wealth. Then, during its retirement years the family spends its wealth. This life-cycle pattern means that much of the wealth is owned by people in theirsixties.

Intergenerational Transfers Some households inherit wealth from the previous generation. Some save more than enough on which to live during retirement and transfer wealth to the next generation. But these intergenerational transfers of wealth do not always increase wealth inequality. If a generation that has a high income saves a large part of that income and leaves wealth to a succeeding generation that has a lower income, this transfer decreases the degree of inequality. But one feature of intergenerational transfers of wealth leads to increased inequality: wealth concentration through marriage.

Marriage and Wealth Concentration People tend to marry within their own socioeconomic class-a phenomenon called assortative mating. In everyday language, "like attracts like." Although there is a good deal of folklore that "opposites attract," perhaps such Cinderella tales appeal to us because they are so rare in reality. Wealthy people seek wealthy partners. Because of assortative mating, wealth becomes more concentrated in a small number of families and $\backslash$ the distribution of wealth becomes more unequal.

## Income Redistribution

The three main ways in which governments in the United States redistribute income are

- Income taxes
- Income maintenance programs
- Subsidized services


## Income Taxes

Income taxes may be progressive, regressive, or proportional. A progressive income tax is one that taxes income at an average rate that increases as income increases. A regressive income tax is one that taxes income at an average rate that decreases as income increases. A proportional income tax (also called flat rate income tax) is one that taxes income at a constantrate, regardless of the level of income.
The income tax rates that apply in the United States are composed of two parts: federal and state taxes. Some cities, such as New York City, also have an income tax. There is variety in the detailed tax arrangements in the individual states, but the tax system, at both the federal and state levels, is progressive. The poorest working households receive money from the government through an earned income tax credit. Successively higher-income households pay 10 percent, 15 percent, 25 percent, 28 percent, 33 percent, and 35 percent of each additional dollar earned.

## Income Maintenance Programs

Three main types of programs redistribute income by making direct payments (in cash, services, or vouchers) to people in the lower part of the income distribution. They are

- Social Security programs
- Unemployment compensation
- Welfare programs

Social Security Programs The main Social Security program is OASDHI-Old Age, Survivors, Disability, and Health Insurance. Monthly cash payments to retired or disabled workers or their surviving spouses and children are paid for by compulsory payroll taxes on both employers and employees.
In 2013, total Social Security expenditure was budgeted at $\$ 1.3$ trillion, and the average monthly Social Security check was $\$ 1,230$ or $\$ 14,760$ per year.
The other component of Social Security is Medicare, which provides hospital and health insurance for the elderly and disabled.

Unemployment Compensation To provide an income to unemployed workers, every state has established an unemployment compensation program. Under these programs, a tax is paid that is based on the income of each covered worker and such a worker receives a benefit when he or she becomes unemployed. The details of the benefits vary from state to state.

Welfare Programs The purpose of welfare is to pro vide incomes for people who do not qualify for Social Security or unemployment compensation. They are

1. Supplementary Security Income (SSI) program, designed to help the neediest elderly, disabled, and blind people
2. Temporary Assistance for Needy Households (TANF) program, designed to help households that have inadequate financial resources
3. Food Stamp program, designed to help the poorest households obtain a basic diet
4. Medicaid, designed to cover the costs of medical care for households receiving help under the SSI and TANF programs

## Subsidized services

A great deal of redistribution takes place in the United States through the provision of subsidized services - services provided by the government at prices below the cost of production. The taxpayers who consume these goods and services receive a transfer in kind from the taxpayers who do not consume them. The two most important areas in which this form of redistribution takes place are healthcare and education both kindergarten through grade 12 and college and university.
In 2013-2014, students enrolled in the University of California system paid annual tuition and fees of $\$ 13,200$. The cost of providing a year's education at the University of California was probably about $\$ 40,000$. So a household with just one member enrolled in one of these institutions received a benefit from the government of $\$ 26,800$ a year.
A dollar collected from a rich person does not translate into a dollar received by a poor person. Some of it gets used up in the process of redistribution. Tax-collecting agencies such as the Internal Revenue Service and welfare-administering agencies (as well as tax accountants and lawyers) use skilled labor, computers, and other scarce resources to do their work. The bigger the scale of redistribution, the greater is the opportunity cost of administering it.
But the cost of collecting taxes and making welfare payments is a small part of the total cost of redistribution. A bigger cost arises from the inefficiency(deadweight loss) of taxes and benefits. Greater equality can be achieved only by taxing productive activities such as work and saving. Taxing people's income from their work and saving lowers the after-tax income they receive. This lower after-tax income makes them work and save less, which in turn resultsin smaller output and less consumption not only for the rich who pay the taxes but also for the poor whoreceive the benefits.
It is not only taxpayers who face weaker incentives to work. Benefit recipients also face weaker incentives. In fact, under the welfare arrangements that prevailed before the 1996 reforms, house- holds that benefited from welfare faced the weakest incentives to work. When a welfare recipient got a job, benefits were withdrawn and eligibility for programs
such as Medicaid ended, so the household in effect paid a tax of more than 100 percent on its earnings. This arrangement locked poor householdsin a welfare trap.
So the agencies that determine the scale and methods of income redistribution must pay close attention to the incentive effects of taxes and benefits. Let's close this chapter by looking at one way in which lawmakers are tackling the big tradeofftoday.

A Major Welfare Challenge Young women who have not completed high school, have a child (or children), live without a partner, and more likely are black or Hispanic are among the poorest people in the United States today. They and their children present a major welfare challenge.
First, their numbers are large. In 2012, there were 15.5 million single-mother families, 5 mil lion of which were in poverty. In 2009 (the most recent year with census data), single mothers were owed $\$ 35$ billion in child support and 30 percent of the women received no support from theirchildren's fathers. The long-term solution to the problem these women face is education and job training-acquiringhuman capital. The short-term solutions are enforcing child support payments by absent fathers and former spouses and providing welfare.

Welfare must be designed to minimize the disincentive to pursue the long-term goal of becoming self-supporting. The current welfare program in the United States tries to walk this fine line.
Passed in 1996, the Personal Responsibility and Work Opportunities Reconciliation Act strengthened the Office of Child Support Enforcement and increased the penalties for nonpayment of support. The act also created the Temporary Assistance for Needy Households (TANF) program. TANF is a block grant paid to the states, which administer payments to individuals. It is not an open-ended entitlement program. An adult member of a household thatis receiving assistance must either work or perform community service, and there is a fiveyear limit forassistance.
We've examined economic inequality in the United States. We've seen how inequality arises and that inequality has been increasing. Economics in the News on pp. 500-501 looks at the increasing inequality that began during the early 1980s and continues to day.

## ECONOMICS IN ACTION

## Income Redistribution: Only the Richest Pay

A household's market income tells us what a household earns in the absence of government redistribution. You've seen that market income is not the official basis for measuring the distribution of income that we've used in this chapter. The Census Bureau's measure is money income (market income plus cash transfers from the government). But market income is the correct starting point for measuring the scale of income redistribution.

We begin with market income and then subtract taxes and add the amounts received in benefits. The result is the distribution of income after taxes and benefits. The data available on benefits exclude the value of subsidized services such as college, so the resulting distribution might understate the total amount of redistribution from the rich to the poor.

The figures show the scale of redistribution in 2001, the most recent year for which the Census Bureau has provided these data. In Fig. 1, the blue Lorenz curve describes the market distribution of income and the green Lorenz curve shows the distribution of income after all taxes and benefits, including Medicaid and Medicare benefits. (The Lorenz curve based on money income in Fig. 19.3 lies between these two curves.)

The distribution after taxes and benefits is less unequal than is the market distribution. The lowest quintile of households received only 0.7 percent of market income but 4.6 percent of income after taxes and benefits. The highest quintile of households received 54 percent of market income, but only 44.4 percent of income after taxes and benefits.

Figure 2 highlights the percentage of total income redistributed among the five groups. The share of total income received by the lowest 60 percent of households increased. The share received by the fourth quintile barely changed, but the share received by the highest quintile fell by 9.6 percent.

Government provision of healthcare services has grown to the scale of private provision. Programs such as Medicaid and Medicare bring high-quality and high-cost healthcare to millions of people who earn too little to buy such services themselves.


Figure 1 Income Distribution Before and After Redistribution


Figure 2 The Scale of Redistribution
Source of dala: U.S. Bureau of the Census, Current Population Survey Arrnual Social and Economic Supplement, Effect of Benefits and Taxes on Income and Poverty, Table 1 Income Distribution Measures, by Definition of Income: 2009.

## The Big Tradeoff

The redistribution of income creates what has been called the big tradeoff, a tradeoff between equity and efficiency. The big tradeoff arises because redistribution uses scarce resources and weakens incentives.

## ECONOMIC ANALYSIS

- The news article reports the extremely high incomes received by two top female executives. These high incomes are not rare today and have been getting less rare.
- Economists Thomas Piketly of l'Ecole d'économie de Paris - Paris School of Economics and Emmanuel Saez of U.C. Berkeley examined the tax returns of the super rich and found the trend shown in Fig. 1.
- Figure 1 shows the income share (percentage of total income) received by the top 0.01 percent of the population.
- The top 0.01 percent includes the top corporate executives and in 2008 was made up of 15,246 families with incomes that exceeded $\$ 9,141,000$.
- The average family in the top 0.01 percent received 296 times the income of the average family in the bottom quintile. This ratio was 27 in 1965. These ratios are in line with the trend reported in the news article.
- Why are executives like Angela Ahrendts paid so much? And why have top incomes risen so much? Isn't there an abundance of talent around? Could not and should not these people be paid much less?
- It is true that there is an abundance of talent. Globatization has made the entire world the talent pool that large corporations tap for their top executive spots.
- But it is because of the abundance of talent that executive pay has become so high.
- You saw on p. 495 that we can view top executives as the winners of a contest among potential superstars.
- Contests induce high effort and productivity from managers at all levels as they compete for the top job.
- How hard people compete (how productive they are) depends on the size of the prize and the probability of winning it.
- You can think of the contest in terms of the pyramids in Fig. 2. The talent pool is the base of the pyramid and the contest delivers a winner who gets to the top.
- When the talent pool is small, as it was in 1965, the chance of being the winner is large enough for a moderate prize to induce enough effort.
- When the talent pool is large, as it is today, the chance of being the winner is very small, so to induce the same amount of effort, the prize is very large.


Figure 1 Income Share of the Top 0.01 Percent


1960s


## Figure 2 Bigger Pyramid Means Bigger Prizes for Getting to Top

Source of data: Figure 1, Emmanuel Saez, "Striking it Richer: The Evolution of Top Incomes in the United States," Pathways, Winter 2008 issue.

- This superstar contest idea explains the direction of change-why top executives' earnings have increased relative to the average wage.
- It also explains why Angela Ahrendts and Marissa Mayer earn such enormous incomes.


## CHAPTER 20: Uncertainty and information

After studying this chapter, you will be able to:

- Explain how people make decisions when they are uncertain about the consequences
- Explain how markets enable people to buy and sell risk
- Explain how markets cope when buyers and sellers have private information
- Explain how uncertainty and incomplete information influence the efficiency of markets
You want cm A in your economics course you want to graduate with a high GPA. But what if grades are inflated and you along with everyone else gets an A? What happens in the job market for college graduates? And how do accurate grades help the market work? The job market for college graduates is an example of a market with uncertainty and incomplete information. Can this market achieve an efficient outcome?
This chapter answers these questions. In Economics in the News at the end of the chapter, you will see why grade inflation is not in your best interest.


## Decisions in the Face of Uncertainty

Tania, a student, is trying to decide which of two summer jobs to take. She can work as a house painter and earn enough to save $\$ 2,000$ by the end of the summer. There is no uncertainty about the income from this job. If Tania takes it, she will definitely have $\$ 2,000$ in her bank account at the end of the summer. The other job, workingas a telemarketer selling subscriptions to a magazine, is risky. If Tania takes this job, her bank balance at the end of the summer will depend on her success at selling. She will earn enough to save $\$ 5,000$ if she is successful but only $\$ 1,000$ if sheturns out to be a poor salesperson. Tania has nevertried selling, so she doesn't know how successful she'll be. But some of her friends have done this job, and 50 percent of them do well and 50 percent do poorly. Basing her expectations on this experience, Tania thinks there is a 50 percent chance that she will earn $\$ 5,000$ and a 50 percent chance that she will earn $\$ 1,000$. Tania is equally as happy to paint as she is to make phone calls. She cares only about the money. Which job does she prefer: the one that provides her with $\$ 2,000$ for sure or the one that offers her a 50 per- cent chance of making $\$ 5,000$ but a 50 percent risk of making only \$1,000?
To answer this question, we need a way of comparing the two outcomes. One comparison is the expected wealth that each job creates.

## Expected Wealth

Expected wealth is the money value of what a person expects to own at a given point in time. An expectation is an average calculated by using a formula that weights each possible outcome with the probability (chance) that it will occur.
For Tania, the probability that she will have
$\$ 5,000$ is 0.5 (a 50 percent chance). The probability that she will have $\$ 1,000$ is also 0.5 . Notice that theprobabilities sum to 1 . Using these numbers, we can calculate Tania's expected wealth, EW, which is
$E W=(\$ 5,000 \times 0.5)+(\$ 1,000 \times 0.5)=\$ 3,000$.
Notice that expected wealth decreases if the risk of a poor outcome increases. For example, if Tania has a20 percent chance of success (and 80 percent chance of failure), her expected wealth falls to $\$ 1,800-(\$ 5,000 \times 0.2)+(\$ 1,000 \times 0.8)=\$ 1,800$.
Tania can now compare the expected wealth from : each job- $\$ 3,000$ for the risky job and $\$ 2,000$ for the non-risky job.
So does Tania prefer the risky job because it gives ;her a greater expected wealth? The answer is we don't know because we don't know how much Tania dis- :likes risk.

## Risk Aversion

Risk aversion is the dislike of risk. Almost everyone is; risk averse but some more than others. In football, running is less risky than passing. Coach John Harbaugh of the Baltimore Ravens, who favors a cautious running game, is risk averse. Denver quarterback Peyton Manning, who favors a risky passing, game, is less risk averse. But almost everyone is risk averse to some degree.
We can measure the degree of risk aversion by the compensation needed to make a given amount of risk acceptable. Returning to Tania: If she needs to bepaid more than $\$ 1,000$ to take on the risk arising from the telemarketing job, she will choose the safe painting job and take the $\$ 2,000$ non-risky income\}But if she thinks that the extra $\$ 1,000$ of expected income is enough to compensate her for the risk, she! will take the risky job.
To make this idea concrete, we need a way of thinking about how a person values different levels wealth. The concept that we use is utility. We apply :the same idea that explains how people make expenditure decisions (see Chapter 8) to explain risk aversion and decisions in the face of risk.

## Utility of Wealth

Wealth (money in the bank and other assets of value); is like all good things. It yields utility. The more wealth a person has, the greater is that person's total utility. But each additional dollar of wealth brings a diminishing increment in total utility - the marginal utility of wealth diminishes as wealth increases. Diminishing marginal utility of wealth means than the gain in utility from an increase in wealth is smaller than the loss in utility from an equal decrease in wealth. Stated differently, the pain from a loss is greater than the pleasurefrom again ofequalsize. Figure 20.1 illustrates Tania's utility of wealth. Each point $A$ through $F$ on Tania's utility of wealth curve corresponds to the value identified by the same letter in the table. For example, at point $C$, Tania's wealth is $\$ 2,000$, and her total utility is 70 units. If Tania's wealth increases, her total utility increases and her marginal utility decreases. Her marginal utility is 25 units when wealth increases from $\$ 1,000$ to $\$ 2,000$, but only 13 units when wealth increases from $\$ 2,000$ to $\$ 3,000$.
We can use a person's utility of wealth curve to calculate expected utility and the cost of risk.

FIGURE 20.1 The Utility of Wealth


The table shows Tania's utility of wealth schedule, and the figure shows her utility of wealth curve. Utility increases as wealth increases, but the marginal utility of wealth diminishes.

## Expected Utility

Expected utility is the utility value of what a person expects to own at a given point in time.
Like expected wealth, it is calculated by using a formulathat weights each possible outcome with the probability that it will occur. But it is the utility outcome, not the money outcome, that is used to calculate expected utility.
Figure 20.2 illustrates the calculation for Tania.
Wealth of $\$ 5,000$ gives 95 units of utility and wealth of $\$ 1,000$ gives 45 units of utility. Each outcome has a probability of 0.5 (a 50 percent chance). Using these numbers, we can calculate Tania's expected utility, $E U$, which is

$$
E U=(95 \times 0.5)+(45 \times 0.5)=70 .
$$

FIGURE 20.2 Expected Utility


Tania has a 50 percent chance of having $\$ 5,000$ of wealth and a total utility of 95 units. She also has a 50 percent chance of having $\$ 1,000$ of wealth and a total utility of 45 units. Tania's expected wealth is $\$ 3,000$ (the average of $\$ 5,000$ and $\$ 1,000$ ) and her expected utility is 70 units (the average of 95 and 45 ). With a wealth of $\$ 3,000$ and no uncertainty, Tania's total utility is 83 units. For a given expected wealth, the greater the range of uncertainty, the smaller is expected utility.
Expected utility decreases if the risk of a poor out- come increases. For example, if Tania has a 20 per-cent chance of success (and an 80 percent chance offailure), her expected utility is 55 units- $(95 \times 0.2)+(45 \times 0.8)=55$.
Notice how the range of uncertainty affects expected utility. Figure 20.2 shows that with $\$ 3,000$ of wealth and no uncertainty, total utility is 83 units. But with the same expected wealth and Tania's uncertainty-a 50 percent chance of having $\$ 5,000$ and a 50 percent chance of having $\$ 1,000$-expected utility is only 70 units. Tania's uncertainty lowers her expected utility by 13 units.
Expected utility combines expected wealth and risk into a single index.

## Making a Choice with Uncertainty

Faced with uncertainty, a person chooses the action that maximizes expected utility. To select the job that gives her the maximum expected utility, Tania must:

1. Calculate the expected utility from the risky telemarketing job
2. Calculate the expected utility from the safe painting job
3. Compare the two expected utilities

Figure 20.3 illustrates the calculations. You've just seen that the risky telemarketing job gives Tania an expected utility of 70 units. The safe painting job also gives Tania a utility of 70. That is, the total utility of $\$ 2,000$ with no risk is 70 units. So with either job, Tania has an expected utility of 70 units. She is indifferent between these two jobs.
If Tania had only a 20 percent chance of success and an 80 percent chance of failure in the telemarketing job, her expected utility would be 55 (calculatedabove). In this case, she would take the painting job and get 70 units of utility. But if the probabilities were reversed and she had an 80 percent chance of success and only a 20 percent chance of failure in the telemarketing job, her expected utility would be 85 units-( $95 \times 0.8$ ) $+(45 \mathrm{X}$ $0.2)=85$. Inthis case, she would take the risky telemarketing job.
We can calculate the cost of risk by comparing the expected wealth in a given risky situation with the wealth that gives the same total utility but no risk.
Using this principle, we can find Tania's cost of bearing the risk that arises from the telemarketing job. That cost, highlighted in Fig. 20.3, is $\$ 1,000$.

FIGURE 20.3 Choice Under Uncertainty


With a 50 percent chance of having $\$ 5,000$ of wealth and a 50 percent chance of having $\$ 1,000$ of wealth, Tania's expected wealth is $\$ 3,000$ and her expected utility is 70 units. Tania would have the same 70 units of total utility with wealth of $\$ 2,000$ and no risk, so Tania's cost of bearing this risk is $\$ 1,000$. Tania is indifferent between the job that pays $\$ 2,000$ with no risk and the job that offers an equal chance of $\$ 5,000$ and $\$ 1,000$.

## Buying and Selling Risk

You've seen at many points in your study of markets how both buyers and sellers gain from trade. Buyers gain because they value what they buy more highly than the price they must pay-they receive a consumer surplus. And sellers gain because they face costs that are less than the price at which they can sell - they receive a producer surplus. Just as buyers and sellers gain from trading goods and services, so they can also gain by trading risk. But risk is a bad, not a good. The good that is traded is risk avoidance. A buyer of risk avoidance can gain because the value of avoiding risk is greater than theprice that must be paid to someone else to get them to bear the risk. The seller of risk avoidance faces alower cost of risk than the price that people are willing to pay to avoid the risk. We're going to put some flesh on the bare bones of this brief account of how people can gain from trading risk by looking at insurance markets.

## Insurance Markets

Insurance plays a huge role in our economiclives. We'll explain

- How insurance reduces risk
- Why people buy insurance
- How insurance companies earn a profit

How Insurance Reduces Risk Insurance reduces the risk that people face by sharing or pooling the risks. When people buy insurance against the risk of an unwanted event, they pay an insurance company a premium. If the unwanted event occurs, the insur-ance company pays out the amount of the insured loss.
Think about auto collision insurance. The probability that any one person will have a serious auto accident is small. But a person who does have an auto accident incurs a large loss. For a large population, the probability of one person having an accident is the proportion of the population that has an accident. But this proportion is known, so the probability of an accident occurring .and the total cost of accidents can be predicted. An insurance company can pool the risks of a large population and enable everyone to share the costs. It does so by collecting premiums from everyone and paying out benefits to those who suffer a loss. An insurance company that remains in business collects at least as much in premiums as it pays out in benefits.

Why People Buy Insurance People buy insurance and insurance companies earn a profit by selling insurance because people are risk averse. To see whypeople buy insurance and why it is profitable, let's consider an example. Dan owns a car worth $\$ 10,000$, and that is his only wealth. There is a 10 percent chance that Dan will have a serious accident that makes his car worth nothing. So there is a 90 percent chance that Dan's wealth will remain at $\$ 10,000$ and a 10 percent chance that his wealth will be zero. Dan's expected wealth is $\$ 9,000-(\$ 10,000 \times 0.9)+(\$ 0 \times 0.1)$.
Dan is risk averse (just like Tania in the previous example). Because Dan is risk averse, he will be better off by buying insurance to avoid the risk that he faces, if the insurance premium isn't too high.

Without knowing some details about just how risk averse Dan is, we don't know the most that he would be willing to pay to avoid this risk. But we do know that he would pay more than $\$ 1,000$. IfDandid pay $\$ 1,000$ to avoid the risk, he would have $\$ 9,000$ of wealth and face no uncertainty about his wealth. If he does not have an accident, his wealth is the $\$ 10,000$ value of his car minus the $\$ 1,000$ hepays the insurance company. If he does lose his car, the insurance company pays him $\$ 10,000$, so he still has $\$ 9,000$. Being risk averse, Dan's expected utility from $\$ 9,000$ with no risk is greater than his expected utility from an expected $\$ 9,000$ with risk. So Dan would be willing to pay more than $\$ 1,000$ to avoid this risk.

How Insurance Companies Earn a Profit For the insurance company, $\$ 1,000$ is the minimum amount at which it would be willing to insure Dan and otherpeople like him. With say 50,000 customers all like Dan, 5,000 customers ( $50,000 \times 0.1$ ) lose their cars and 45,000 don't. Premiums of $\$ 1,000$ give the insurance company a total revenue of $\$ 50,000,000$. With 5,000 claims of $\$ 10,000$, the insurance company pays out $\$ 50,000,000$. So a premium of $\$ 1,000$ enables the insurance company to break even (make zero economic profit) on this business. But Dan (and everyone else) is willing to pay more than $\$ 1,000$, so insurance is a profitable business andthere is a gain from trading risk.
The gain from trading risk is shared by Dan (and the other people who buy insurance) and the insurance company. The exact share of the gain depends on the state of competition in the market for insurance.
If the insurance market is a monopoly, the insurance company can take all the gains from trading risk. But if the insurance market is competitive, economic profit will induce entry and profits will be competed away. In this case, Dan (and the other buyers of insurance) gets the gain.

## A Graphical Analysis of Insurance

We can illustrate the gains from insurance by using a graph of Dan's utility of wealth curve. We begin, in Fig. 20.4, with the situation if Dan doesn't buy insurance and decides to bear the risk he faces.

Risk-Taking Without Insurance With no accident, Dan's wealth is $\$ 10,000$ and his total utility is 100 units. If Dan has an accident, his car is worthless: he has no wealth and no utility. Because the chance of an accident is 10 percent (or 0.1 ), the chance of not having an accident is 90 percent (or 0.9 ). Dan's expected wealth is $\$ 9,000-(\$ 10,000 \times 0.9)+$ ( $\$ 0 \times 0.1$ )-and his expected utility is 90 units-( $100 \times 0.9$ ) + ( $0 \times 0.1$ ).
You've just seen that without insurance, Dan gets 90 units of utility. But Dan also gets 90 units of utility if he faces no uncertainty with a smaller amount of wealth. We're now going to see how much Dan will pay to avoid uncertainty.

The Value and Cost of Insurance Figure 20.5 shows the situation when Dan buys insurance. You can see that for Dan, having $\$ 7,000$ with no risk is just asgood as facing a 90 percent chance of having $\$ 10,000$ and a 10 percent chance of having no wealth. So if Dan pays $\$ 3,000$ for insurance, he has $\$ 7,000$ of wealth, faces no uncertainty, and gets 90 units of utility. The
amount of $\$ 3,000$ is the maxi- mum that Dan is willing to pay for insurance. It is the value of insurance to Dan.
Figure 20.5 also shows the cost of insurance. With a large number of customers each of whom has a 10 percent chance of making a $\$ 10,000$ claim for the loss of a vehicle, the insurance company can provide insurance at a cost of $\$ 1,000$ (10 percent of $\$ 10,000$ ). If Dan pays only $\$ 1,000$ for insurance, his wealth is $\$ 9,000$ (the $\$ 10,000$ value of his car minus': the $\$ 1,000$ he pays for insurance), and his utility from $\$ 9,000$ of wealth with no uncertainty is about 98 units.

FIGURE 20.4 Taking a Risk Without Insurance


Dan's wealth (the value of his car) is $\$ 10,000$, which gives him 100 units of utility.

With no insurance, if Dan has a crash, he has no wealth and no utility.

With a 10 percent chance of a crash, Dan's expected wealth is $\$ 9,000$ and his expected utility is 90 units.

Gains from Trade Because Dan is willing to pay up $i$ to $\$ 3,000$ for insurance that costs the insurance company $\$ 1,000$, there is a gain from trading risk of $\$ 2,000$ per insured person. How the gains are shared $\bullet$ depends on the nature of the market. If the insurance market is competitive, entry will increase supply and lower the price to \$1,000 (plus normal profit and operating costs). Dan (and the other buyers of insurance) enjoys a consumer surplus. If the insurance market is a monopoly, the insurance company takes the $\$ 2,000$ per insured person as economic profit.

FIGURE 20.5 The Gains from Insurance


If Dan pays $\$ 3,000$ for insurance, his wealth is $\$ 7,000$ and his utility is 90 units - the same utility as with no insurance-so $\$ 3,000$ is the value of insurance for Dan.

If Dan pays $\$ 1,000$ for insurance, which is the insurance company's cost of providing insurance, his wealth is $\$ 9,000$ and his utility is about 98 units.

Dan and the insurance company share the gain from insurance.

## Risk That Can't Be Insured

The gains from auto collision insurance that we've studied here apply to all types of insurance. Examples are property and casualty insurance, life insurance, and healthcare insurance. One person's risks associated with driving, life, and health are independent of other persons'. That's why insurance is possible. The risks are spread across a population. But not all risks can be insured. To be insurable, risks must be independent. If an event causes every- one to be a loser, it isn't possible to spread and pool the risks. For example, flood insurance is often not available for people who live on a floodplain because if one person incurs a loss, most likely alldo.
Also, to be insurable, a risky event must be observable to both the buyer and seller of insurance. But much of the uncertainty that we face arises because we know less (or more) than others with whom we do business. In the next section, we look at the way markets cope when buyers and sellers have different information.

## ECONOMICS IN ACTION

## Insurance in the United States

We spend 7 percent of our income on private insurance. That's more than we spend on cars or food. The figure shows the relative sizes of the four main types of private insurance. More than 80 percent of Americans have life insurance, and most have private health insurance.

In addition, we buy Medicare, Medicaid, Social Security, and unemployment insurance through our taxes.


The U.S. Insurance Industry

## Private Information

In all the markets that you've studied so far, the buyers and the sellers are well informed about the good, service, or factor of production being traded. But in some markets, either the buyers or the sellers - usually the sellers - are better informed about the value of the item being traded than the person on the other side of the market. Information about the value of the item being traded that is possessed by only buyers or sellers is called private information. And a market in which the buyers or sellers have private information has asymmetric information.

## Asymmetric Information: Examples and Problems

Asymmetric information affects many of your own economic transactions. One example is your knowledge about your driving skills and temperament.
You know much more than your auto insurancecompany does about how carefully and defensively you drive-about your personal risk of having an accident that would cause the insurance company to pay a claim. Another example is your knowledge about your work
effort. You know more than your employer about how hard you are willing to work. Yet another example is your knowledge about the quality of your car. You know whether it's a lemon, but the person to whom you are about to sell it does not know and can't find out until after he or she has bought it.
Asymmetric information creates two problems:

- Adverse selection
- Moral hazard

Adverse Selection is the tendency for people to enter into agreements in which they canuse their private information to their own advantage and to the disadvantage of the uninformed party.
For example, if Jackie offers her salespeople a fixed wage, she will attract lazy salespeople. Hardworking salespeople will prefer not to work for Jackie becausethey can earn more by working for someone who pays by results. The fixed-wage contract adverselyselects those with private information (knowledge about their work habits) who can use that knowledge to their own advantage and to the disadvantage of theother party.

Moral Hazard is the tendency for people with private information, after entering into an agreement, to use that information for their own benefit and at the cost of the less-informed party.
For example, Jackie hires Mitch as a salesperson and pays him a fixed wage regardless of how much hesells. Mitch faces a moral hazard. He has an incentiveto put in the least possible effort, benefiting himself and lowering Jackie's profits. For this reason, sales-people are usually paid by a formula that makes their income higher, the greater is the volume (or value) of,their sales.
A variety of devices have evolved that enable markets to function in the face of moral hazard and adverse selection. We've just seen one, the use of incentive payments for salespeople. We're going to look at how three markets cope with adverse selectionand moral hazard. They are:

- The market for used cars
- The market for loans
- The market for insurance


## The Market for Used Cars

When a person buys a car, it might turn out to be a lemon. If the car is a lemon, it is worth less to the buyer than if it has no defects. Does the used car market have two prices reflecting these two values-;-a:• low price for lemons and a higher price for cars with:' out defects? Itturns out that it does. But it needs some help to do so and to overcome what is called the lemons problem - the problem that in a market in which it is not possible to distinguish reliable products from lemons, there are too many lemons and too few reliable products traded.
To see how the used car market overcomes the ,lemons problem, we'll first look at a used car market that has a lemons problem.

The Lemons Problem in a Used Car Market Toexplain thelemons problem as clearly as possible, we'll assume that there are only two kinds of cars: defective cars-lemons-and cars without defects that we'll callgood cars. Whether or not a car is a lemon is private information that is available onlytd. the current owner. The buyer of a used car can't tell whether he is buying a lemon until after he has bought the car and learned as much about it as its current owner knows.
Some people with low incomes and the time and ability to fix cars are willing to buy lemons as long as they know what they're buying and pay an appropriately low price. Suppose that a lemon is worth $\$ 5,000$ to a buyer. More people want to buy a good car and we'll assume that a good car is worth $\$ 25,000$ to a buyer.
But the buyer can't tell the difference between a lemon and a good car. Only the seller has this information. And telling the buyer that a car is not a lemon does not help. The seller has no incentive to tell the truth.
So the most that the buyer knows is the probability of buying a lemon. If half of the used cars sold turn out to be lemons, the buyer knows that he has a 50 percent chance of getting a good car and a 50 percent chance of getting a lemon.
The price that a buyer is willing to pay for a car of unknown quality is more than the value of a lemon because the car might be a good one. But the price is less than the value of a good car because it might turn out to be a lemon.
Now think about the sellers of used cars, who know the quality of their cars. Someone who owns a good car is going to be offered a price that is less than the value of that car to the buyer. Many owners will be reluctant to sell for such a low price. So the quantity of good used cars supplied will not be as large as it would be if people paid the price they are worth.
In contrast, someone who owns a lemon is going to be offered a price that is greater than the value of that car to the buyer. So owners of lemons will be eager to sell and the quantity of lemons supplied will be greater than it would be if people paid the price that a lemon is worth.
Figure 20.6 illustrates the used car market that we've just described. Part (a) shows the demand for used cars, $D$, and the supply of used cars, $S$. The market equilibrium occurs at a price of $\$ 10,000$ per car with 400 cars traded each month.
Some cars are good ones and some are lemons, but buyers can't tell the difference until it is too late to influence their decision to buy. But buyers do know what a good car and a lemon are worth to them, and sellers know the quality of the cars they are offering for sale. Figure 20.6(b) shows the demand curve for good cars, DG, and the supply curve of good cars, $S G$. Figure 20.6(c) shows the demand curve for lemons, $D L$, and the supply curve of lemons, Sr.
At the market price of $\$ 10,000$ per car, owners of good cars supply 200 cars a month for sale. Owners of lemons also supply 200 cars a month for sale. Theused car market is inefficient because there are too many lemons and not enough good cars. Figure 20.6makes this inefficiency dear by using the concept ofdeadweight loss (see Chapter 5, pp. 190-191).

FIGURE 20.6 The Lemons Problem


Buyers can't tell a good used car from a lemon. Demand and supply determine the price and quantity of used cars traded in part (a). In part (b), $D_{G}$ is the demand curve for good used cars and $S_{G}$ is the supply curve. At the market price, too few
good cars are available, which brings a deadweight loss. In part (c), $D_{l}$ is the demand curve for lemons and $S_{l}$ is the supply curve. At the market price, too many lemons are available, which brings a deadweight loss.

At the quantity of good cars supplied, buyers are willing to pay $\$ 25,000$ for a good car. They are willing to pay more than a good car is worth to its cur- rent owner for all good cars up to 400 cars a month. The gray triangle shows the deadweight loss that results from there being too few good used cars.
At the quantity of lemons supplied, buyers are willing to pay $\$ 5,000$ for a lemon. They are willing to pay less than a lemon is worth to its current owner for all lemons above 150 cars a month. The gray tri- angle shows the deadweight loss that results from there being too many lemons.
You can see adverse selection in this used car market because there is a greater incentive to offer a lemon for sale. You can see moral hazard because the owner of a lemon has little incentive to take good care of it. The market for used cars is not working well. Too many lemons and too few good cars are traded.

A Used Car Market with Dealers' Warranties How can used car dealers convince buyers that a car isn't a lemon? The answer is by giving a guarantee in the form of a warranty. By providing warranties only on good cars, dealers signal which cars are good ones and which are lemons.

Signaling occurs when an informed person takes actions that send information to uninformed persons. The grades and degrees that a university awardsstudents are signals. They inform potential (uninformed) employers about the ability of the people they are considering hiring.
Inthe market for used cars, dealers send signals by giving warranties on the used cars they offer forsale. The message in the signal is that the dealer agrees to pay the costs of repairing the car if it turns out to have a defect.

Buyers believe the signal because the cost of sending a false signal is high. A dealer who gives a warranty on a lemon ends up bearing a high cost of repairs-and gains a bad reputation. A dealer who gives a warranty only on good cars has no repair costs and a reputation that gets better and better. It pays dealers to send an accurate signal, and it is rational for buyers to believe the signal.
So a car with a warranty is a good car; a car without a warranty is a lemon. Warranties solve the lemons problem and enable the used car market to function efficiently with two prices: one for lemonsand one for good cars.
Figure 20.7 illustrates this outcome. In part (a) the demand for and supply of good cars determine the price of a good car. In part (b), the demand for and supply of lemons determine the price of a lemon.
Both markets areefficient.
Pooling Equilibrium and Separating Equilibrium You've seen two outcomes in the market for used cars. Without warranties, there is only one message visible to the buyer: All cars look the same. So there is one price regardless of whether the car is a good car or a lemon. We call the equilibrium in a market when only one message is available and an uninformed person cannot determine quality a pooling equilibrium.
But in a used car market with warranties, there are two messages: Good cars have warranties and lemons don't. So there are two car prices for the two types of cars. We call the equilibrium in a market when signaling provides full information to a previously uninformed person a separating equilibrium.

## The Market for Loans

When you buy a tank of gasoline and swipe your credit card, you are taking a loan from the bank that issuedyour card. You demand and your bank supplies a loan. Have you noticed the interest rate on an unpaid creditcard balance? In 2014, it ranged between 7 percent ayear and 36 percent a year. Why are these interest rates so high? And why is there such a huge range? The answer is that when banks make loans, they face the risk that the loan will not be repaid. The risk that a borrower, also known as a creditor, might not repay a loan is called credit risk or default risk. For credit card borrowing, the credit risk is high and it varies among borrowers. The highest-risk borrowers pay the highest interest rate. Interest rates and the price of credit risk are determined in the market for loans. The lower the interest rate, the greater is the quantity of loans demanded and for a given level of credit risk, the higher the interest rate, the greater is the quantity of loans sup- plied. Demand and supply determine the interest rate and the price of credit risk. If lenders were unable to charge different interest rates to reflect different degrees of credit risk, there would be a pooling equilibrium and an inefficient loans market.

Inefficient Pooling Equilibrium Tosee why a pooling equilibrium would be inefficient, suppose that bankscan't identify the individual credit risk of their borrowers: they have no way of knowing how likely it isthat a given loan will be repaid. In this situation, every borrower pays the same interest rate and the market is in a pooling equilibrium.

If all borrowers pay the same interest rate, the market for loans has the same problem as the used car market. Low-risk customers borrow less than they would if they were offered the low interest rate appropriate for their low credit risk. High-risk customers borrow more than they would if they faced the high interest rate appropriate for their high credit risk. So banks face an adverse selection problem. Too many borrowers are high risk and too few are low risk.

FIGURE 20.7 Warranties Make the Used Car Market Efficient

(a) Good cars

With dealers' warranties as signals, the equilibrium price of a good used car is $\$ 20,000$ and 400 cars are traded. The market for good used cars is efficient. Because the signal

(b) Lemons
enables buyers to spot a lemon, the price of a lemon is $\$ 6,667$ and 150 lemons are traded. The market for lemons is efficient.

Signaling and Screening in the Market for Loans Lenders don't know how likely it is that a given loan will be repaid, but the borrower does know. Low-riskborrowers have an incentive to signal their risk by providing lenders with relevant information. Signals might include information about the length of time aperson has been in the current job or has lived at thecurrent address, home ownership, marital status, age,and business record. High-risk borrowers might be identified simply as those who have failed to signal low risk. Theseborrowers have an incentive to mislead lenders; andlenders have an incentive to induce high-risk borrowers to reveal their risk level. Inducing an informed party to reveal relevant private information is called screening.
By not lending to people who refuse to reveal relevant information, banks are able to screen as well as receive signals that help them to separate their borrowers into a number of credit-risk categories. If lenders succeed, the market for loans comes to a separating equilibrium with a high interest rate for high- risk borrowers and a low interest rate for low-risk borrowers. Signaling and screening in the market for loans act like warranties in the used car market and work to avoid the deadweight loss of a pooling equilibrium.

## The Market for Insurance

People who buy insurance face moral hazard, and insurance companies face adverse selection. Moral hazard arises because a person with insurance against' a loss has less incentive than an uninsured person to avoid the loss. For example, a business with fire insurance has less incentive to install a fire alarm or sprinkler system than a business with no fire insurance does. Adverse selection arises because people who' create greater risks are more likely to buy insurance. For example, a person with a history of driving infringements is more likely to buy collision insurance<: than is a person with a history of good driving. Insurance companies have an incentive to find ..ways around the moral hazard and adverse selection .problems. By doing so, they can lower premiums for low-risk people and raise premiums for high-risk people.

One way in which auto insurance companies separate high-risk and low-risk customers is with a no-claim bonus. A driver accumulates a no-claim bonus by driving safely and avoiding accidents. The greater the bonus, the greater is the incentive to drive carefully. Insurance companies also use a deductible. A deductible is the amount of a loss that the insured person agrees to bear. The premium is smaller, the larger is the deductible, and the decrease in the premium is more than proportionate to the increase in the deductible. By offering insurance with full coverage - no deductible - on terms that are attractive only to the high-risk people and by offering coverage with a deductible on more favorable terms that are attractive to low-risk people, insurance companies can do profitable business with everyone. High-risk people choose policies with a low deductible and a high premium; lowrisk people choose policies with a high deductible and a low premium.

## Uncertainty, Information, and the Invisible Hand

A recurring theme throughout microeconomics is thebig question: When do choices made in the pursuit ofself-interest also promote the social interest? When does the invisible hand work well and when does it fail us?
You've learned about the concept of efficiency, amajor component of what we mean by the social interest. And you've seen that while competitive markets generally do a good job in helping to achieve efficiency, impediments such as monopoly and the absence of well-defined property rights can prevent the attainment of an efficient use of resources.
How do uncertainty and incomplete information affect the ability of self-interested choices to lead to a social interest outcome? Are these features of economic life another reason why markets fail and why some type of government intervention is required to achieve efficiency? These are hard questions, and there are no definitive answers. But there are some useful things that we can say about the effects of uncertainty and a lack of complete information on the efficiency of resource use. We'll begin our brief review of this issue by thinking about information as just another good.

## ECONOMICS IN ACTION

## The Sub-Prime Credit Crisis

A sub-prime mortgage is a loan to a homebuyer who has a high risk of default. Figure 1 shows that between 2001 and 2005, the price of risk was low. Figure 2 shows why: The supply of credit, $S_{0}$, was large and so was the amount of risk taking. In 2007, the supply of credit decreased to $S_{1}$. The price of risk jumped and, faced with a higher interest rate, many sub-prime borrowers defaulted. Defaults in the subprime mortgage market spread to other markets that supplied the funds that financed mortgages.


Figure 1 The Price of Commercial Credit Risk


## Information asaGood

More information is generally useful, and less uncertainty about the future is generally useful. Think about information as one of the goods that we want more of.
The most basic lesson about efficiency that you learned in Chapter 2 can be applied to information. Along our production possibilities frontier, we face a tradeoff between information and all other goods and services. Information, like everything else, can be produced at an increasing opportunity cost-an increasing marginal cost. For example, we could get more accurate weather forecasts, but only at increasing marginal cost, as we increased the amount of information that we gather from the atmosphere and the amount of money that we spend on supercomputers to process the data.
The principle of decreasing marginal benefit also applies to information. More information is valuable, bur the more you know, the less you value another increment of information. For example, knowing that it will rain tomorrow is valuable information. Knowing the amount of rain to within an inch is even more useful. But knowing the amount of rain to within a millimeter probably isn't worth much more.
Because the marginal cost of information is increasing and the marginal benefit is decreasing, there is an efficient amount of information. It would be inefficient to be overinformed.
In principle, competitive markets in information might deliver this efficient quantity. Whether they actually do so is hard to determine.

## Monopoly in Markets that Cope with Uncertainty

There are probably large economies of scale in pro- viding services that cope with uncertainty and incomplete information. The insurance industry, for example, is highly concentrated. Where monopoly elements exist, exactly the same inefficiency issues arise as occur in markets where uncertainty and incomplete information are not big issues. So it is likely that in some information markets, including insurance markets, there is underproduction arising from the attempt to maximize monopoly profit. You've seen how people make decisions when faced with uncertainty and how markets work when there is asymmetric information. Economics inthe News on pp. 520-521 looks at the way markets in human capital and labor use grades as signals that sort students by ability so that employers can hire the type of labor they seek. You'll see why grade deflation can be efficient and grade inflation is inefficient.
Discriminating grades are in the social interest and in the self-interest of universities and students.

- Accurate grades provide valuable information to students and potential employers about a student's ability.
- Harvard, Princeton, and Yale (and most schools) want to provide accurate information and avoid grade inflationawarding a high grade to most students.
- The labor market for new college graduates works badly with grade inflation and works well with accurate grading.
- Figure 1 shows a labor market for new college graduates when there is grade inflation.
- Students with high ability are not distinguished from other students, and the supply curve represents the supply of students of all ability levels.
- The demand curve shows the employers' willingness to hire new workers without knowledge of their true ability.
- Students get hired for a low wage rate. Eventually, they get sorted by ability as employers discover the true ability of their workers from on-the-job performance.
- Figures 2 and 3 show the outcome with accurate grading.
- In Fig. 2, students with high grades get high-wage jobs and in Fig. 3, students with low grades get low-wage jobs.


Figure 1 Market with Grade Inflation

- The outcomes in Figs. 2 and 3 that arise immediately with accurate grading occur eventually with grade inflation as employers accumulate information about the abilities of the workers.
- But the cost to the student and the employer of discovering true ability is greater with grade inflation than with accurate grading.


Figure 2 The Market for A Students


Figure 3 The Market for D Students

# PART SEVEN: MONITORING MACROECONOMIC PERFORMANCE 

## CHAPTER 21: Measuring GDP and Economic Growth

After studying this chapter, you will be able to:

- Define GDP and explain why GDP equals aggregate expenditure and aggregate income
- Explain how the Bureau of Economic Analysis measures U.S. GDP and real GDP
- Explain the uses and limitations of real GDP as a measure of economic well-being

Will oureconomy start to expand more rapidly in 2015 or will growth remain slow? Or worse, will the economy slip into recession? U.S. businesses, both small and large, want to know the answers to these questions. To assess the state of the economy and to make big decisions about business expansion, firms such as Google and Amazon use forecasts of GDP. What is GDP andwhat does it tell us about the state of the economy? In this chapter, you will find out how economic statisticians at the Bureau of Economic Analysis measure GDP and its rate of growth. You will also learn about the uses and the limitations of these measures. In Economics in the News at the end of the chapter, you can see what GDP tells us about the state of the U.S. economytoday.

## Gross Domestic Product

What exactly is GDP, how is it calculated, what does it mean, and why do we care about it? You are going to discover the answers to these questions in thischapter. First, what is GDP?

## GDP Defined

GDP, or gross domestic product, is the market value of the final goods and services produced within a country in a given time period. This definition has four parts:

- Market value
- Final goods and services
- Produced within a country
- In a given time period

We'll examine each in turn.

Market Value To measure total production, we must add together the production of apples and oranges, computers and popcorn. Just counting the items doesn't get us very far. For example, which is the greater total production: 100 apples and 50 oranges or 50 apples and 100 oranges?
GDP answers this question by valuing items at their market values-the prices at which items are traded in markets. If the price of an apple is 10 cents, then the market value of 50 apples is $\$ 5$. If the price of an orange is 20 cents, then the market value of 100 oranges is
\$20. By using market prices to value production, we can add the apples and oranges together. The market value of 50 apples and 100 oranges is $\$ 5$ plus $\$ 20$, or $\$ 25$.

Final Goods and Services To calculate GDP, we value the final goods and services produced. A final good (or service) is an item that is bought by its final user during a specified time period. It contrasts with an intermediate good (or service), which is an item that is produced by one firm, bought by another firm, and used as a component of a final good or service.
For example, a Ford truck is a final good, but a Firestone tire on the truck is an intermediate good. An iPad is a final good, but an Apple ASX chip inside it is an intermediate good.
If we were to add the value of intermediate goods and services produced to the value of final goods and services, we would count the same thing many times-a problem called double counting. The value of a truck already includes the value of the tires, and ! the value of an iPad already includes the value of the chip inside it. Some goods can be an intermediate good in some situations and a final good in other situations. For example, the ice cream that you buy on a hot summer day is a final good, but the ice cream that a restaurant buys and uses to make sundaes is an inter-) mediate good. The sundae is the final good. So whether a good is an intermediate good or a final good depends on what it is used for, not what it is.
The purchase of a secondhand good - for example, a used car or existing home - isn't part of GDP; it's part of GDP in the year in which it was produced.

Produced Within a Country Only goods and services that are produced within a country count as part of that country's GDP. Nike Corporation, a U.S. firm, produces sneakers in Vietnam, and the market value of those shoes is part of Vietnam's GDP, not part of U.S. . GDP. Toyota, a Japanese firm, produces automobiles in Georgetown, Kentucky, and the value of this production is part of U.S. GDP, not part of Japan's GDP.

In a Given Time Period GDP measures the value of ' production in a given time period normally either a quarter of a year - called the quarterly GDP data - ora year - called the annual GDP data.

## GDP and the Circular Flow of Expenditure and Income

GDP is a measure of the value of total production. We can measure this value either by the total income earned producing GDP or the total expenditure on GDP. The equality between the value of total production and total income is important because it shows the direct link between productivity and living standards. Our standard of living rises when our income rise and we can afford to buy more goods and services. But we must produce more goods and services:" if we are to be able to buy more goods and services. ; Rising incomes and a rising value of production together. They are two aspects of the same phenomenon. You're now going to see why.

The Circular Flow Model Figure 21.1 illustrates the circular flow model. The economy consists of house-holds, firms, governments, and the rest of the world (the rectangles), which trade in factor markets and goods (and services) markets. We focus first on households and firms.

Households and Firms Households sell and firms buy the services of labor, capital, and land in factor markets. For these factor services, firms pay income tohouseholds: wages for labor services, interest for theuse of capital, and rent for the use of land. A fourth factor of production, entrepreneurship, receives profit.
Firms' retained earnings - profits that are not distributed to households - are part of the household sector's income. You can think of retained earnings as being income that households save and lend back to firms. Figure 21.1 shows the total income-aggregate income received by households, including retained earnings, as the blue flow labeled $Y$. Firms sell and households buy consumer goods and services in goods markets. The total payment for these goods and services is consumption expenditure, shown by the red flow labeled C.
Firms buy and sell new capital equipment-such as computer systems, airplanes, trucks, and assembly line equipment-in goods markets. Some of what firms produce is not sold but is added to inventory. For example, if GM produces 1,000 cars and sells 950 of them, the other 50 cars remain in GM'sinventory of unsold cars, which increases by 50 cars. When a firm adds unsold output to inventory, we can think of the firm as buying goods from itself The purchase of new plant, equipment, and buildings and the additions to inventories are investment, shown by the flow labeled

Governments buy goods and services from firms-government expenditure. In Fig. 21.1, government expenditure is shown as the flow $G$. Governments finance their expenditure with taxes and make financial transfers to households, such as Social Security benefits and unemployment benefits, and pay subsidies to firms. But taxes and financial transfers are not part of the circular flow of expenditure and income.

Rest of the World Firms in the United States sell goods and services to the rest of the world- exports-and buy goods and services from the rest of the world-imports. The value of exports $(X)$ minus the value of imports $(M)$ is called net exports, the red flow $X-M$ in Fig 21.1. If net exports are positive, the net flow is from U.S. firms to the rest of the world. If net exports are negative, the net flow is from the rest of the world to U.S. firms.

GDP Equals Expenditure Equals Income Gross domestic product can be measured in two ways: By the total expenditure on goods and services or by the total income earned producing goods and services.
The total expenditure-aggregate expenditure-is the sum of the red flows in Fig. 21.1. Aggregate expenditure equals consumption expenditure plus investmentplus government expenditure plus net exports.
Aggregate income is equal to the total amount paid for the services of the factors of production used to produce final goods and services-wages, interest,rent, and profit. The blue flow in Fig. 21.1 shows aggregate income. Because firms pay out as incomes (including retained profits) everything they receive from the sale of their output, aggregate income (the blue flow) equals aggregate expenditure (the sum of the dark flows). That is,

$$
Y=C+I+G+X-M .
$$

The table in Fig. 21.1 shows the value of each expenditure in 2014 and that their sum is $\$ 17,044$ billion, which also equals aggregate income.

FIGURE 21.1 The Circular Flow of Expenditure and Income


Because aggregate expenditure equals aggregate income, the two methods of measuring GDP give the same answer. So GDP equals aggregate expenditure and equals aggregate income.
The circular flow model is the foundation on which the national economic accounts are built.

## Why "Domestic" and Why "Gross"?

What do the words "domestic" and "gross" mean in the term gross domestic product? Domestic product is production within acountry. It contrasts with a related concept, national product, which is the value of goods and services produced anywhere in the world by the residents of a nation. For example, Nike's income from shoe factories that it owns in Vietnam is part of U.S. nationalproduct. But it is part of Vietnam's domestic product. Gross national product, GNP, equals GDP plus net income from factors of production owned in other countries.

Gross means before subtracting the depreciation of capital. The opposite of gross is net, which means after subtracting the depreciation of capital.
Depreciation is the decrease in the value of a firm's capital that results from wear and tear and obsolescence. The total amount spent both buying new capital and replacing
depreciated capital is called gross investment. The amount by which the value of capital increases is called net investment. Netinvestment equals gross investment minus depreciation. Gross investment is one of the expenditures included in the expenditure approach to measuring GDP. So the resulting value of total product is a gross measure. Gross profit, which is a firm's profit before subtracting depreciation, is one of the incomes included in the income approach to measuring GDP. So again, the resulting value of total product is a gross measure.

## Measuring U.S. GDP

The Bureau of Economic Analysis (BEA) uses the concepts in the circular flow model to measure GDP and its components in the National Income and Product Accounts. Because the value of aggregate production equals aggregate expenditure and aggregate income, there are two approaches available for measuring GDP, and both are used. They are

- The expenditure approach
- The income approach


## The Expenditure Approach

The expenditure approach measures GDP as the sum of consumption expenditure (C), investment (!), government expenditure on goods and services (G), and net exports of goods and services ( $X-M$ ). These expenditures correspond to the red flows through the goods markets in the circular flow model in Fig. 21.1 and Fig. 21.2. Table 21.1 shows these expenditures and GDP for 2014.
Personal consumption expenditures (the flow C in Fig. 21.2) are the expenditures by U.S. households ongoods and services produced in the United States andin the rest of the world. They include goods such as soda and books and services such as banking and legaladvice. They also include the purchase of consumer durable goods such as TVs and microwave ovens. But they do not include the purchase of new homes, which the BEA counts as part of investment.
Gross private domestic investment (the red flow / in Fig. 21.2) is expenditure on capital equipment andbuildings by firms and the additions to business inventories. It also includes expenditure on new homesby households.
Government expenditure on goods and services (the red flow G in Fig. 21.2) is the expenditure by all levels of government on goods and services such as national defense and garbage collection. It does not include transfer payments, such as unemploymentbenefits, because they are not expenditures on goodsand services.
Net exports of goods and services (the red flow $X-M$ in Fig. 21.2) are the value of exports minus the value of imports. This item includes the value of airplanes that Boeing sells to British Airways (a U.S. export), and the value of Japanese DVD players that Best Buy purchases from Sony (a U.S. import).

## FIGURE 21.2 Aggregate Expenditure



Aggregate expenditure is the sum of the dark arrows.
TABLE 21.1 GDP: The Expenditure Approach

| Item | Symbol | Amount in 2014 (billions of dollars) | Percentage of GDP |
| :---: | :---: | :---: | :---: |
| Personal consumption expenditures | C | 11,729 | 68.8 |
| Gross private domestic investment | I | 2,714 | 15.9 |
| Government expenditure on goods and services | G | 3,139 | 18.4 |
| Net exports of goods and services | $X-M$ | -538 | -3.2 |
| Gross domestic product | $Y$ | 17,044 | 100.0 |

The expenditure approach measures GDP as the sum of personal consumption expenditures, $C$, gross private domestic investment, $I$, government expenditure on goods and services, $G$, and net exports, $X-M$. In 2014, GDP measured by the expenditure approach was $\$ 17,044$ billion. Expenditure on personal consumption goods and services is more than two thirds of aggregate expenditure.

Source of data: U.S. Department of Commerce, Bureau of Economic
Analysis. The data are for the first quarter of 2014 at an annual rate.

## The Income Approach

The income approach measures GDP by summing the incomes that firms pay households for the services of the factors of production they hire-wages for labor, interest for capital, rent for land, and profit for entrepreneurship. These incomes correspond to the blue flows through the factor markets in the circular flow model in Fig. 21.1 and Fig. 21.3. Table 21.2 shows these incomes and GDP for 2014.
Compensation of employees (the blue flow Win Fig. is the payment for labor services. It includes net wages and salaries (called "take-home pay") that workers receive plus taxes withheld on earnings plus fringe benefits such as Social Security and pension fund contributions. This item is more than 50 per-cent of total income.
Net interest, rentalincome, corporateprofits, and proprietors 'income are earned by capital and land. These other factor incomes are in the blue flow OF! in Fig. 21.3.
The factor incomes sum to net domestic income at factor cost, which is the cost of the factors of production used to produce final goods. The expenditures on final goods are valued at market prices, which differ from factor cost because of indirect taxes and subsidies.
An indirect tax is a tax such as a sales tax or a tax on gasoline. Market price includes indirect taxes, so exceeds factor cost. A subsidy is a payment, such as a farm subsidy, by the government to a producer. Subsidies make market price less than factor cost.

FIGURE 21.3 Aggregate Income


The aggregate income is the sum of the dark arrows.
To get from factor cost to market price, we add indirect taxes and subtract subsidies and get net domestic income at market prices. We still must get from) a net to a gross measure. Total expenditure is a gross number because it includes gross investment. Net domestic income at market prices is a net income measure because corporate profits are measured after deducting depreciation. They are a net income measure. To get from net income to gross income, we must add depreciation.
We've now arrived at GDP using the income approach. This number is not exactly the same as GDP using the expenditure approach because all the numbers are estimates. The gap between the two measures of GDP, called the statistical discrepancy, is never large. In 2014, it was 1.1 percent of GDP.

## TABLE 21.2 GDP: The Income Approach

| Item | Amount in 2014 (billions of dollars) | Percentage of GDP |
| :---: | :---: | :---: |
| Compensation of employees | 9,109 | 53.4 |
| Net interest | 685 | 4.0 |
| Rental income | 623 | 3.7 |
| Corporate profits | 1,514 | 8.9 |
| Proprietors' income | 1,351 | 7.9 |
| Net domestic income at factor cost | 13,282 | 77.9 |
| Indirect taxes less subsidies | 1,244 | 7.3 |
| Net domestic income at market prices | 14,526 | 85.2 |
| Depreciation | 2,699 | 15.8 |
| GDP (income approach) | 17,225 | 101.1 |
| Statistical discrepancy | -181 | -1.1 |
| GDP (expenditure approach) | 17,044 | 100.0 |

The sum of factor incomes equals net domestic income at factor cost. GDP equals net domestic income at factor cost plus indirect taxes minus subsidies plus depreciation.

In 2014, GDP measured by the income approach was $\$ 17,225$ billion. This amount is $\$ 181$ billion more than GDP measured by the expenditure approach-a statistical discrepancy of $\$ 181$ billion or 1.1 percent of GDP.

Compensation of employees-labor income-is by far the largest component of aggregate income.

## Nominal GDP and Real GDP

Often, we want to compare GDP in two periods, say 2000 and 2014. In 2000, GDP was $\$ 10,285$ billion and in 2014, it was $\$ 17,044$ billion-66 percenthigher than in 2000 . This increase in GDP is a combination of an increase in production and a rise in prices. To isolate the increase in production from the rise in prices, we distinguish between real GDP and nominal GDP Real GDP is the value of final goods and services produced in a given year when valued at the prices of a reference base year. By comparing the value of production in the two years at the same prices, we reveal the change in production. Currently, the reference base year is 2009 and we describe real GDP as measured in 2009 dollars-in terms of what the dollar would buy in 2009.

Nominal GDP is the value of final goods and services produced in a given year when valued at the prices of that year. Nominal GDP is just a more precise name for GDP.
Economists at the Bureau of Economic Analysis calculate real GDP using the method described in the Mathematical Note on pp. 546-547. Here, we'll explain the basic idea but not the technical details.

## Calculating Real GDP

We'll calculate real GDP for an economy that produces one consumption good, one capital good, and one government service. Net exports are zero.
Table 21.3 shows the quantities produced and the prices in 2009 (the base year) and in 2014. In part (a), we calculate nominal GDP in 2009. For each item, wemultiply the quantity produced in 2009 by its price in 2009 to find the total expenditure on the item. We sum the expenditures to find nominal GDP, which in 2009 is $\$ 100$ million. Because 2009 is the base year, both real GDP and nominal GDP equal $\$ 100$ million.
In Table 21.3(b), we calculate nominal GDP in 2014, which is $\$ 300$ million. Nominal GDP in 2014 is three times its value in 2009. But by how much has production increased? Real GDP will tell us.

In Table 21.3(c), we calculate real GDP in 2014. The quantities of the goods and services produced are those of 2014, as in part (b). The prices are those in the reference base year2009, as in part (a).
For each item, we multiply the quantity produced in 2014 by its price in 2009. We then sum these expenditures to find real GDP in 2014, which is $\$ 160$ million. This number is what total expenditure would have been in 2014 if prices had remained the same as they were in 2009.
Nominal GDP in 2014 is three times its value in 2009, but real GDP in 2014 is only 1.6 times its 2009 value-a 60 percent increase in production.

## The Uses and Limitations of Real GDP

Economists use estimates of real GDP for two main purposes:

- To compare the standard of living over time
- To compare the standard of living across countries


## The Standard of Living Over Time

One method of comparing the standard of living over time is to calculate real GDP per person in different years. Real GDP per person is real GDP divided by the population. Real GDP per person tells us the value ofgoods and services that the average person can enjoy.By using real GDP, we remove any influence that rising prices and a rising cost of living might have had on our comparison. We're interested in both the long-term trends and the shorterterm cycles in the standard of living.

TABLE 21.3 Calculating Nominal GDP and Real GDP

|  | Item | Quantity (millions) | Price (dollars) | Expenditure (millions of dollars) |
| :---: | :---: | :---: | :---: | :---: |
| (a) In 2009 |  |  |  |  |
| C | T-shirts | 10 | 5 | 50 |
| 1 | Computer chips | 3 | 10 | 30 |
| G | Security services | 1 | 20 | 20 |
| $Y$ | Real GDP in 2009 |  |  | 100 |
| (b) In 2014 |  |  |  |  |
| C | T-shirts | 4 | 5 | 20 |
| 1 | Computer chips | 2 | 20 | 40 |
| G | Security services | 6 | 40 | 240 |
| $\gamma$ | Nominal GDP in 2 | 2014 |  | 300 |
| (c) Quantities of 2014 valued at prices of 2009 |  |  |  |  |
| $C$ | T-shirts | 4 | 5 | 20 |
| , | Computer chips | 2 | 10 | 20 |
| G | Security services | 6 | 20 | 120 |
| $Y$ | Real GDP in 2014 |  |  | 160 |

In 2009, the reference base year, real GDP equals nominal GDP and was $\$ 100$ million. In 2014, nominal GDP increased to $\$ 300$ million, but real GDP, which is calculated by using the quantities in 2014 in part (b) and the prices in 2009 in part (a), was only $\$ 160$ million-a 60 percent increase from 2009.

Long-Term Trend A handy way of comparing real GDP per person over time is to express it as a ratio of some reference year. For example, in 1960, real GDP per person was $\$ 17,210$ and in 2013, it was $\$ 49,658$. So real GDP per person in 2013 was almost 3 times its 1960 level ( $\$ 49,658+\$ 17,210=2.9$ ).
To the extent that real GDP per person measures the standard of living, people were 2.9 times as well offin 2013 as their grandparents had been in 1960.
Figure 21.4 shows the path of U.S. real GDP per person for the 53 years from 1960 to 2013 and high- lights two features of our expanding living standard:

- The growth of potential GDP per person
- Fluctuations of real GDP per person

The Growth of Potential GDP Potential GDP is the maximum level of real GDP that can be produced while avoiding shortages of labor, capital, land, and entrepreneurial ability that would bring rising inflation. Potential GDP per person, the smoother black line inFig. 21.4, grows at a steady pace because the quantities of the factors of production and their
productivities grow at a steady pace. But potential GDP per person doesn't grow at a constant pace. During the 1960s, it grew at 2.8 percent per year but slowed to only 2.3 percent per year during the 1970s. This slowdown might seem small, but it had big consequences, as you'll soon see.

FIGURE 21.4 Rising Standard of Living in the United States


Fluctuations of Real GDP You can see that real GDP !; shown by the red line in Fig. 21.4 fluctuates around(potential GDP, and sometimes real GDP shrinks. Let's take a closer look at the two features of our \} expanding living standard that we've just outlined.

Productivity Growth Slowdown How costly was the slowdown in productivity growth after 1970? The answer is provided by the Lucas wedge, which is the dollar value of the accumulated gap between what real GDP per person would have been if the 1960s growth rate had persisted and what real GDP per person turned out to be. (Nobel Laureate Robert K. Lucas Jr. drew attention to this gap.)

Figure 21.5 illustrates the Lucas wedge. The wed started out small during the 1970s, but by 2013 re\# GDP per person was $\$ 31,000$ per year lower than.it, would have been with no growth slowdown and the accumulated gap was an astonishing \$400,000 per person.

FIGURE 21.5 The Cost of Slower Growth:
The Lucas Wedge


The black line projects the 1960s growth rate of real GDP per person to 2013. The Lucas wedge arises from the slowdown of productivity growth that began during the 1970s. The cost of the slowdown is $\$ 400,000$ per person.

Real GDP Fluctuations-The Business Cycle We call the fluctuations in the pace of expansion of real GDP the business cycle. The business cycle is a periodic but irregular up-and-down movement of total production and other measures of economic activity. The business cycle isn't a regular predictable cycle like the phases of the moon, but every cycle has twophases:

1. Expansion
2. Recession
and two turning points:
3. Peak
4. Trough

Figure 21.6 shows these features of the most recent U.S. business cycle.
An expansion is a period during which real GDP increases. In the early stage of an expansion, real GDP returns to potential GDP; and as the expansion progresses, potential GDP grows and real GDP eventually exceeds potential GDP.
A common definition of recession is a period during which real GDP decreases-its growth rate is negative-for at least two successive quarters. The definition used by the National Bureau of Economic Research, which dates the U.S. business cyclephases and turning points, is "a period of significant decline in total output, income, employment, and trade, usually lasting from six months to a year, and marked by contractions in many sectors of the economy."

An expansion ends and recession begins at a business cycle peak, which is the highest level that real GDP has attained up to that time. A recession ends at a trough, when real GDP reaches a temporary low point and from which the next expansion begins.
In 2008, the U.S. economy went into an unusually severe recession. Starting from a long way below potential GDP, a new expansion began in mid-2009. Estimates of the exact depth of the recession have changed with revisions of the real GDP data.

FIGURE 21.6 The Most Recent U.S. Business Cycle


A recession began at a peak in the fourth quarter of 2007 and ended at a trough in the second quarter of 2009. A slow expansion then began, but real GDP has remained below potential GDP for more than five years.

## The Standard of living Across Countries

Two problems arise in using real GDP to compare living standards across countries. First, the real GDP of one country must be converted into the same currency units as the real GDP of the other country.
Second, the goods and services in both countries must be valued at the same prices. Comparing the United States and China provides a striking example of these two problems.

China and the United States in U.S. Dollars $\operatorname{In} 2013$, nominal GDP per person in the United States was $\$ 53,000$ and in China it was 42,000 yuan. The yuan is the currency of China and the price at which the dollar and the yuan exchanged, the market exchange rate, was 6.2 yuan per $\$ 1$ U.S. Using this exchangerate, 42,000 yuan converts to $\$ 6,775$. On these numbers, GDP per person in the United States in 2013 was 7.9 times that in China.

The bottom line in Fig. 21.7 shows real GDP per person in China from 1980 to 2013 when the market exchange rate is used to convert yuan to U.S. dollars.

China and the United States at PPP Figure 21.7 shows a second estimate of China's real GDP per person thatvalues China's production on the same terms as U.S. production. Ituses purchasing power parity or PPP prices, which are the sameprices for both countries.
figure 21.7 Two Views of Real GDP in China


Real GDP per person in China has grown rapidly. But how rapidly it has grown and to what level depends on how red GDP is valued. When GDP in 2013 is valued at the markel exchange rate, U.S. income per person is 7.9 times that in China. China looks like a poor developing country. But this comparison is misleading. When GDP is valued at purchas. ing power parity prices (PPP), U.S. income per person is onlv 5.4 times that in China.

The prices of some goods are higher in the United States than in China, so these items get a smaller weight in the calculation of China's real GDP than they get in U.S. real GDP. An example is a Big Mac) that costs $\$ 4.80$ in Chicago. In Shanghai, a Big Mac costs 16.93 yuan which is the equivalent of $\$ 2.73$.in China's real GDP, a Big Mac gets about half the weight that it gets in U.S. real GDP.
Some prices in China are higher than in the United States but more prices are lower, so Chinese prices put a lower value on China's production than do U.S. prices.
According to the PPP comparisons (same prices in both countries), real GDP per person in the United States in 2013 was 5.4 times that of China, not 7.9 times.

Limitations of Real GDP Real GDP measures the value of goods and servicesthat are bought in markets. Some of the factors thatinfluence the standard of living and that are not partof GDP are:

- Household production
- Underground economic activity
- Leisure time
- Environmental quality

Household Production Preparing meals, changing a light bulb, cutting grass, and caring for a child are all examples of household production. Because these productive activities are not traded in markets, they are not included in GDP.
The omission of household production from GDP means that GDP underestimates total production. But it also means that the growth rate of GDP overestimates the growth rate of total production. The reason is that some of the growth rate of market production (included in GDP) is a replacement for home production. So part of the increase in GDP arises from a decrease in home production.

Underground Economic Activity The underground economy is the part of the economy that is purposelyhidden from the view of the government to avoid taxes and regulations or because the goods and services being produced are illegal. Because underground economic activity is unreported, it is omitted from GDP. The underground economy in the United States ranges between 9 and 30 percent of GDP ( $\$ 1,500$ billion to $\$ 5,000$ billion).

Leisure Time Leisure time is an economic good that adds to our economic well-being and the standard ofliving. Other things remaining the same, the more leisure we have, the better off we are. Our working time is valued as part of GDP, but our leisure time is not. Yet that leisure time must be at least as valuable to us as the wage that we earn for the last hour worked. If it were not, we would work instead of taking leisure. Over the years, leisure time has steadily increased. The workweek has become shorter, more people take early retirement, and the number of vacation days has increased. These improvements in economic well-being are not reflected in real GDP.

Environmental Quality Economic activity directly influences the quality of the environment. Burningoil and coal brings global warming and climate change. Depleting nonrenewable resources, clearing forests, and polluting lakes and rivers are other environmental consequences of industrial production.
Resources used to protect the environment are valued as part of GDP. For example, the value of catalytic converters that help to protect the atmosphere from automobile emissions is part of GDP. But the cost of pollution is not subtracted from GDP. An industrial society possibly produces more atmospheric pollution than an agricultural society does. But pollution does not always increase as we become wealthier. Wealthy people value a clean environment and are willing to pay for one. Compare the pollution in China today with pollution in the United States. China, a poor country, pollutes its rivers, lakes, and atmosphere in a way that is unimaginable in the United States.

## ECONOMICS IN ACTION

## A Broader Indicator of Economic Well-Being

The limitations of real GDP reviewed in this chapter affect the standard of living and general well-being of every country. So to make international comparisons of the general state of economic well-being, we must look at real GDP and other indicators.

The United Nations has constructed a broader measure called the Human Development Index (HDI), which combines real GDP, life expectancy and health, and education. Real GDP per person (measured on the PPP basis) is a major component of the HDI so, as you can see in the figure, the two are strongly correlated.

The figure shows the data for 2012. In that year, Norway had the highest HDI and Australia had the second highest, but Qatar had the highest real GDP per person. The United States had the fifth highest HDI.

The HDI of the United States is lower than that of Norway and Australia because the people of those countries live longer and have better access to healthcare and education than do Americans.
African nations have the lowest levels of economic well-being. The figure shows that Niger had the lowest


The Human Development Index
Source of data: United Nations hdr.undp.org/en/statistics/data.

HDI and the Democratic Republic of Congo had the lowest real GDP per person.

The Bottom Line Do we get the wrong message about the level and growth in economic wellbeing and the standard of living by looking at the growth of real GDP? The influences that are omitted from real GDP are probably large. Developing countries have a larger amount of household production thando developed countries, so the gap between their living standards is exaggerated. Also, as real GDP grows, part of the growth is a switch from home production

## ECONOMIC ANALYSIS

- The news article reports the first estimates of real GDP and its expenditure components for the second quarter of 2014.
- A news article on this topic appears every three months.
- In the second quarter of 2014, real GDP increased by $\$ 154$ billion from $\$ 15,832$ billion to $\$ 15,986$ billion.
- The increase in real GDP during the second quarter was a bit less than 1 percent ( $\$ 154$ is 0.97 percent of $\$ 15,832$ ).
- If this growth rate is maintained for a full year, real GDP will be 4 percent higher at the end of the year. That is what the news article means when it reports that real GDP grew at an annualized rate of 4 percent.
- The growth in the second quarter of 2014 came after a quarter in which real GDP shrank, so some of the second quarter growth was returning to the level at the end of 2013.
- Figure 1 shows the increases in real GDP and the expenditure components for the second quarter of 2014 as the blue bars and for the full year from the second quarter of 2013 as the red bars.
- Year-over-year, the red bars in Fig. 1, real GDP increased by $\$ 379$ billion and consumption expenditure increased most.
- In the second quarter of 2014, as reported in the news article and shown by the blue bars in Fig.1, business inventories (shown separately from business fixed investment) and consumption contributed most to the expansion. Government expenditure barely changed, and net exports decreased aggregate expenditure (were "a drag" says the news report).
- The news article says that a big contributor to expansion came from a business "inventory build-up, which reversed a big decline in the first quarter but is not sustainable."
- Figure 2 shows whether the news article is correct.
- It is correct that an increase in business inventories were a large component of the second quarter expansion. They increased by $\$ 93$ billion, which is 60 percent of the increase in real GDP.
- But the news article is not correct that inventories decreased in the first quarter. They increased by $\$ 35$ billion despite a fall in real GDP.
- The news article might also be wrong about sustainability. Figure 2 shows that in almost every quarter (except


Figure 1 Aggregate Expenditure Changes


Figure 2 Quarterly Changes in Real GDP and Business Inventories
one) since 2010 the change in business inventories has been positive-business inventories have increased.

- Figure 2 also shows that business inventories and real GDP generally change in the same direction but that inventories sometimes lag the change in real GDP.


## APPENDIX

Graphs in Macroeconomics

## After studying this appendix, you will be able to:

- Make and interpret a time-series graph
- Make and interpret a graph that uses a ratio scale


## The Time-Series Graph

In macroeconomics we study the fluctuations and trends in the key variables that describe macroeconomic performance and policy. These variables include GDP and its expenditure and income components that you've learned about in this chapter. They also include variables that describe the labor market and consumer prices that you study in Chapter 22.

Regardless of the variable of interest, we want to be able to compare its value today with that in the past; and we want to describe how the variable has changed over time. The most effective way to do these things is to make a time-series graph.

## Making a Time-Series Graph

A time-series graph measures time (for example, years, quarters, or months) on the $x$-axis and the variable or variables in which we are interested on the $y$-axis. Figure A21.1 is an example of a time-series graph. It provides some information about unemployment in the United States since 1994. In this figure, we measure time in years starting in 1994. We measure the unemployment rate (the variable that we are interested in) on the $y$-axis.

A time-series graph enables us to visualize how a variable has changed over time and how its value in one period relates to its value in another period. It conveys an enormous amount of information quickly and casily.

Let's see how to "read" a time-series graph.

## Reading a Time-Series Graph

To practice reading a time-series graph, take a close look at Fig. A21.1. The graph shows the level, change, and speed of change of the variable.
figure A21.1 A Time-Series Graph


A time-series graph plots the level of a variable on the $y$-axis against time (here measured in years) on the $x$-axis. This graph shows the unemployment rate each year from 1994 to 2014. The graph shows when unemployment was high, when it was low, when it increased, when it decreased, and when it changed quickly and slowly.

MyEconLab Animation.

- The level of the variable: It tells us when unemployment is high and low. When the line is a long distance above the $x$-axis, the unemployment rate is high, as it was, for example, in 2009. When the line is close to the $x$-axis, the unemployment rate is low, as it was, for example, in 2001.
- The change in the variable: It tells us how unemployment changes-whether it increases or decreases. When the line slopes upward, as it did in 2008 and 2009 , the unemployment rate is rising. When the line slopes downward, as it did in 1997, the unemployment rate is falling.
- The speed of change in the variable: It tells us whether the unemployment rate is rising or falling quickly or slowly. If the line is very steep, then the unemployment rate increases or decreases quickly. If the line is not steep, the unemployment rate increases or decreases slowly. For example, the unemployment rate rose quickly in 2008 and slowly in 2003 and it fell quickly in 2012 and slowly in 1997.


## Ratio Scale Reveals Trend

A time-series graph also reveals whether a variable has a cycle, which is a tendency for a variable to alternate between upward and downward movements, or a trend, which is a tendency for a variable to move in one general direction.

The unemployment rate in Fig. A21.1 has a cycle but no trend. When a trend is present, a special kind of time-series graph, one that uses a ratio scale on the $y$-axis, reveals the trend.

## A Time-Series with a Trend

Many macroeconomic variables, among them GDP and the average level of prices, have an upward trend. Figure A21.2 shows an example of such a variable: the average prices paid by consumers.

In Fig. A21.2(a), consumer prices since 1974 are graphed on a normal scale. In 1974 the level is 100 . In other years, the average level of prices is measured as a percentage of the 1974 level.

The graph clearly shows the upward trend of prices. But it doesn't tell us when prices were rising fastest or whether there was any change in the trend. Just looking at the upward-sloping line in Fig. A21.2(a) gives the impression that the pace of growth of consumer prices was constant.

## Using a Ratio Scale

On a graph axis with a normal scale, the gap between 1 and 2 is the same as that between 3 and 4 . On a graph axis with a ratio scale, the gap between 1 and 2 is the same as that between 2 and 4 . The ratio 2 to 1 equals the ratio 4 to 2 . By using a ratio scale, we can "see" when the growth rate (the percentage change per unit of time) changes.

Figure A21.2(b) shows an example of a ratio scale. Notice that the values on the $y$-axis get closer together but the gap between 400 and 200 equals the gap between 200 and 100: The ratio gaps are equal.
Graphing the data on a ratio scale reveals the trends. In the case of consumer prices, the trend is much steeper during the 1970s and early 1980s than in the later years. The steeper the line in the ratioscale graph in part (b), the faster are prices rising. Prices rose rapidly during the 1970s and early 1980s and more slowly in the later 1980s and 1990s. The ratio-scale graph reveals this fact. We use ratio-scale graphs extensively in macroeconomics.

FIGURE A21.2 Ratio Scale Reveals Trend

(a) Normal scale

(b) Ratio scale

The graph shows the average of consumer prices from 1974 to 2014. The level is 100 in 1974, and the values for the other years are percentages of the 1974 level. Consumer prices normally rise each year, so the line slopes upward. In part (a), where the $y$-axis scale is normal, the rate of increase appears to be constant.

In part (b), where the $y$-axis is a ratio scale (the ratio of 400 to 200 equals the ratio 200 to 100), prices rose faster during the 1970s and early 1980s and slower in the later years. The ratio scale reveals this trend.

MyEconLab Animation

## Chapter 22: Monitoring Jobs and Inflation

After studying this chapter, you will beable to:

- Explain why unemployment is a problem and how we measure the unemployment rate and other labor market indicators
- Explain why unemployment occurs and why it is present even at full employment
- Explain why inflation is a problem and how we measure the inflationrate

Each month, we chart the course of unemployment inflation as measures of U.S. economic health. How dowe measure the unemployment rate and the inflation rate and are they reliable vital signs for the economy?
As the U.S. economy slowly expanded after a recession in 2008 and 2009, job growth was weak and questions about the health of the labor market became of vital importance to millions of Americanfamilies. Economics in the News, at the end of this chapter, puts the spotlight on the labor market through recession and a weak 2010-2014 expansion.

## Employment and Unemployment

What kind of job market will you enter when you graduate? Will there be plenty of good jobs to choose among, or will jobs be so hard to find that you end up taking one that doesn't use your education and pays a low wage? The answer depends, to a large degree, on the total number of jobs available and on the number of people competing for them. The class of 2014 had a tough time in the jobs market. In July 2014, four years after a recession, 10 million Americans wanted a job but couldn't find one, and another 8 million had either given up the search for a job or had reluctantly settled for a part-time job. Despite the high unemployment, the U.S. economy is an incredible job-creating machine. Even in 2009 at the depths of recession, 139 million people had jobs- 22 million more than in 1989. But in recent years, population growth has outstripped jobs growth, so unemployment is a serious problem.

## ECONOMICS IN ACTION

## What Kept Ben Bernanke Awake at Night

The Great Depression began in October 1929, when the U.S. stock market crashed. It reached its deepest point in 1933, when 25 percent of the labor force was unemployed, and lasted until 1941, when the United States entered World War II. The depression quickly spread globally to envelop most nations.
The 1930s were and remain the longest and worst period of high unemployment in history. Failed banks, shops, farms, and factories left millions of Americans without jobs, homes, and food. Without the support of government and charities, millions would have starved.
The Great Depression was an enormous political event: It fostered the rise of the German and Japanese militarism that were to bring the most devastating war humans have ever fought. It also led to President Franklin D. Roosevelt's "New Deal," which
enhanced the role of government in economic life and made government intervention in markets popular and the market economy unpopular.
The Great Depression also brought a revolution in economics. British economist John Maynard Keynes published his General Theory of Employment, Interest, and Money and created what we now call macroeconomics.

Why Unemployment Is a Problem Unemployment is a serious personal and social economic problem for two main reasons. It results in

- Lost incomes and production
- Lost human capital

Lost Incomes and Production The loss of a job brings a loss of income and lost production. These losses are devastating for the people who bear them and they make unemployment a frightening prospect for everyone. Unemployment benefits create a safety net, but they don't fully replace lost earnings.
Lost production means lower consumption and a lower investment in capital, which lowers the living standard in both the present and the future.

Lost Human Capital Prolonged unemployment permanently damages a person's job prospects by destroying human capital.

Think about a manager who loses his job when his employer downsizes. The only work he can find is driving a taxi. After a year in this work, he discovers that he can't compete with new MBA graduates.
Eventually, he gets hired as a manager but in a small firm and at a lower wage than before. He has lost some of his human capital.
The cost of unemployment is spread unequally, which makes it a highly charged political problem as well as a serious economic problem.
Governments make strenuous efforts to measure unemployment accurately and to adopt policies to moderate its level and ease its pain. Here, we'll learn how the U.S. government monitors unemployment.

## Current Population Survey

Every month, the U.S. Census Bureau surveys 60,000 households and asks a series of questions about the age and job market status of the members of each household. This survey is called the Current Population Survey (or CPS). The Census Bureau uses the answers to chart the course of the laborforce.
Figure 22.1 shows the population categories used by the Census Bureau and the relationships among the categories.
The population divides into two broad groups: the working-age population and others who are too young to work or who live in institutions and are unable to work. The working-age population is the total number of people aged 16 years and over whoare not in jail, hospital, or some other form of institutional care.

The Census Bureau divides the working-age population into two groups: those in the labor force and those not in the labor force. It also divides the labor force into two groups: the employed and the unemployed. So the labor force is the sum of the employed and the unemployed.
To be counted as employed in the Current Population Survey, a person must have either a fulltime jobor a part-time job. To be counted as unemployed, aperson must be available for work and must be in one of three categories:

1. Without work but has made specific efforts tofind a job within the previous four weeks
2. Waiting to be called back to a job from which he or she has been laid off
3. Waiting to start a new job within 30 days

Anyone surveyed who satisfies one of these three criteria is counted as unemployed. People in the working-age population who are neither employed nor unemployed are classified as not in the labor force.
In June 2014, the population of the United States was 318 million; the working-age population was 248 million. Of this number, 92 million were not in the labor force. Most of these people were in school full time or had retired from work. The remaining 156 million people made up the U.S. labor force. Of these, 146.3 million were employed and 9.7 million were unemployed.

## Three Labor Market Indicators

The Census Bureau calculates three indicators of the state of the labor market. They are

- The unemployment rate
- The employment-to-population ratio
- The labor force participation rate

The Unemployment Rate The amount of unemployment is an indicator of the extent to which people who want jobs can't find them. The unemploymentrate is the percentage of the people in the labor forcewho are unemployed. That is,

$$
\text { Unemployment rate }=\frac{\begin{array}{c}
\text { Number of people } \\
\text { unemployed } \tag{100}
\end{array}}{\text { Labor force }} \times
$$

and

$$
\text { Labor force }=\begin{aligned}
& \text { Number of people employed }+ \\
& \text { Number of people unemployed. }
\end{aligned}
$$

In June 2014, the number of people employed was
146.3 million and the number unemployed was 9.7 million. By using the above equations, you can verify that the labor force was 156 million ( 146.3 million plus 9.7 million) and the unemployment rate was 6.2 percent ( 9.7 million divided by 156.0 million, multi- plied by 100).

Figure 22.2 shows the unemployment rate from 1980 to 2014. The average unemployment rate during this period is 6.5 percent-equivalent to 10.1 million people being unemployed in 2014. The unemployment rate fluctuates over the business cycle and reaches a peak
value after a recession ends. Each peak unemployment rate in the recessions of:' 1982, 1990-1991, and 2001 was lower than the previous one. But the recession of 2008-2009 ended the downward trend.

The Employment-to-Population Ratio The number of people of working age who have jobs is an indicator of both the availability of jobs and the degree ofmatch between people's skills and jobs. The employment-to-population ratio is the percentage of people of working age who have jobs. That is,


In June 2014, the number of people employed was 146.3 million and the working-age population was 248 million. By using the above equation, . you can verify that the employment-to-population ratio was 59 percent ( 146.3 million divided by 248 i; million, multiplied by 100).
Figure 22.3 shows the employment-to-population; ratio. This indicator followed an upward trend before 2000 and then a downward trend. The increase before 2000 means that the U.S. economy created jobs at a faster rate than the working-age population grew. This indicator also fluctuates: It falls during a recession and increases during an expansion.

FIGURE 22.2 The Unemployment Rate: 1980-2014


FIGURE 22.3 Labor Force Participation and Employment: 1980-2014


The trend in the labor force participation rate and the employmentto-population ratio is upward before 2000 and downward after 2000.

The employment-topopulation ratio fluctuates more than the labor force participation rate over the business cycle and reflects cyclical fluctuations in the unemployment rate.

The fall in both measures was steep during 2008 and 2009.

The Labor Force Participation Rate The number of people in the labor force is an indicator of the willingness of people of working age to take jobs. The laborforce participation rate is the percentage of the working-age population who are members of the labor force. That is,

$$
\underset{\text { Labor force }}{\text { participation rate }}=\frac{\text { Labor force }}{\begin{array}{c}
\text { Working-age } \\
\text { population }
\end{array}} \times 100 \text {. }
$$

In June 2014, the labor force was 156.0 million and the working-age population was 248.0 million. By using the above equation, you can verify that the labor force participation rate was 62.9 percent ( 156.0 million divided by 248.0 million, multiplied by 100).
Figure 22.3 shows the labor force participation rate. Like the employment-to-population ratio, this indicator has an upward trend before 2000 and then a downward trend. It also has mild fluctuations around the trend but a steep decrease in 2008and 2009. Unsuccessful job seekers left the laborforce during the recession and didn't reenter duringthe weak expansion that began in 2010.

## Other Definitions of Unemployment

Do fluctuations in the labor force participation rate over the business cycle mean that people who leave the labor force during a recession should be counted as unemployed? Or are they correctly counted as being not in the labor force?
The Bureau of Labor Statistics believes that the official unemployment definition gives the correct measure. But it provides data on two types of under-employed labor excluded from the official measure. They are

- Marginally attached workers
- Part-time workers who want full-time jobs

A Marginally Attached Worker is a person who currently is neither working nor looking for work but has indicated that he or shewants and is available for a job and has looked for work
sometime in the recent past. A marginally attached worker who has stopped looking for a job because of repeated failure to find one is called a discouraged worker.
The official unemployment measure excludes marginally attached workers because they haven't made specific efforts to find a job within the past four weeks. In all other respects, they are unemployed.

Part-Time Workers Who Want Full-Time Jobs Many part-time workers want to work part time. This arrangement fits in with the other demands on theirtime. But some part-time workers would like full- time jobs and can't find them. In the official statistics, these workers are called economic part-time workers and they are partly unemployed.

## Most Costly Unemployment

All unemployment is costly, but the most costly is long-term unemployment that results from job loss.
People who are unemployed for a few weeks and then find another job bear some costs of unemploy-ment. But these costs are low compared to the costs borne by people who remain unemployed for many weeks.
Also, people who are unemployed because they voluntarily quit their jobs to find better ones or because they have just entered or reentered the labor market bear some costs of unemployment. But these costs are lower than those borne by people who lose their job and are forced back into the job market.
The unemployment rate doesn't distinguish among these different categories of unemployment. If most of the unemployed are long-term job losers, the situation is much worse than if most are short-term voluntary job searchers.

## Alternative Measures of Unemployment

To provide information about the aspects of unemployment that we've just discussed, the Bureau of Labor Statistics reports sixalternative measures of the unemployment rate: two that are narrower than the official measure and three that are broader. The narrower measures focus on the personal cost of unemploymentand the broader measures focus on assessing the full amount of underemployed labor resources. Figure 22.4 shows these measures from 1994 (the first year for which all six are available) to 2014. U-3 is the official unemployment rate. Long-term unemployment (U-1) and unemployed job losers ( $\mathrm{U}-2$ ) are about 40 percent of the unemployed on average but 60 percent in a deep recession. Adding discouraged workers (U-4) makes very little difference to the unemployment rate, but adding all other marginally attached workers (U-5) adds one percentage point. A big difference is made by adding the economic part- time workers ( $U$ 6). In June 2014, after adding theseworkers the underemployment rate was 12 percent.

FIGURE 22.4 Six Alternative Measures of Unemployment


U-1 are those unemployed for 15 weeks or more, and U-2 are job losers. U-3 is the official unemployment rate. U-4 adds discouraged workers, and U-5 adds all other marginally attached workers. The broadest measure, U-6, adds parttime workers who want full-time jobs. Fluctuations in all the alternafive measures are similar to those in the official measure, $\mathrm{U}_{3} \mathbf{3}_{2}$

## Unemployment and Full Employment

There is always someone without a job who is searching for one, so there is always some unemployment. The key reason is that the economy is a complex mechanism that is always changing-it experiences frictions, structural change, and cycles.

## Frictional Unemployment

There is an unending flow of people into and out of the labor force as people move through the stages of life - from being in school to finding a job, to working, perhaps to becoming unhappy with a job and looking for a new one, and finally, to retiring from fulltime work.
There is also an unending process of job creation and job destruction as new firms are born, firms expand or contract, and some firms fail and go out of business.
The flows into and out of the labor force and the processes of job creation and job destruction create the need for people to search for jobs and for businesses to search for workers. Businesses don't usually hire the first person who applies for a job, and unemployed people don't usually take the first job thatcomes their way. Instead, both firms and workersspend time searching for what they believe will be the best available match. By this process of search, people can match their own skills and interests with the available jobs and find a satisfying job and a good income.

The unemployment that arises from the normal labor turnover we've just described-from people entering and leaving the labor force and from the ongoing creation and destruction of jobs- is calledfrictional unemployment. Frictional unemployment is a permanent and healthy phenomenon in a dynamic, growing economy.

## Structural Unemployment

The unemployment that arises when changes in technology or international competition change the skills needed to perform jobs or change the locations of jobs is called structural unemployment. Structural unemployment usually lasts longer than frictionalunemployment because workers must retrain andpossibly relocate to find a job. When a steel plant in Gary, Indiana, is automated, some jobs in that city disappear. Meanwhile, new jobs for security guards, retail clerks, and life-insurance salespeople are created in Chicago and Indianapolis. The unemployed former steelworkers remain unemployed for several months until they move, retrain, and get one of thesejobs. Structural unemployment is painful, especiallyfor older workers for whom the best available option might be to retire early or take a lower-skilled, lower- paying job.

## Cyclical Unemployment

The higher than normal unemployment at a business cycle trough and the lower than normal unemployment at a business cycle peak is called cyclical unemployment. A worker who is laid off because the economy is in a recession and who gets rehired some months later when the expansion begins has experienced cyclical unemployment.

## "Natural Unemployment"

Natural unemployment is the unemployment that arises from frictions and structural change when thereis no cyclical unemployment-when all the unemployment is frictional and structural. Natural unemployment as a percentage of the labor force is called the natural unemployment rate.

Full Employment is defined as a situation in which the unemployment rate equals the natural unemployment rate.
What determines the natural unemployment rate?
Is it constant or does it change over time?
The natural unemployment rate is influenced by many factors but the most important ones are

- The age distribution of the population
- The scale of structural change
- The real wage rate
- Unemployment benefits

The Age distribution of the Population An economy with a young population has a large number of new job seekers every year and has a high level of frictional unemployment. An economy with an aging population has fewer new job seekers and a low levelof frictional unemployment.

The Scale of Structural Change The scale of structural change is sometimes small. The same jobs usingthe same machines remain in place for many years.
But sometimes there is a technological upheaval. The old ways are swept aside and millions of jobs are lost and the skill to perform them loses value. The amount of structural unemployment fluctuates withthe pace and volume of technological change and the change driven by fierce international competition, especially from fast-changing Asian economies. A high level of structural unemployment is present in many parts of the United States today (as you can see in Economics in Action below).

The Real Wage Rate The natural unemployment rate is influenced by the level of the real wage rate. Real wage rates that bring unemployment are a mini- mum wage and an efficiency wage. Chapter 6 (see pp. 169-171) explains how the minimum wage creates unemployment. An efficiency wage is a wage set above the going market wage to enable firms to attract the most productive workers, get them to work hard, and discourage them from quitting.

Unemployment Benefits Unemployment benefits increase the natural unemployment rate by lowering the opportunity cost of job search. European countries have more generous unemployment benefits and higher natural unemployment rates than the United States. Extending unemployment benefits increases the natural unemployment rate.
There is no controversy about the existence of a natural unemployment rate. Nor is there disagreement that the natural unemployment rate changes.
But economists don't know its exact size or the extent to which it fluctuates. The Congressional Budget Office estimates the natural unemployment rate and its estimate for 2012 was 6 percent-about. 70 percent of the unemployment in that year.

## Real GDP and Unemployment Over the Cycle

The quantity of real GDP at full employment is potential GDP (Chapter 21, p. 536). Over the business cycle, real GDP fluctuates around potential GDP. The gap between real GDP and potential GD 'is called the output gap. As the output gap fluctuates. over the business cycle, the unemployment rate fluctuates around the natural unemployment rate.

## ECONOMICS IN ACTION

## Structural and Cyclical Unemployment in Michigan

In 2010, 13.6 percent of Michigan's labor force was unemployed-the nation's highest official unemployment rate - and when marginally attached workers and part-time workers who want full-time jobs are added, almost 22 percent of the state's labor force was unemployed or underemployed. And 8.4 percent of Michigan's labor force were unemployed for long spells. One of Michigan's problems was structural - a collapse of manufacturing jobs centered on the auto industry. These jobs had been disappearing steadily as robot technologies spread to do ever more of the tasks in the assembly of automobiles. The 2008-2009 recession accelerated this rate of job loss. But by 2014, Michigan's unemployment rate had fallen to 7.5 percent, a fall larger than the fall in the average.

FIGURE 22.5 The Output Gap and the Unemployment Rate

(a) Output gap

(b) Unemployment rate

As real GDP fluctuates around potential GDP in part (a), the unemployment rate fluctuates around the natural unemployment rate in part (b). In recessions, cyclical unemployment increases and the output gap becomes negative. At business cycle peaks, the unemployment rate falls below the natural rate and the output gap becomes positive. The natural unemployment rate decreased during the 1980s and 1990s.

Around 11,000 businesses in Michigan produce high-tech scientific instruments and components for defense equipment, energy plants, andmedical equipment. In2010, these businesses employed around 400,000 people, which was more than 10 percen.t of the state's labor force and two thirds of all manufacturing jobs. Although the recession hit these firms they cut employment by only 10 percent, compared, with a 24 percent cut in manufacturing jobs in the rest of the Michigan economy. And these businesses and)some new ones together with Michigan's traditional) auto industry added jobs at a rapid pace after 2010. By mid-2014, more than 100,000 new manufacturing jobs had been created in Michigan.

## The Price Level, Inflation, and Deflation

What will it really cost you to pay off your studentloan? What will your parents' life savings buy whenthey retire? The answers depend on what happens tothe price level, the average level of prices, and the value of money. A persistently rising price level iscalled inflation; a persistently falling price level is called deflation.
We are interested in the price level, inflation, and deflation for two main reasons. First, we want to measure the annual percentage change of the price level-the inflation rate or deflation rate. Second, wewant to distinguish between the money values and real values of economic variables such as your student loan and your parents' savings.
We begin by explaining why inflation and deflation are problems. Then we'll look at how we mea- sure the price level and the inflation rate. Finally, we'll return to the task of distinguishing real values from moneyvalues.

Why Inflation and deflation Are Problems Low, steady, and anticipated inflation or deflation isn't a problem, but an unexpected burst of inflationor period of deflation brings big problems and costs.
An unexpected inflation or deflation

- Redistributes income
- Redistributes wealth
- Lowers real GDP and employment
- Diverts resources from production

Redistributes Income Workers and employers sign wage contracts that last for a year or more. An unexpected burst of inflation raises prices but doesn't immediately raise the wages. Workers are worse off because their wages buy less than they bargained forand employers are better off because their profits rise.
An unexpected period of deflation has the opposite effect. Wage rates don't fall but the prices fall. Workers are better off because their fixed wages buy more than they bargained for and employers are worse off with lower profits.

Redistributes Wealth People enter into loan contracts that are fixed in money terms and that pay an interestrate agreed as a percentage of the money borrowed and lent. With an unexpected burst of inflation, the money that the borrower repays to the lender buysless than the money originally loaned. The borrowerwins and the lender loses. The interest paid
on the loan doesn't compensate the lender for the loss in the value of the money loaned. With an unexpected deflation, the money that the borrower repays to the lender buys more than the money originally loaned. The borrower loses and the lender wins.

Lowers Real GDP and Employment Unexpected inflation that raises firms' profits brings a rise in investment and a boom in production and employment. Real GDP rises above potential GDP and theunemployment rate falls below the natural rate. But this situation is temporary. Profitable investment dries up, spending falls, real GDP falls below potential GDP and the unemployment rate rises. Avoiding these swings in production and jobs means avoiding unexpected swings in the inflation rate.
An unexpected deflation has even greater consequences for real GDP and jobs. Businesses and households that are in debt (borrowers) are worse off\}and they cut their spending. A fall in total spending 'brings a recession and rising unemployment.

Diverts Resources from Production Unpredictable inflation or deflation turns the economy into a casino and diverts resources from productive activities to forecasting inflation. It can become more profitable to forecast the inflation rate or deflation rate correctly,, than to invent a new product. Doctors, lawyers, accountants, farmers-just about everyone-can make themselves better off, not by specializing in the profession for which they have been trained but by spending more of their time dabbling as amateur economists and inflation forecasters and managing their investments.
From a social perspective, the diversion of talent that results from unpredictable inflation is like throwing scarce resources onto a pile of garbage. This waste of resources is a cost of inflation.
At its worst, inflation becomes hyperinflation-an inflation rate of 50 percent a month or higher that grinds the economy to a halt and causes a society to collapse. Hyperinflation is rare, but Zimbabwe in recent years and several European and Latin American countries have experiencedit.
We pay close attention to the inflation rate, even when its rate is low, to avoid its consequences. We monitor the price level every month and devote considerable resources to measuring it accurately. You're now going to see how we do this.

## The consumer Price Index

Every month, the Bureau of Labor Statistics (BLS)measures the price level by calculating the Consumer Price Index (CPI), which is a measure of the average of the prices paid by urban consumers for a fixed basketof consumer goods and services. What you learn herewill help you to make sense of the CPI and relate it to your own economic life. The CPI tells you about the value of the money in your pocket.

## Reading the CPI Numbers

The CPI is defined to equal 100 for a period called the reference base period. Currently, the reference base period is 1982-1984. That is, for the average of the 36 months from January 1982 through December 1984, the CPI equals 100.

In June 2014, the CPI was 237.7. This number tells us that the average of the prices paid by urban consumers for a fixed market basket of consumer goods and services was 137.7 percent higher in June2014 than it was on average during 1982-1984.

## Constructing the CPI

Constructing the CPI involves three stages:

- Selecting the CPI basket
- Conducting the monthly price survey
- Calculating the CPI

The CPI Basket The first stage in constructing the CPI is to select what is called the CPI basket. This basket contains the goods and services represented inthe index, each weighted by its relative importance. The idea is to make the relative importance of the items in the CPI basket the same as that in thebudget of an average urban household. For example, because people spend more on housing than on bus rides, the CPI places more weight on the price of housing than on the price of a bus ride.
To determine the CPI basket, the BLS conducts a Consumer Expenditure Survey. Today's CPI basket is based on data gathered in the Consumer Expenditure Survey of 2012-2013. Figure 22.6 shows the CPI basket. As you look at the relative importance of the items in the CPI basket, remember that it applies to the average household. Individual households' baskets are spread around the average. Think about what you buy and compare your basket with the CPI basket.

The Monthly Price Survey Each month, BLS employees check the prices of the 80,000 goods and services in the CPI basket in 30 metropolitan areas. Because the CPI aims to measure price changes, it is important that the prices recorded each month refer to exactly the same item. For example, suppose the price of a box of jelly beans has increased but a box now contains more beans. Has the price of jelly beans increased? The BLS employee must record the details of changes in quality or packaging so that price changes can be isolated from other changes. Once the raw price data are in hand, the next task is to calculate the CPI.

## Calculating the CPI To calculate the CPI, we

- Find the cost of the CPI basket at base-period prices.
- Find the cost of the CPI basket at current-period prices.
- Calculate the CPI for the base period and the current period.

We'll work through these three steps for the simple artificial economy in Table 22.1, which shows the quantities in the CPI basket and the prices in the base period (2014) and current period (2015).
Part (a) contains the data for the base period. In that period, consumers bought 10 oranges at $\$ 1$ eachand 5 haircuts at $\$ 8$ each. To find the cost of the CPIbasket in the base-period prices, multiply the quantities in the CPI basket by the base-period prices. Thecost of oranges is $\$ 10$ ( 10 at $\$ 1$ each), and the cost of haircuts is $\$ 40$ (5 at $\$ 8$ each). So the total cost
of the CPI basket in the base period at base-period prices is $\$ 50(\$ 10+\$ 40)$. Part (b) contains the price data for the current period. The price of an orange increased from $\$ 1$ to $\$ 2$, which is a 100 percent increase- $(\$ 1+\$ 1) \times 100=100$. The price of a haircut increased from $\$ 8$ to $\$ 10$, which is a 25 percent increase $-(\$ 2+\$ 8) X 100=25$.
The CPI provides a way of averaging these price increases by comparing the cost of the basket rather than the price of each item. To find the cost of the CPI basket in the current period, 2015, multiply the quantities in the basket by their 2015 prices. The cost of oranges is $\$ 20$ (IO at $\$ 2$ each), and the cost of haircuts is $\$ 50$ ( 5 at $\$ 10$ each). So total cost of the fixed CPI basket at current-period prices is $\$ 70(\$ 20+\$ 50)$.

FIGURE 22.6 The CPI Basket


The CPI basket consists of the items that an average urban household buys. It consists mainly of housing ( 41.4 percent), transportation ( 16.4 percent), and food and beverages ( 14.9 percent). All other items sum to 27.3 percent of the total.

You've now taken the first two steps toward calculating the CPI: calculating the cost of the CPI basketin the base period and the current period. The thirdstep uses the numbers you've just calculated to find the CPI for 2014 and 2015. The formula for the CPI is

$$
\mathrm{CPI}=\frac{\begin{array}{c}
\text { Cost of CPI basket at } \\
\text { current prices }
\end{array}}{\begin{array}{c}
\text { Cost of CPI basket at } \\
\text { base-period prices }
\end{array}} \times 100 .
$$

| TABLE 22.1 The CPI: <br> A Simplified Calculation |  |  |  |
| :---: | :---: | :---: | :---: |
| (a) The cost of the CPI basket at base-period prices: 2014 |  |  |  |
| CPI basket |  |  | Cost of |
| Item | Quantity | Price | CPI Basket |
| Oranges | 10 | \$1.00 | \$10 |
| Haircuts | 5 | \$8.00 | \$40 |
| Cost of CPI basket at base-period prices |  |  | \$50 |

(b) The cost of the CPI basket at current-period prices: 2015

| CPI basket |  |  | Cost of <br> Item |
| :--- | :---: | :---: | :---: |
| Quantity | Price | Basket |  |

In Table 22.1, you established that in 2014 (the base period), the cost of the CPI basket was $\$ 50$ and in 2015, it was $\$ 70$. If we use these numbers in the CPIformula, we can find the CPI for 2014 and 2015. For 2014, the CPI is

$$
\mathrm{CPI} \text { in } 2014=\frac{\$ 50}{\$ 50} \times 100=100
$$

For 2015, the CPI is

$$
\mathrm{CPI} \text { in } 2015=\frac{\$ 70}{\$ 50} \times 100=140
$$

The principles that you've applied in this simplified CPI calculation apply to the more complex calculations performed every month by the BLS.

## Measuring the Inflation Rate

A major purpose of the CPI is to measure changes in the cost of living and in the value of money. To measure these changes, we calculate the inflation rate as the annual percentage change in the CPI. To calculate the inflation rate, we use the formula:

$$
\underset{\text { rate }}{\text { Inflation }}=\frac{\text { CPI this year }- \text { CPI last year }}{\text { CPI last year }} \times 100 .
$$

We can use this formula to calculate the inflation rate in 2014. The CPI in June 2014 was 237.7, andthe CPI in June 2013 was 232.9. So the inflation rate during the 12 months to June 2014 was

$$
\underset{\text { rate }}{\text { Inflation }}=\frac{(237.7-232.9)}{232.9} \times 100=2.1 \% .
$$

## Distinguishing High Inflation from a High Price Level

Distinguishing High Inflation from a High Price Level
Figure 22.7 shows the CPI and the inflation rate in the United States between 1970 and 2014. The two parts of the figure are related and emphasize the distinction between high inflation and high prices.
When the price level in part (a) rises rapidly, (I970 through 1982), the inflation rate in part (b) is high. When the price level in part (a) rises slowly, (after 1982), the inflation rate in part (b) is low.
A high inflation rate means that the price level is rising rapidly. A high price level means that there has been a sustained period of rising prices.
When the price level in part (a) falls (2009), the inflation rate in part (b) is negativedeflation.
The CPI is not a perfect measure of the price level and changes in the CPI probably overstate the inflation rate. Let's look at the sources of bias.

## The Biased CPI

The main sources of bias in the CPI are

- New goods bias
- Quality change bias
- Commodity substitution bias
- Outlet substitution bias

New Goods Bias If you want to compare the price level in 2014 with that in 1970, you must somehow compare the price of a computer today with that of a typewriter in 1970. Because a PC is more expensive than a typewriter was, the arrival of the PC puts an upward bias into the CPI and its inflation rate.

Quality Change Bias Cars and many other goods get better every year. Part of the rise in the prices of thesegoods is a payment for improved quality and is not inflation. But the CPI counts the entire price rise asinflation and so overstates inflation.

FIGURE 22.7 The CPI and the Inflation Rate

(a) CPI

(b) Inflation rate

When the price level rises rapidly, the inflation rate is high; when the price level rises slowly, the inflation rate is low. When the price level falls, the inflation rate is negative.

From 1970 through 1982, the price level increased rapidly in part (a) and the inflation rate was high in part (b). After 1982, the price level rose slowly in part (a) and the inflation rate was low in part (b). In 2009, the price level fell and the inflation rate was negative-there was deflation.

Commodity Substitution Bias Changes in relative prices lead consumers to change the items they buy. For example, if the price of beef rises and the price of chicken remains unchanged, people buy more chicken and less beef This switch from beef to chicken might provide the same amount of meat and the same enjoyment as before and expenditure is the same as before. The price of meat has not changed. But because the CPI ignores the substitution of chicken for beef, it says the price of meat has increased.

Outlet Substitution Bias When confronted with higher prices, people use discount stores more frequently and convenience stores less frequently. Thisphenomenon is called outlet substitution. The CPI surveys do not monitor outlet substitutions.

## TheMagnitude oftheBias

You've reviewed the sources of bias in the CPL But how big is the bias? This question was tackled in1996 by a Congressional Advisory Commission on the Consumer Price Index chaired by Michael Boskin, an economics professor at Stanford University. This commission said that the CPI overstates inflation by 1.1 percentage points a year. That is, if the CPI reports that inflation is 3.1 per- cent a year, most likely inflation is actually 2 percent a year.

## Some Consequences of the Bias

The bias in the CPI distorts private contracts and increases government outlays. Many private agreements, such as wage contracts, are linked to the CPL For example, a firm and its workers might agree to a three-year wage deal that increases the wage rate by 2 percent a year plus the percentage increase in the CPL Such a deal ends up giving the workers more real income than the firm intended.
Close to a third of federal government outlays, including Social Security checks, are linked directly to the CPL And while a bias of 1 percent a yearseems small, accumulated over a decade it adds up to almost a trillion dollars of additional expenditures.

## Alternative Price Indexes

The CPI is just one of many alternative price levelindex numbers and because of the bias in the CPI;other measures are used for some purposes. We'lldescribe three alternatives to the CPI and explain when and why they might be preferred to the CPLThe alternatives are

- Chained CPI
- Personal consumption expenditure deflator
- GDP deflator

Chained CPI The chained CPI is a price index that is calculated using a similar method to that used to calculate chained-dollar real GDP described in Chapter 21 (see pp. 546-547).
The chained CPI overcomes the sources of bias in the CPL It incorporates substitutions and new goods bias by using current and previous period quantities rather than fixed quantities from an earlier period. The chained CPI measures a lower inflation rate, on average, than the standard CPL Between 2000 and 2014, the average inflation rate as measured by . the chained CPI is only 0.7 percentage points lower than the standard CPI-1.7 percent versus 2.4 percent per year.

Personal Consumption Expenditure Deflator The personal consumption expenditure deflator (or PCB deflator) is calculated from data in the national income accounts that you studied in Chapter 21. When the Bureau of Economic Analysis calculates real GDP, it also calculates the real values of its expenditure components: real consumption expenditure, real investment, real government expenditure, and real net exports. These calculations are done in the same way as that for real GDP described in simplified terms on p. 535 and more technically on pp. 546-547 in Chapter 21.
To calculate the PCE deflator, we use the formula:

$$
\text { PCE deflator }=(\text { Nominal } C \div \text { Real } C) \times 100 \text {, }
$$

where $C$ is personal consumption expenditure.
The basket of goods and services included in the PCE deflator is broader than that in the CPI because it includes all consumption expenditure, not only the items bought by a typical urban family. The difference between the PCE deflator and the CPI is small. Since 2000, the inflation rate measured by the PCE deflator is 1.9 percent per year, 0.5 percentage points lower than the CPI inflation rate.

GDP Deflator The GDP deflator is a bit like the PCE deflator except that it includes all the goods and services that are counted as part of GDP. So it is an index of the prices of the items in consumption, investment, government expenditure, and net exports

$$
\underset{\text { deflator }}{\mathrm{GDP}}=(\text { Nominal GDP } \div \text { Real GDP }) \times 100
$$

This broader price index is appropriate for macroeconomics because it is a comprehensive measure of the cost of the real GDP basket of goods and services. Since 2000, the GDP deflator has increased and average rate of 2.0 percent per year, 0.4 percentage :; points below the CPI inflation rate.

## Core Inflation

No matter whether we calculate the inflation rate using the CPI, the chained CPI, the PCE deflator, or the GDP deflator, the number bounces around a good deal from month to month or quarter to quarter. To determine the trend in the inflation rate, we need to strip the raw numbers of their volatility. The core inflation rate is a measure of the inflation rate that excludes volatile prices in an attempt to reveal the underlying inflation trend. (The inflation rate that includes all prices is called the headline inflation rate.)
As a practical matter, the core inflation rate is calculated as the percentage change in a price indexexcluding the prices of food and fuel. The prices of these two items are among the most volatile.
While the core PCE inflation rate removesthe volatile elements in inflation, it can give a misleading view of the true underlying inflation rate. If the relative prices of the excluded
items are changing, the core PCE inflation rate will give a biased measure of the true underlying inflation rate.
Such a misleading account was given during the years between 2003 and 2008 when the relative prices of food and fuel were rising. The result was a core inflation rate that was systematically below the headline inflation rate.
Figure 22.8 graphs the core and headline inflation rates since 2000 and shows how core inflation removes the extreme swings in the headline rate.

## The Real Variables in Macroeconomics

Yousaw in Chapter 21 how we measure real GDP. And you've seen in this chapter how we can use nominal GDP and real GDP to provide another measure of the price level-the GDP deflator. But viewing real GDP as nominal GDP deflated, opens up the idea of other real variables. By using the GDP deflator, we can deflate other nominal variables to find their real values. For example, the real wage rate is the nominal wage rate divided by the GDP deflator. We can adjust any nominal quantity or price variable for inflation by deflating it by dividing it by the price level.
There is one variable that is a bit different-an interest rate. A real interest rate is not a nominal interest rate divided by the price level. You'll learn how to adjust the nominal interest rate for inflation to find the real interest rate in Chapter 24. But all the other real variables of macroeconomics are calculated by dividing a nominal variable by the price level.

FIGURE 22.8 Core Inflation


The core inflation rate excludes volatile price changes of food and fuel. Since 2003, the core inflation rate has mostly been below the CPI inflation rate because the relative prices of food and fuel have been rising.

## ECONOMIC ANALYSIS

- This news article reports and comments on some labor market data for July 2014.
- The 209,000 jobs added during July 2014 are measured by a survey of payroll jobs at business establishments, called the Current Employment Survey (CES).
- The CES measures the change in the number of non-farm jobs.
- The CES is different from the Current Population Survey (CPS), a survey of households described on p. 517, which measures the number of people with a job-the number employed.
- In July 2014, the CPS reported that employment increased by 131,000.
- Although the two surveys can give a conflicting account in a single month, as they did in July 2014, over a longer period, they give the same message.
- To lower unemployment and to increase the labor force participation rate, the number of jobs must increase by more than the increase in population, which for the population (aged 16 and over) is about 200,000 per month.
- Figure 1 shows the change in the number of jobs the blue curve) and the change in the population (the red curve) over the year from July 2013 to July 2014.
- You can see that in most months, the number of jobs created exeeded the increase in population.
- Figure 2 shows how job creation has changed the unemployment rate. With the exceptions of February and July 2014, the unemployment rate has fallen every month from 7.3 percent in July 2013 to 6.2 percent in July 2014.
- Figure 3 shows how job creation has changed the labor force participation rate, which fell from July to November 2013, but then increased through March 2014 before falling again.
- Look closely at the numbers on the $y$-axis of Fig. 3. The labor force participation rate changed by a very small amount from 63.4 percent in July 2013 to 62.9 percent in July 2014.
- To create enough jobs to employ the growing labor force plus the unemployed and the underemployed and to bring others back into the labor force, the pace of job creation will need to increase to beyond its level during 2014.


Figure 1 Jobs Created


Figure 2 Unemployment Rate


Figure 3 Labor Force Participation Rate

## PART EIGHT: MACROECONOMIC TRENDS

## CHAPTER 23: Economic Growth

After studying this chapter, you will be able to:

- Define and calculate the economic growth rate and explain the implications of sustained growth
- Describe the economic growth trends in the United States and other countries and regions
- Explain what makes potential GDP grow
- Explain the sources of labor productivity growth
- Explain the theories of economic growth and policies to increase its rate
U.S.real GDP per person and the standard living tripled between 1964 and 2014. We see even more dramatic change in China, where incomes havetripled not in 50 years but in the 14 years since 2000. Incomes are also growing rapidly in some African economies, one of which is the small but dynamic Botswana. In this chapter, we study the forces that make real GDP grow; and in Economics in the News at the end of the chapter, we look at lessons we can learn from the slow growth of South Africa and its fast growingneighbor, Botswana.


## The Basics of Economic Growth

Economic growth is the expansion of production possibilities. A rapid pace of economic growth maintained over a number of years can transform a poor nation into a rich one. Such have been the stories of Hong Kong, South Korea, and some other Asian economies.Slow economic growth or the absence of growth can condemn a nation to devastating poverty. Such has been the fate of Sierra Leone, Somalia, Zambia, and much of the rest of Africa. The goal of this chapter is to help you to understand why some economies expand rapidly and others stagnate. We'll begin by learning how to calculate a growth rate, by distinguishing between economic growth and a business cycle expansion, and by discovering the magic of sustained growth.

## Calculating Growth Rates

We express a growth rate as the annual percentage change of a variable - the change in the level expressed as a percentage of the initial level. The growth rate of real GDP, for example, is calculated as:

$$
\begin{aligned}
& \begin{array}{l}
\text { Real GDP } \\
\text { growth rate }
\end{array}=\frac{\begin{array}{c}
\text { Real GDP } \\
\text { in current year }
\end{array}-\begin{array}{c}
\text { Real GDP } \\
\text { in previous year }
\end{array}}{\text { Real GDP in previous year }} \times 100 .
\end{aligned}
$$

Using some numbers, if real GDP in the current year is $\$ 11$ trillion and if real GDP in the previous year was $\$ 10$ trillion, then the economic growth rate is 10 percent.

The growth rate of real GDP tells us how rapidly the total economy is expanding. This measure is useful for telling us about potential changes in the balance of economic power among nations. But it does not tell us about changes in the standard of living. The standard of living depends on real GDP per person (also called per capita real GDP), which is real GDP divided by the population. So the contribution of real GDP growth to the change in the standard of living depends on the growth rate of real GDP per person. We use the above formula to calculate this growthrate, replacing real GDP with real GDP per person.
Suppose, for example, that in the current year, when real GDP is $\$ 11$ trillion, the population is 202 million. Then real GDP per person is $\$ 11$ trilliondivided by 202 million, which equals $\$ 54,455$. And suppose that in the previous year, when real GDP was $\$ 10$ trillion, the population was 200 million. Then real GDP per person in that year was $\$ 10$ trillion divided by 200 million, which equals $\$ 50,000$.
Use these two values of real GDP per person within the growth formula above to calculate the growth rate of real GDP per person. That is,

$$
\begin{aligned}
& \begin{array}{l}
\text { Real GDP } \\
\text { per person } \\
\text { growth rate }
\end{array}
\end{aligned}=\frac{\$ 54,455-\$ 50,000}{\$ 50,000} \times 100=8.9
$$

The growth rate of real GDP per person can also; be calculated (approximately) by subtracting the population growth rate from the real GDP growth rate. In the example you've just worked through, the growth rate of real GDP is 10 percent. The population changes from 200 million to 202 million, so the population growth rate is 1 percent. The growth rate of real GDP per person is approximately equal to 10 percent minus 1 percent, which equals 9 percent. Real GDP per person grows only if real GDP grows faster than the population grows. If the growth of rate of the population exceeds the growth rate of real' GDP, then real GDP per person falls.

## Economic Growth Versus Business Cycle expansion

Real GDP can increase for two distinct reasons: Theyeconomy might be returning to full employment in an expansion phase of the business cycle or potential GDP might be increasing. The return to full employment in an expansion phase of the business cycle isn't economic growth. It is just taking up the slack that resulted from the previous recession. The expansion of potential GDP is economic growth.
Figure 23.1 illustrates this distinction using the production possibilities frontier (the PPP that you studied in Chapter 2). A return to full employment in a business cycle expansion is a movement from inside the $P P P$ at a point such as $A$ to a point on the $P P P$ such as $B$. Economic growth is the expansion of production possibilities. It is an outward movement of the PPP such as the shift from $P P P_{0}$ to $P P P_{1}$ and the movement from point $B$ on $P P P_{0}$ to point $C$ on $P P P_{1}$. The growth rate of potential GDP measures the pace of expansion of production possibilities and smoothes out the business cycle fluctuations in the growth rate of real GDP.
figure 23.1 Economic Growth and a Business Cycle Expansion


The increase in aggregate production in the move from point $A$ inside $P P F_{0}$ to point $B$ on $P P F_{0}$ is an expansion phase of the business cycle and it occurs with no change in production possibilities. Such an expansion is not economic growth. The increase in aggregate production in the move from point $B$ on PPFF to point $C$ on PPF $_{1}$ is economic growth-an expansion of production possibilities shown by an outward shift of the PPF.

Figure 23.2 shows how the growth rate of potential GDP (red curve) smoothes the more erratic fluctuations in the growth rate of real GDP. Business cycle fluctuations in the real GDP growth rate mask the underlying trend growth rate revealed by the growth rate of potential GDP.

## The Magic of Sustained Growth

Sustained growth of real GDP per person can transform a poor society into a wealthy one. The reason is that economic growth is like compound interest.

Compound Interest Suppose that you put $\$ 100$ in the bank and earn 5 percent a year interest on it. After one year, you have $\$ 105$. If you leave that $\$ 105$ in the bank for another year, you earn 5 percent interest on the original $\$ 100$ and on the $\$ 5$ interest that you earned last year. You are now earning interest on interest! The next year, things get even better.

Then you earn 5 percent on the original $\$ 100$ and on the interest earned in the first year and the second year. You are even earning interest on the interest that you earned on the interest of the first year.

FIGURE 23.2 Growth Rates of Real GDP and Potential GDP


The annual growth rate of real GDP fluctuates widely over the business cycle and masks changes in the underlying trend growth rate. The annual growth rate of potential GDP provides information about changes in the trend growth rate. Both the growth rate of potential GDP and the trend growth rate of real GDP have fallen since 2000.

Your money in the bank is growing at a rate of 5 percent a year. Before too many years have passed, your initial deposit of $\$ 100$ will have grown to $\$ 200$. But after how many years?
The answer is provided by a formula called the Rule of 70 , which states that the number of years it takes for the level of any variable to double is approximately 70 divided by the annual percentage growth rate of the variable. Using the Rule of 70, you can now calculate how many years it takes your $\$ 100$ to become $\$ 200$. It is 70 divided by 5 , which is 14 years.

FIGURE 23.3 The Rule of 70


## Applying the Rule of 70

The Rule of 70 applies to any variable, so it applies to real GDP per person. Figure 23.3 shows the doubling time for growth rates of 1 percent per year to 12 per- cent peryear. You can see that real GDP per person doubles in 70 years ( 70 divided by 1 )-an average human life span-if the growth rate is 1 percent a year. It doubles in 35 years if the growth rate is 2 percent a year and in just 10 years if the growth rate is 7 percent a year. We can use the Rule of 70 to answer other questions about economic growth. For example, in 2010 U.S. real GDP per person was approximately 4 times that of China. China's recent growth rate of real GDP per person was 10 percent a year. If this growth rate were maintained, how long would it take China's real GDP per person to reach that of the United States in 2010? The answer, provided by the Rule of 70, is 14 years. China's real GDP per person doubles in 7 years ( 70 divided by 10). It doubles again to 4 times its 2010 level in another 7 years. So after 14 years of growth at 10 percent a year, China's real GDP per person is 4 times its 2010 level and equals that of the United States in 2010. Of course, after 14 years, U.S. real GDP per person would have increased, so China would still not have caught up to the United States. But at these growth rates, China's real GDP per person would equal that of the United States in 2010 by 2024.

## Long-Term Growth trends

You have just seen the power of economic growth to increase incomes. At a 1 percent growth rate, it takes a human life span to double the standard of living. But at a 7 percent growth rate, the standard of living doubles every decade. How fast has our economy grown over the long term? How fast are other economies growing? Are poor countries catching up to rich ones, or do the gaps between the rich and poor persist or evenwiden?

## Long-Term Growth in the U.S. Economy

Figure 23.4 shows real GDP per person in the United States for the hundred years from 1914 to 2014. The thick line is actual real GDP and the thin line (that starts in 1949) is potential GDP. The trend in potential GDP tells us about economic growth. Fluctuations around potential GDP tell us about the business cycle.
Two extraordinary events dominate the graph: the Great Depression of the 1930s, when growth stopped for a decade, and World War II of the 1940s, when growth briefly exploded.
For the century as a whole, the average growth rate was 2 percent a year. But the growth rate has notremained constant. From 1910 to the onset of the Great Depression in 1929, the average growth rate was a bit lower than the century average at 1.8 percent a year. Between 1930 and 1950, averaging out the GreatDepression and World War II, the growth rate was 2.4 percent a year. After World War II, the growth rate started out at 2 percent a year. It then increased and growth averaged 3 percent a year during the 1960s. In1973, and lasting for a decade, the growth rate slowed.Growth picked up somewhat during the 1980s and even more during the 1990s dot.com expansion. But the growth rate never returned to the pace achieved during the fast-growing 1960s.
A major goal of this chapter is to explain why our economy grows and why the growth rate changes.
Another goal is to explain variations in the economic growth rate across countries. Let's now look at some other countries' growth rates.
figure 23.4 A Hundred Years of Economic Growth in the United States


## Real GDP Growth in the World Economy

Figure 23.5 shows real GDP per person in the United States and in other countries between 1960 and 2010. Part (a) looks at the seven richest countries known as the G7 nations. Among these nations, the United States has the highest real GDP per person. In 2010,

Canada had the second-highest real GDP per person, ahead of Japan and France, Germany, Italy, and the United Kingdom (collectively the Europe Big 4).
During the fifty years shown here, the gaps between the United States, Canada, and the Europe Big 4 have been almost constant. But starting from a long way below, Japan grew fastest. It caught up to Europe in 1970 and to Canada in 1990. But duringthe 1990s, Japan's economy stagnated. Many other countries are growing more slowly than and falling farther behind, the United States. Figure 23.5(b) looks at some of these countries.
Real GDP per person in Central and South America was 28 percent of the U.S. level in 1960. It grew more quickly than the United States and reached 30 percent of the U.S. level by 1980, but then growth slowed and by 2010, real GDP per person in these countries was 23 percent of the U.S. level. In Eastern Europe, real GDP per person has grown more slowly than anywhere except Africa, and fell from 32 percent of the U.S. level in 1980 to 19 percent in 2003 and then increased again to 22 percent in 2010. Real GDP per person in Africa, the world's poorest continent, fell from 10 percent of the U.S. level in 1960 to 5 percent in 2007 and then increased slightly to 6 percent in 2010.
Even modest differences in economic growth rates sustained over a number of years bring enormous differences in the standard of living. And some of the differences that you've just seen are enormous. So thefacts about economic growth in the United States and around the world raise some big questions.
What are the preconditions for economic growth? What sustains economic growth once it gets going? How can we identify the sources of economic growth and measure the contribution that each source makes? What can we do to increase the sustainable rate of economicgrowth?
We're now going to address these questions and discover the causes of economic growth. We start by seeing how potential GDP is determined and what makes it grow. You will see that labor productivity growth is the key to rising living standards and go on to explore the sources of this growth.

## ECONOMICS IN ACTION

## Fast Trains on the Same Track

Five Asian economies, Hong Kong, Korea, Singapore, Taiwan, and China, have experienced spectacular growth, which you can see in the figure. During the 1960s, realGDP per person in these economies ranged from 3 to 28 percent of that in the United States. But by 2010, real GDP per person in Singapore and Hong Kong had surpassed that of the United States.
The figure also shows that China is catching up rapidly but from a long way behind. China's real GDP per person increased from 3 percent of the U.S. level in 1960 to 26 percent in 2010.

The Asian economies shown here are like fast trains running on the same track at similar speeds and with a roughly constant gap between them. Singapore and Hong Kong are hooked together as the lead train, which runs about 20 years in front of Taiwan and Korea and 40 years in front of China.
Real GDP per person in Korea in 2010 was similar to that in Hong Kong in 1988, and real GDP in China in 2010 was similar to that of Hong Kong in 1976. Between 1976 and 2010,

Hong Kong transformed itself from a poor developing economy into one of the richest economies in the world.
The rest of China is now doing what Hong Kong has done. China has a population 200 times that of Hong Kong and more than 4 times that of theUnited States. So if China continues its rapid growth, the world economy will change dramatically. As these fast-growing Asian economies catch up with the United States, we can expect their growthrates to slow. But it will be surprising if China's growth rate slows much before it has closed the gapon the United States.

FIGURE 23.5 Economic Growth Around the World: Catch-Up or Not?

(a) Catch-up?

Real GDP per person has grown throughout the world. Among the rich industrial countries in part (a), real GDP per person has grown slightly faster in the United States than in Canada and the four big countries of Europe (France, Germany, Italy, and the United Kingdom). Japan had the fastest growth rate before 1973 but then growth slowed and Japan's economy stagnated during the 1990s.

(b) No catch-up?

Among a wider range of countries shown in part (b), growth rates have been lower than that of the United States. The gaps between the real GDP per person in the United States and in these countries have widened. The gap between the real GDP per person in the United States and Africa has widened by a large amount.

## How Potential GDP Grows

Economic growth occurs when real GDP increases. But a one-shot rise in real GDP or a recovery from recession isn't economic growth. Economic growth is asustained, year-afteryear increase in potential GDP. So what determines potential GDP and what are the forces that make it grow?
Labor, capital, land, and entrepreneurship produce real GDP, and the productivity of the factors of production determines the quantity of real GDP thatcan be produced. The quantity of land is fixed and on any given day, the quantities of entrepreneurial ability and capital are also fixed and their productivities are given. Thequantity of labor employed is the only variable factor of production. Potential GDP is the level of real GDP when the quantity of labor employed is the full-employment quantity.


To determine potential GDP, we use a model with two components:

- An aggregate production function
- An aggregate labor market

Aggregate Production Function When you studied the limits to production in Chapter 2 (see p. 70), you learned that the production possibilities.frontier is the boundary between the combinations of goods and services that can be produced and those that cannot.
We're now going to think about the production possibilities frontier for two special "goods": real GDP and the quantity of leisure time. Think of real GDP as a number of big shopping carts. Each cart contains some of each kind of different goods and services produced, and one cartload of items costs $\$ 1$ trillion. To say that real GDP is $\$ 13$ trillion means that it is 13 very big shopping carts of goods and services.
The quantity of leisure time is the number of hours spent not working. Each leisure hour could be spent working. If we spent all our time taking leisure, we would do no work and produce nothing. Real GDP would be zero. The more leisure we forgo, the greater is the quantity of labor we supply and the greater is the quantity of real GDP produced.
But labor hours are not all equally productive. We use our most productive hours first, and as more hours are worked, these hours are increasingly less productive. So for each additional hour of leisure forgone(each additional hour of labor), real GDP increases but by successively smaller amounts.

The aggregate production function is the relationship that tells us how real GDP changes as the quantity of labor changes when all other influences on production remain the same. Figure 23.6 on page 584 shows this relationship - the curve labeled PF. An increase in thequantity of labor (and a corresponding decrease in leisure hours) brings a movement along the production function and an increase in real GDP.

Aggregate Labor Market In macroeconomics, we pretend that there is one large labor market that determines the quantity of labor employed and the quantity of real GDP produced. To see how this aggregate labor market works, we study the demand for labor, the supply of labor, and labor market equilibrium.

The Demand for Labor The demand for labor is the relationship between the quantity of labor demandedand the real wage rate. The quantity of labor demanded is the number of labor hours hired by all the firms in the economy during a given period. This quantity depends on the price of labor, which is the real wage rate.

The real wage rate is the money wage rate divided by the price level. The real wage rate is the quantity of goods and services that an hour of labor earns. It contrasts with the money wage rate, which is the number of dollars that an hour of labor earns.
The real wage rate influences the quantity of labor demanded because what matters to firms is not the number of dollars they pay (money wagerate) but how much output they must sell to earnthose dollars.
The quantity of labor demanded increases as the real wage rate decreases-the demand for labor curveslopes downward. Why? The answer lies in the shape of the production function. You've seen that along the production function, each additional hour of labor increases real GDP by successively smaller amounts. This tendency has a name: the law of diminishing returns. Because of diminishing returns, firms will hire more labor only if the real wage rate falls to match the fall in the extra output produced by that labor.

TheSupply of Labor The supply oflabor is the relation- ship between the quantity of labor supplied and the real wage rate. The quantity of labor supplied is the number of labor hours that all the households in the economy plan to work during a given period. This quantity depends on the real wage rate. The real wage rate influences the quantity of labor supplied because what matters to households is not the number of dollars they earn (money wage rate) but what they can buy with those dollars. The quantity of labor supplied increases as the real wage rate increases-the supply of labor curve slopes upward. At a higher real wage rate, more people choose to work and more people choose to work longer hours if they can earn more per hour.

Labor Market Equilibrium The price of labor is the real wage rate. The forces of supply and demand operate in labor markets just as they do in the markets for goods and services to eliminate a shortage or a surplus. But a shortage or a surplus of labor brings only a gradual change in the real wage rate. If there is a shortage of labor, the real wage rate rises to eliminate it; and if there is a surplus of labor, the real wage rate eventually falls to eliminate
it. When there is neither a shortage nor a surplus, the labor market is in equilibrium-fullemployment equilibrium.
Figure 23.7 illustrates labor market equilibrium. The demand for labor curve is $L D$ and the supply of labor curve is LS. This labor market is in equilibrium at a real wage rate of $\$ 35$ an hour and 200 billion hours a year areemployed.
If the real wage rate exceeds $\$ 35$ an hour, the quantity of labor supplied exceeds the quantity demanded and there is a surplus of labor. When there is a surplusof labor, the real wage rate falls toward the equilibrium real wage rate where the surplus is eliminated.
If the real wage rate is less than $\$ 35$ an hour, the quantity of labor demanded exceeds the quantitysupplied and there is a shortage of labor. When there is a shortage of labor, the real wage rate rises toward the equilibrium real wage rate where the shortage is eliminated.

FIGURE 23.7 Labor Market Equilibrium


Labor market equilibrium occurs when the quantity of labor demanded equals the quantity of labor supplied. The equilibrium real wage rate is $\$ 35$ an hour, and equilibrium employment is 200 billion hours per year.

At a wage rate above $\$ 35$ an hour, there is a surplus of labor and the real wage rate falls to eliminate the surplus. At a wage rate below $\$ 35$ an hour, there is a shortage of labor and the real wage rate rises to eliminate the shortage.
If the real wage rate is $\$ 35$ an hour, the quantity of labor demanded equals the quantity supplied and there is neither a shortage nor a surplus of labor. In this situation, there is no pressure in either direction on the real wage rate. So the real wage rateremains constant
and the market is in equilibrium. At this equilibrium real wage rate and level of employment, the economy is atfull employment.

Potential GDP You've seen that the production function tells us the quantity of real GDP that a given amount of labor can produce-see Fig. 23.6. The quantity of real GDP produced increases as the quantity of labor increases. At the equilibrium quantity of labor, the economy is at full employment, and the quantity of real GDP at full employment is potential GDP. So the full-employment quantity of labor produces potential GDP. Figure 23.8 illustrates the determination of potential GDP. Part (a) shows labor market equilibrium. At the equilibrium real wage rate, equilibrium employment is 200 billion hours. Part (b) shows the production function. With 200 billion hours of labor, the economy can produce a real GDP of $\$ 13$ trillion.
This amount is potential GDP.

## What Makes Potential GDP Grow?

We can divide all the forces that make potential GDP grow into two categories:

- Growth of the supply of labor
- Growth of labor productivity

Growth of the Supply of Labor When the supply of labor grows, the supply of labor curve shifts right-ward. The quantity of labor at a given real wage rateincreases. The quantity of labor is the number of workers employed multiplied by average hours perworker. The number employed equals the employment-to- population ratio multiplied by the workingage population, divided by 100 (see Chapter 22, p. 556). Sothe quantity of labor changes as a result of changes in

1. Average hours per worker
2. The employment-to-population ratio
3. The working-age population

Average hours per worker have decreased as the workweek has become shorter, and the employment- to-population ratio has increased as more women have entered the labor force. The combined effect of these two factors has kept the average hours per working-age person (approximately) constant.
Growth in the supply of labor has come from growth in the working-age population. In the long run, the working-age population grows at the same rate as the total population.

TheEffects of Population Growth Population growth brings growth in the supply of labor, but it does not change the demand for labor or the production function. The economy can produce more output by using more labor, but there is no change in the quantity of real GDP that a given quantity of labor can produce.
With an increase in the supply of labor and no change in the demand for labor, the real wage rate falls and the equilibrium quantity of labor increases. The increased quantity of labor produces more out- put and potential GDP increases.

Illustrating the Effects of population Growth Figure 23.9 illustrates the effects of an increase in the population. In Fig. 23.9(a), the demand for labor curve is $L D$ and initially the supply of labor curve is $L S_{0}$. The equilibrium real wage rate is $\$ 35$ an hour and the quantity of labor is 200 billion hours a year. In Fig. 23.9(b), the production function (PF) shows that with 200 billion hours of labor employed, potential GDP is $\$ 13$ trillion at point $A$.
An increase in the population increases the supply of labor and the supply of labor curve shifts right- ward to $L S_{1}$ At a real wage rate of $\$ 35$ an hour, there is now a surplus of labor. So the real wage rate falls.
In this example, the real wage rate will fall until it reaches $\$ 25$ an hour. At $\$ 25$ an hour, the quantity of labor demanded equals the quantity of labor sup-plied. The equilibrium quantity of labor increases to 300 billion a year.
Figure 23.9(b) shows the effect on real GDP. As the equilibrium quantity of labor increases from 200 billion to 300 billion hours, potential GDP increases along the production function from $\$ 13$ trillion to $\$ 16$ trillion at point $B$.

So an increase in the population increases the full- employment quantity of labor, increases potential GDP, and lowers the real wage rate. But the population increase decreases potential GDP per hour of labor. Initially, it was $\$ 65$ ( $\$ 13$ trillion divided by 200 billion). With the population increase, potential GDP per hour of labor is $\$ 53.33$ ( $\$ 16$ trillion divided by 300 billion). Diminishing returns are the source of the decrease in potential GDP per hour of labor.

Growth of Labor Productivity Labor productivity is the quantity of real GDP produced by an hour of labor. It is calculated by dividing real GDP by aggregate labor hours. For example, if real GDP is $\$ 13$ trillion and aggregate hours are 200 billion, labor productivity is $\$ 65$ per hour. When labor productivity grows, real GDP per person grows and brings a rising standard of living. Let's see how an increase in labor productivity changes potential GDP.

Effects of an Increase in Labor Productivity If labor productivity increases, production possibilities expand. The quantity of real GDP that any given quantity oflabor can produce increases. If labor is more productive, firms are willing to pay more for a given number of hours of labor so the demand for labor also increases.
With an increase in the demand for labor and no change in the supply oflabor, the real wage rate rises and the quantity of labor supplied increases. The equilibrium quantity of labor also increases.

So an increase in labor productivity increases potential GDP for two reasons: Labor is more productive and more labor is employed.

Illustrating the Effects of an Increase in Labor Productivity Figure 23.10 illustrates the effects of an increase in labor productivity. In part (a), the production function initially is $P F_{0}$. With 200 billion hours of labor employed, potential GDP is $\$ 13$ trillion at point $A$.

FIGURE 23.8 The Labor Market and Potential GDP


## (b) Potential GDP

The economy is at full employment when the quantity of labor demanded equals the quantity of labor supplied, in part (a). The real wage rate is $\$ 35$ an hour, and employment is 200 billion hours a year. Part (b) shows potential GDP. It is the quantity of real GDP determined by the production function at the fullemployment quantity of labor.

FIGURE 23.9 The Effects of an Increase in Population

(b) Potential GDP

An increase in the population increases the supply of labor. In part (a), the supply of labor curve shifts rightward. The real wage rate falls and aggregate labor hours increase. In part (b), the increase in aggregate labor hours brings an increase in potential GDP. But diminishing returns bring a decrease in potential GDP per hour of labor.

In part (b), the demand for labor curve is $L D_{0}$ and the supply of labor curve is $L S$. The real wage rate is $\$ 35$ an hour, and the equilibrium quantity of labor is 200 billion hours a year. Now labor productivity increases. In Fig. 23.10(a), the increase in labor productivity shifts the production function upward to $P F_{1}$. At each quantity of labor, more real GDP can be produced. For example, at 200 billion hours, the economy can now produce $\$ 18$ trillion of real GDP at point B.
In Fig. 23.10(b), the increase in labor productivity increases the demand for labor and the demand for labor curve shifts rightward to $L D_{1}$. At the initial real wage rate of $\$ 35$ an hour, there is now a shortage of labor. The real wage rate rises. In this example, the real wage rate will rise until it reaches $\$ 45$ an hour.
At $\$ 45$ an hour, the quantity of labor demanded equals the quantity of labor supplied and the equilibrium quantity of labor is 225 billion hours a year.

Figure 23.10(a) shows the effects of the increase in labor productivity on potential GDP. There are two effects. At the initial quantity of labor, real GDP increases to point $B$ on the new production function. But as the equilibrium quantity of labor increases from 200 billion to 225 billion hours, potential GDP increases to $\$ 19$ trillion at point C. Potential GDP per hour of labor also increases. Initially, it was $\$ 65$ ( $\$ 13$ trillion divided by 200 billion). With the increase in labor productivity, potential GDP per hour of labor is \$84.44 (\$19 trillion divided by 225 billion).
The increase in aggregate labor hours that you have just seen is a consequence of an increase in labor productivity. This increase in aggregate labor hours and labor productivity is an example of the interaction effects that economists seek to identify in their search for the ultimate causes of economic growth.
In the case that we've just studied, aggregate labor hours increase but that increase is a consequence, not a cause, of the growth of potential GDP. The source of the increase in potential GDP is an increase in labor productivity.

Labor productivity is the key to increasing output per hour of labor and rising living standards. Butwhat brings an increase in labor productivity? The next section answers this question.

## Why labor Productivity Grows

You've seen that labor productivity growth makes potential GDP grow; and you've seen that labor productivity growth is essential if real GDP per person and the standard of living are to grow. But why does labor productivity grow? What are the preconditions that make labor productivity growth possible and what are the forces that make it grow? Why does labor productivity grow faster at some times and in some places than others?

FIGURE 23.10 The Effects of an Increase in Labor Productivity

(b) The labor market

An increase in labor productivity shifts the production func: tion upward from $P F_{0}$ to $P F_{1}$ in part (a) and shifts the demand for labor curve rightward from $L D_{0}$ to $L D_{1}$ in part (b). The real wage rate rises to $\$ 45$ an hour, and aggregate labor hours increase from 200 billion to 225 billion. Potential GDP increases from $\$ 13$ trillion to $\$ 19$ trillion.

## Preconditions for Labor Productivity Growth

The fundamental precondition for labor productivity growth is the incentive system created by firms, markets, property rights, and money. These four social institutions are the same as those described in Chapter 2 (see pp. 82-83) that enable people to gain by specializing and trading.
It was the presence of secure property rights in Britain in the middle 1700s that got the Industrial Revolution going (see Economics inAction on p. 591). And it is their absence in some parts of Africa today that is keeping labor productivity stagnant. With the preconditions for labor productivity growth in place, three things influence its pace:

- Physical capital growth
- Human capital growth
- Technological advances


## Physical Capital Growth

As the amount of capital per worker increases, labor productivity also increases. Production processes that use hand tools can create beautiful objects, but production methods that use large amounts of capital per worker are much more productive. The accumulation of capital on farms, in textile factories, in iron foundries and steel mills, in coal mines, on building sites, in chemical plants, in auto plants, in banks and insurance companies, and in shopping malls has added incredibly to the labor productivity of our economy. The next time you see a movie that is set in the Old West or colonial times, look carefully at the small amount of capital around. Try to imagine howproductive you would be in such circumstances com- pared with your productivity today.

## ECONOMICS IN ACTION

## Women Are the Better Borrowers

Economic growth is driven by the decisions of billions of individuals to save and invest, and to borrow andlend. In developing countries, most people are too poor to save and too big a risk to be able to borrow from abank. But they can get a microloan to start a business, employ a few people, and earn an income. And many of the most successful microloan borrowers are women.
Microloans originated in Bangladesh and have spread throughout the developing world. Kiva.org and M1croPlace.com (owned by eBay) are Web sites that enable people to lend money that is used to make microloans in developing economies.
Microloans are helping many women to feed and clothe their families and to grow their businesses. Butnot all microloan-financed businesses succeed. And the evidence from controlled experiments conducted by Esther Duflo and her colleagues in the Abdul LatifJameel Poverty Action Lab is that gains in consumption are temporary. A few years after getting a microloan, borrowers are no better off on average than they were before taking a loan. Making poor people less poor requires more than access to microloans.

## Human Capital Growth

Human capital-the accumulated skill and knowledge of human beings-is the fundamental source oflabor productivity growth. Human capital grows when a new discovery is made and it grows as more and more people learn how to use past discoveries.
The development of one of the most basic human skills - writing - was the source of some of the earliest major gains in productivity. The ability to keep written records made it possible to reap ever larger gains from specialization and trade. Imagine how hard it would be to do any kind of business if all the accounts, invoices, and agreements existed only in people's memories.
Later, the development of mathematics laid the foundation for the eventual extension of knowledge about physical forces and chemical and biological processes. This base of scientific knowledge was the foundation for the technological advances of the Industrial Revolution and of today's information revolution.
But a lot of human capital that is extremely productive is much more humble. It takes the form of millions of individuals learning and becoming remarkably more productive by repetitively doing simple production tasks. One much-studied example of this type of human capital growth occurred in World war II. With no change in physical capital, thousands of workers and managers in U.S. shipyards learned from experience and accumulated human capital that more than doubled their productivity in less than two years.

## Technological Advances

The accumulation of physical capital and human capital have made a large contribution to labor productivity growth. But technological change - the discovery and the application of new technologies - has made an even greater contribution.
Labor is many times more productive today than it was a hundred years ago but not because we have: more steam engines and more horse-drawn carriages per person. Rather, it is because we have transportation equipment that uses technologies that were unknown a hundred years ago and that are more productive than the old technologies were. Technological advance arises from formal research and development programs and from informal trial, and error, and it involves discovering new ways ofgetting more out of our resources.
Toreap the benefits of technological change, capital must increase. Some of the most powerful and far-reaching fundamental technologies areembodied in human capital - for example, language, writing, and mathematics. But most technologies are embodied physical capital. For example, to reap the benefits $d$.. the internal combustion engine, millions of horse-drawn carriages had to be replaced with automobiles and to reap the benefits of digital music, millions of Discmans had to be replaced by iPods

## ECONOMICS IN ACTION

## Intellectual Property Rights Propel Growth

In 1760, when the states that 16 years later would become the United States of America were developingagricultural economies, England was on the cusp of an economic revolution, the Industrial Revolution.

For 70 dazzling years, technological advances in the use of steam power, the manufacture of cotton, wool, iron, and steel, and in transportation, accompanied by massive capital investment associated with these technologies, transformed the economy of England. Incomes rose and brought an explosion in an increasingly urbanized population. By 1825, advances in steam technology had reached a level of sophistication that enabled Robert Stevensonto build the world's first steam-powered rail engine (the Rocket, pictured on page 591 in the Science Museum, London) and the birth of the world's first railroad.
Why did the Industrial Revolution happen? Why did it start in 1760? And why in England? Economic historians say that intellectual property rights - England's patent system provides the answer.
England's patent system began with the Statute of Monopolies of 1624, which gave inventors a monopoly to use their idea for a term of 14 years. For about 100 years, the system was used to reward friends of the royal court rather than true inventors. But from around 1720 onward, the system started to work well. To be granted a 14-year monopoly, an inventor only had to pay the required $£ 100$ fee (about $\$ 22,000$ in today's money) and register his or her invention. The inventor was not required to describe the invention in too much detail, so registering and getting a patent didn't mean sharing the invention with competitors. This patent system, which is essentially the same as today's, aligned the self-interest of entrepreneurialinventors with the social interest and unleashed a flood of inventions, the most transformative of which wassteam power and, by 1825 , the steam locomotive.

FIGURE 23.11 The Sources of Economic Growth


> Labor supply growth and labor productivity growth combine to determine real GDP growth. Real GDP per person growth depends on real GDP growth and population growth.

Figure 23.11 summarizes the sources of labor productivity growth and more broadly, of real GDP growth. The figure also emphasizes that for real GDP per person to grow, real GDP must grow faster than the population.
Economics in the News on the next page provides an example of today's labor productivity growth arising from the spread of robot technologies.

## Is Economic Growth Sustainable? Theories, Evidence, and Policies

You've seen how population growth and labor productivity growth make potential GDP grow. You've also seen that the growth of physical capital and human capital and technological advances make labor productivity grow. But what causes economic growth? Why do growth rates vary? How do population growth, capital accumulation, and technological change interact to determine the economic growth rate? What can we say about the future of economic growth? Is growth sustainable? Will the rich economies and
the economies of the developing world keep growing, or will growth end to be followed by stagnation or even a falling standard ofliving?
Economists have wrestled with these questions for the past 250 years and made progress in answering them. We're now going to look at the evolutionof ideas about the sustainability of economic growth and the policies that might achieve faster growth. We start by studying the three main theories about the process of economic growth:

- Classical growth theory
- Neoclassical growth theory
- New growth theory

Classical Growth Theory is the view that the growth of real GDP per person is temporary and that when it rises above the subsistence level, a population explosion eventually brings it back to the subsistence level. Adam Smith, Thomas Robert Malthus, and DavidRicardo the leading economists of the late eighteenth and early nineteenth centuries - proposed this theory, but the view is most closely associated with the name of Malthus and is sometimes called the Malthusian theory. Charles Darwin's ideas about evolution by natural selection were inspired by the insights of Malthus.

Modern-Day Malthusians Many people today are Malthusians. They say that if today's global population of 7.2 billion explodes to 11 billion by 2050 and perhaps 35 billion by 2300, we will run out of resources, real GDP per person will decline, and we will return. to a primitive standard of living. We must, say Malthusians, contain population growth. Modern-day Malthusians also point to global warming and climate change as reasons to believe that, eventually, real GDP per person will decrease.

Neoclassical Growth Theory is the proposition that real GDP per person grows because technological change induces saving and investment that make capital per hour of labor grow. Growth ends if technological change stops because of diminishing marginal returns to both labor and capital. Robert Solow of MIT suggested the most popular version of this growth theory in the 1950s.
Neoclassical growth theory's big break with its classical predecessor is its view about population growth.

Neoclassical Theory of population Growth The population explosion of eighteenth century Europe that created the classical theory of population growth eventually ended. The birth rate fell, and while the populationcontinued to increase, its rate of increase moderated.
The key economic influence that slowed the population growth rate is the opportunity cost of a woman's time. As women's wage rates increase and their job opportunities expand, the opportunity cost of having children increases. Faced with a higher opportunity cost, families choose to have fewerchildren and the birth rate falls.
Technological advances that bring higher incomes also bring advances in healthcare that extendlives.
So as incomes increase, both the birth rate and the death rate decrease. These opposing forces offset each other and result in a slowly rising population.

This modern view of population growth and the historical trends that support it contradict the views of the classical economists. They also call into question the modern doomsday view that the planet will be swamped with more people than it can support.

Technological Change and Diminishing Returns In neoclassical growth theory, the pace of technological change influences the economic growth rate but economic growth does not influence the pace of technological change. Neoclassical growth theory assumes that technological change results from chance. When we're lucky, we have rapid technological change, and when bad luck strikes, the pace of technological advance slows.
To understand neoclassical growth theory, imagine the world of the mid-1950s, when Robert Solow is explaining his idea. Income per person is around $\$ 12,000$ a year in today's money. The population is growing at about 1 percent a year. Saving and investment are about 20 percent of GDP, which is enough to keep the quantity of capital per hour of labor constant. Income per person is growing but not very quickly.
Then technology begins to advance at a more rapid pace across a range of activities. The transistor revolutionizes an emerging electronics industry.
New plastics revolutionize the manufacture of household appliances. The interstate highway sys- tem revolutionizes road transportation. Jet airliners start to replace pistonengine airplanes and speed air transportation.
These technological advances bring new profit opportunities. Businesses expand, and new businesses are created to exploit the newly available profitabletechnologies. Investment and saving increase. The economy enjoys new levels of prosperity and growth. But will the prosperity last? And will the growth last?
Neoclassical growth theory says that the prosperity will last but the growth will not last unless technology keeps advancing.
According to neoclassical growth theory, the prosperity will persist because there is no classical population growth to induce the wage rate to fall. So the gains in income per person are permanent.
But growth will eventually stop if technology stops advancing because of diminishing marginal returns to capital. The high profit rates that result from technological change bring increased saving and capital accumulation. But as more capital is accumulated, more and more projects are under- taken that have lower rates of return-diminishingmarginal returns. As the return on capital falls, the incentive to keep investing weakens. With weaker incentives to save and invest, saving decreases and the rate of capital accumulation slows. Eventually, the pace of capital accumulation slows so that it is only keeping up with population growth. Capital per worker remains constant.

A Problem with Neoclassical Growth Theory All economies have access to the same technologies, and capital is free to roam the globe, seeking the highest available real interest rate. Capital will flowacross regions until rates of return are equal, and rates of return will be equal when capital per hour of labor is equal. Real GDP growth rates and income levels per person around the world will converge. Figure 23.5 on p .582 shows that while there is some sign of convergence among the rich countries in part (a), convergence is slow, and part (b) shows that it does not appear to be imminent for all countries. New growth theory
overcomes this shortcoming of neo- classical growth theory. It also explains what determines the pace of technological change.

New Growth Theory holds that real GDP per persongrows because of the choices people make in the pursuit of profit and that growth will persist indefinitely. Paul Romer of Stanford University developed this theory during the 1980s, based on ideas of Joseph Schumpeter during the 1930s and 1940s.
According to the new growth theory, the pace at which new discoveries are made-and at whichtechnology advances-is not determined by chance. It depends on how many people are looking for a new technology and how intensively they are looking. The search for new technologies is driven by incentives.
Profit is the spur to technological change. The forces of competition squeeze profits, so to increase profit, people constantly seek either lower-cost methods of production or new and better products for which people are willing to pay a higher price. .Inventors can maintain a profit for several years by taking out a patent or a copyright, but eventually, a new discovery is copied, and profits disappear. So more research and development is undertaken in the hope of creating a new burst of profitable investmentsand growth.
Two facts about discoveries and technological knowledge play a key role in the new growth theory:. Discoveries are (at least eventually) a public capital <good; and knowledge is capital that is not subject to diminishing marginal returns. Economists call a good a public good when no one can be excluded from using it and when one person. use does not prevent others from using it. National defense is the classic example of a public good. The programming language used to write apps for the iPhone is another.
Because knowledge is a public good, as the benefits of a new discovery spread, free resources become available. Nothing is given up when they are used: They have a zero opportunity cost. When a student in Austin writes a new iPhone app, his use of the programming language doesn't prevent another student in Seattle from using it.
Knowledge is even more special because it is not subject to diminishing returns. But increasing the stock of knowledge makes both labor and machines more productive. Knowledge capital does not bring diminishing returns. Biotech knowledge illustrates this idea well. Biologists have spent a lot of time developing DNA sequencing technology. As more has been discovered, the productivity of this knowledge capital has relentlessly increased. In 1990, it cost about $\$ 50$ to sequence one DNA base pair. That cost had fallen to $\$ 1$ by 2000 and to $1 / 10,000$ th of apenny by 2010.
The implication of this simple and appealing observation is astonishing. Unlike the other two theories, new growth theory has no growth-stopping mechanism. As physical capital accumulates, the return to capital - the real interest rate - falls. But the incentive to innovate and earn a higher profit becomes stronger. So innovation occurs, capital becomes more productive, the demand for capital increases, and the real interest rate rises again.
Labor productivity grows indefinitely as people discover new technologies that yield a higher real interest rate. The growth rate depends only on people's incentives and ability to innovate.

A Perpetual Motion Economy New growth theory sees the economy as a perpetual motion machine, which Fig. 23.12 illustrates.
No matter how rich we become, our wants exceed our ability to satisfy them. We always want a higherstandard of living. In the pursuit of a highest standard ofliving, human societies have developed incentive systems-markets, property rights, and money - that enable people to profit from innovation. Innovationleads to the development of new and better techniques of production and new and better products. To take advantage of new techniques and to produce newproducts, new firms start up and old firms go out of business-firms are born and die. As old firms die and new firms are born, some jobs are destroyed and others are created. The new jobs created are better than the old ones and they pay higher real wage rates. Also, with higher wage rates and more productive techniques, leisure increases. New and better jobs and new and better products lead to more consumption goods and services and, combined with increased leisure, bring a higher standard of living.
But our insatiable wants are still there, so the process continues: Wants and incentives create innovation, new and better products, and a yet higher standard of living.

FIGURE 23.12 A Perpetual Motion Machine


People want a higher standard of living and are spurred by profit incentives to make the innovations that lead to new and better techniques and new and better products.

These new and better techniques and products, in turn, lead to the birth of new firms and the death of some old firms, new and better ¡obs, and more leisure and more consumption goods and services.

The result is a higher standard of living, but people want a still higher standard of living, and the growth process continues.

## New Growth Theory Versus Malthusian Theory

The contrast between the Malthusian theory and new growth theory couldn't be more sharp. Malthusians see the end of prosperity as we know it today and new growth theorists see unending plenty. The contrast becomes clearest by thinking about the differing views about population growth.
To a Malthusian, population growth is part of the problem. To a new growth theorist, population growth is part of the solution. People are the ultimate economic resource. A larger population brings forth morewants, but it also brings a greater amount of scientificdiscovery and technological advance. So rather than being the source of falling real GDP per person,
population growth generates faster labor productivity growth and rising real GDP per person. Resources are limited, but the human imagination and ability to increase productivity are unlimited.

## Sorting Out the Theories

Which theory is correct? None of them tells us the whole story, but each teaches us something of value.
Classical growth theory reminds us that our physical resources are limited and that without advances in technology, we must eventually hit diminishing returns.
Neoclassical growth theory reaches the same conclusion but not because of a population explosion. Instead, it emphasizes diminishing returns to capital and reminds us that we cannot keep growth going just by accumulating physical capital. We must also advance technology and accumulate human capital. We must become more creative in our use of scarce resources. New growth theory emphasizes the capacity of human resources to innovate at a pace that offsets diminishing returns. New growth theory fits the facts of today's world more closely than do either of the other two theories.

## The Empirical Evidence on the Causes of Economic Growth

Economics makes progress by the interplay between theory and empirical evidence. A theory makes pre- dictions about what we will observe if the theory is correct. Empirical evidence, the data generated by history and the natural experiments that it performs, provides the data for testing the theory.
Economists have done an enormous amount of research confronting theories of growth with the empirical evidence. The way in which this research has been conducted has changed over the years.
In 1776, when Adam Smith wrote about "the nature and causes of the Wealth of Nations" in his celebrated book, empirical evidence took the form of carefully selected facts described in words and stories. Today, large databases, sophisticated statistical methods, and fast computers provide numerical measurements of the causes of economic growth. Economists have looked at the growth rate data for more than 100 countries for the period since1960 and explored the correlations between the growth rate and more than 60 possible influences onit. The conclusion of this data crunching is that mostof these possible influences have variable and unpredictable effects, but a few of them have strong andclear effects. Table 23.1 summarizes these morerobust influences. They are arranged in order of difficulty (or in the case of region, impossibility) of changing. Political and economic systems are hard to change, but market distortions, investment, and openness to international trade are features of a nation's economy that can be influenced by policy. Let's now look at growth policies.

## Policies for Achieving Faster Growth

Growth theory supported by empirical evidencetells us that to achieve faster economic growth, we must increase the growth rate of physical capital, the paceof technological advance, or the growth rate of human capital and openness to international trade. The main suggestions for achieving these objectives are

- Stimulate saving
- Stimulate research and development
- Improve the quality of education
- Provide international aid to developing nations
- Encourage international trade

Stimulate Saving Saving finances investment so stimulating saving increases economic growth. The East Asian economies have the highest growth ratesand the highest saving rates. Some African economies have the lowest growth rates and the lowest savingrates.
Tax incentives can increase saving. Individual Retirement Accounts (IRAs) are a tax incentive to save. Economists claim that a tax on consumption rather than income provides the best saving incentive.

TABLE 23.1 The Influences on Economic Growth

| Influence | Good for Economic Growth | Bad for Economic Growth |
| :---: | :---: | :---: |
| Region | - Far from equator | - Sub-Saharan Africa |
| Politics | - Rule of law | - Revolutions |
|  | - Civil liberties | - Military coups |
|  |  | - Wars |
| Economic system | - Capitalist |  |
| Market distortions | . | - Exchange rate distortions |
|  |  | - Price controls and black markels |
| Investment | - Human capital |  |
|  | - Physical capital |  |
| International trade | - Open to trade |  |
| Source of data: Xavier Sala | o Million Regressions," The America | (1ew, Vol. 87, No. 2, (May 1997), pp. 178-183. |

Stimulate Research and Development Everyone can use the fruits of basic research and developmentefforts. For example, all biotechnology firms can useadvances in gene-splicing technology. Because basicinventions can be copied, the inventor's profit is limited and the market allocates too few resources to thisactivity. Governments can direct public funds towardfinancing basic research, but this solution is not fool-proof It requires a mechanism for allocating the public funds to their highest-valued use.

Improve the Quality of Education The free market produces too little education because it brings benefits beyond those valued by the people who receive the education. By funding basic education and by ensuring high standards in basic skills such as language, mathematics, and science, governments can contribute to a nation's growth potential. Education can also be stimulated and improved by using tax incentives to encourage improved private provision.

Provide International Aid to Developing Nations It seems obvious that if rich countries give financial aid to developing countries, investment and growth will increase in the recipient countries. Unfortunately, the obvious does not routinely happen. A large amount of datadriven research on the effects of aid on growth has turned up a zero and even negative effect. Aid often gets diverted and spent on consumption.

Encourage International Trade Trade, not aid, stimulates economic growth. It works by extracting the available gains from specialization and trade. Thefastest-growing nations are those most open to trade. If the rich nations truly want to aid economic development, they will lower their trade barriers against developing nations, especially in farm products. The World Trade Organization's efforts to achieve more open trade are being resisted by the richer nations.
To complete your study of economic growth, take a look at Economics in the News on pp. 560-561, which compares the contrasting growth performance of two African nations.

## ECONOMIC ANALYSIS

- South Africa's economic growth rate has not been spectacular.
- Before 1994, South Africa's economy was hit by sanctions aimed at ending apartheid and real GDP per person decreased.
- Since 1995, real GDP per person has increased but at a rate of 3.4 percent per year.
- South Africa's growth compares unfavorably with that of some other African nations, one of which is its neighbor Botswana, that are growing more rapidly.
- Figure 1 shows real GDP per person in South Africa and Botswana from 1980 to 2012. You can see that real GDP per person in Botswana has grown much more quickly than in South Africa.
- A key reason Botswana's real GDP per person has grown more rapidly than South Africa's is the pace of investment in new capital.
- Figure 2 shows that Botswana invests double the percentage of GDP invested by South Africa.
- The growth of physical capital and human capital and technological change are proceeding at a rapid pace in Botswana and bringing rapid growth in real GDP per person.
- Figure 3 illustrates how the production function is changing in these economies. It is shifting upward at a more rapid pace in Botswana than in South Africa.
- Why is Botswana more successful than South Africa and are the policies proposed in the news article enough to raise South Africa's growth rate to the desired 8 percent per year?
- Economists Daron Acemoglu, Simon Johnson, and James Robinson say that Botswana had the right institutions for growth - well defined and widely respected private property right*.
- The proposals in the news article don't directly address strengthening private property rights, but they do have that effect.
- The labor market reforms described in the article would increase human capital and labor productivity.
- The labor market and capital market reforms together would make capital accumulation and technological change more profitable and further contribute to labor productivity growth.
- The specific target of 8 percent growth is probably too ambitious.


Figure 1 Real GDP in Two African Economies


Figure 2 Investment in Two African Economies


Figure 3 Labor Productivity Growth in Two African Economies

## CHAPTER 24: Finance, Saving, and Investment

After studying this chapter, you will be able to:

- Describe the flows of funds in financial markets
- Explain how saving and investment decisions interact in financial markets
- Explain how governments influence financial markets

Interest rates fell during 2014, and by mid-year the U.S. government could borrow at 2.5 percent per . year. In 201 2, when the economy was still feeling the effects of a financial meltdown in which billions of dollars had been lost, interest rates were even lower. Behind the drama and headlines that interest rates create, financial markets play a crucial, unseen rolefunneling funds from savers and lenders to investors and borrowers. This chapter explains how financial markets work, and Economics in the News at the end of the chapter looks at the forces at work during 2014 that led to lower interest rates.

## Financial Institutions and Financial Markets

The financial institutions and markets that we study in this chapter provide the channels through which saving flows to finance the investment in new capital that makes the economy grow. In studying financial institutions and markets, we distinguish between

- Finance and money
- Capital and financial capital


## Finance and Money

We use the term finance to describe the activity of providing the funds that finance expenditures on capital. The study of finance looks at how households and firms obtain and use financial resources and how they cope with the risks that arise in this activity. Money is what we use to pay for goods and services and factors of production and to make financial transactions. The study of money looks at how households and firms use it, how much of it they hold, how banks create and manage it, and how its quantity influences the economy.
Finance and money are closely interrelated and some of the main financial institutions, such as banks, provide both financial services and monetary services. Nevertheless, by distinguishing between finance and money and studying them separately, we will better understand our financial and monetary markets and institutions.
For the rest of this chapter, we study finance.
Money is the topic of the next chapter.

## Capital and Financial Capital

Economists distinguish between capital and financial capital. Capital consists of physical capital - tools, instruments, machines, buildings, and inventories - and human capital. When economists use the term capital, they mean physical capital.

Financial Capital consists of the funds that firms use to buy physical capital and that households use to buy a home or to invest in human capital.
You're going to see, in this chapter, how investment, saving, borrowing, and lending decisions influence the quantity of capital and make it grow and, as a consequence, make real GDP grow. We begin by describing the links between capital and investment and between wealth and saving.

## Capital and Investment

The quantity of capital changes because of investment and depreciation. Investment increases the quantity of capital and depreciation decreases it (see Chapter 21, p. 532). The total amount spent on new capital is called gross investment. The change in the value of capital is called net investment. Net investment equals gross investment minus depreciation.
Figure 24.1 illustrates these terms. On January 1, 2014, Ace Bottling Inc. had machines worth $\$ 30,000-$ Ace's initial capital. During 2014, the market value of Ace's machines fell by 67 percent - $\$ 20,000$. After this depreciation, Ace's machines were valued at $\$ 10,000$. During 2014, Ace spent $\$ 30,000$ on new machines. This amount is Ace's gross investment. By December 31, 2014, Ace Bottling had capital valued at $\$ 40,000$, so its capital had increased by $\$ 10,000$. This amount is Ace's net investment. Ace's net investment equals its gross investment of $\$ 30,000$ minus depreciation of its initial capital of $\$ 20,000$.

## Wealth and Saving

Wealth is the value of the things that people own. It contrasts with income, which is what people earn during a given time period from supplying the services of the resources they own. Saving is the amount of income that is not paid in taxes or spent on consumption goods and services. Saving increases wealth.
Wealth also increases when the market value of assets rises-called capital gains-and decreases when the market value of assets falls-called capital losses.
For example, if at the end of the school year you have $\$ 250$ in the bank and a coin collection worth $\$ 300$, then your wealth is $\$ 550$. During the summer (suppose that you earn $\$ 5,000$ (net of taxes) andspend $\$ 1,000$ on consumption goods and services, so.; your saving is $\$ 4,000$. Your bank account increases $\$ 4,250$ and your wealth becomes $\$ 4,550$. The 4,000 increase in wealth equals saving. If coins rise. in value and your coin collection is now worth $\$ 500$, you have a capital gain of $\$ 200$, which is also added to your wealth. National wealth and national saving work like this personal example. The wealth of a nation at the end ' of a year equals its wealth at the start of the year plus: its saving during the year, which equals income minus consumption expenditure.

FIGURE 24.1 Capital and Investment


On January 1, 2014, Ace Bottling had capital worth $\$ 30,000$. During the year, the value of Ace's capital fell by $\$ 20,000$-depreciation -and Ace spent $\$ 30,000$ on new capital-gross investment. Ace's net investment was $\$ 10,000(\$ 30,000$ gross investment minus $\$ 20,000$ depreciation) so that at the end of 2014, Ace had capital worth \$40,000.
To make real GDP grow, saving and wealth must be transformed into investment and capital. This transformation takes place in the markets for financial capital and through the activities of financial institutions. We're now going to describe these markets and institutions.

## Financial Capital Markets

Saving is the source of the funds that are used to finance investment, and these funds are supplied and demanded in three types of financial markets:

- Loan markets
- Bond markets
- Stock markets

Loan Markets Businesses often want short-term finance to buy inventories or to extend credit to their customers. Sometimes they get this finance in the form of a loan from a bank. Households often want finance to purchase big ticket items, such as automobiles or household furnishings and appliances. They get this finance as bank loans, often in the form of outstanding credit card balances. Households also get finance to buy new homes. (Expenditure
on new homes is counted as part of investment.) These funds are usually obtained as a loan that is secured by a mortgage--a legal contract that gives ownership of a home to the lender in the event that the borrower fails to meet the agreed loan payments (repayments and interest). All of these types of financing take place in loan markets.

Bond Markets When Wal-Mart expands its business and opens new stores; itgets the finance it needs by selling bonds. Governments - federal, state, and municipal - also raise finance by issuing bonds.
A bond is a promise to make specified payments on specified dates. For example, you can buy a Wal- Mart bond that promises to pay $\$ 5.00$ every year until 2024 and then to make a final payment of $\$ 100$ in 2025.
The buyer of a bond from Wal-Mart makes a loan to the company and is entitled to the paymentspromised by the bond. When a person buys a newly issued bond, he or she may hold the bond until the borrower has repaid the amount borrowed or sell it to someone else. Bonds issued by firms and governments are traded in the bondmarket.
The term of a bond might be long (decades) or short (just a month or two). Firms often issue very short-term bonds as a way of getting paid for their sales before the buyer is able to pay. For example, when GM sells $\$ 100$ million of railroad locomotives to Union Pacific, GM wants to be paid when the items are shipped. But Union Pacific doesn't want to pay until the locomotives are earning an income. In this situation, Union Pacific might promise to pay GM \$101 million three months in the future. A bankwould be willing to buy this promise for (say) $\$ 100$ million. GM gets $\$ 100$ million immediately and thebank gets $\$ 101$ million in three months when UnionPacific honors its promise. The U.S. Treasury issues promises of this type, called Treasury bills.
Another type of bond is a mortgage-backed security, which entitles its holder to the income from a package of mortgages. Mortgage lenders create mortgage-backed securities. They make mortgage loans to homebuyers and then create securities that they sell to obtain more funds to make more mortgage loans. The holder of a mortgage-backed security is entitled to receive payments that derive from the payments received by the mortgage lender from the home-buyer-borrower. Mortgage-backed securities were at the center of the storm in the financial markets in 2007-2008.

Stock Markets When Boeing wants finance to expand its airplane building business, it issues stock.A stock is a certificate of ownership and claim to thefirm's profits. Boeing has issued about 900 million shares of its stock. So if you owned 900 Boeing shares, you would own one millionth of Boeing andbe entitled to receive one millionth of its profits. Unlike a stockholder, a bondholder does not own part of the firm that issued the bond. A stock market is a financial market in which shares of stocks of corporations are traded. The New York Stock Exchange, the London Stock Exchange (in England), the Tokyo Stock Exchange (in Japan), and the Frankfurt Stock Exchange (in Germany) are all examples of stock markets.

## Financial Institutions

Financial markets are highly competitive because of the role played by financial institutions in those markets. A financial institution is a firm that operates on both sides of the markets for financial capital. A financial institution is a borrower in one market anda lender in another. Financial institutions also stand ready to trade so that households with funds to lend and firms or households seeking funds can always findsomeone on the other side of the market with whom to trade. The key financial institutions are

- Commercial banks
- Government-sponsored mortgage lenders
- Pension funds
- Insurance companies
- The Federal Reserve

Commercial Banks Commercial banks are financial institutions that accept deposits, provide payment services, and make loans to firms and households. The bank that you use for your own banking services and that issues your credit card is a commercial bank.
These institutions play a central role in the monetary • system. We study commercial banks, along with the Federal Reserve that regulates them, in detail in Chapter 25.

Government-Sponsored Mortgage Lenders Two large financial institutions, the Federal National Mortgage Association, or Fannie Mae, and the Federal Home Loan Mortgage Corporation, or Freddie Mac, are enterprises that buy mortgages from banks, package them into mortgage-backed securities, and sell them. In September 2008, Fannie and Freddie owned or guaranteed $\$ 6$ trillion worth of mortgages (half of the U.S. $\$ 12$ trillion of mortgagesand were taken over by the federal government.

Pension Funds Pension funds are financial institutions that use the pension contributions of firms and workers to buy bonds and stocks. The mortgage-backed securities of Fannie Mae and Freddie Mac are among the assets of pension funds. Some pension funds are very large, and they play an active role in the firms whose stock they hold.

Insurance Companies Insurance companies enable households and firms to cope with risks such as accident, theft, fire, ill-health, and a host of other misfortunes. They receive premiums from their customers and pay claims. Insurance companies use the funds they have received but not paid out as claims to buy bonds and stocks on which they earn interest income.
In normal times, insurance companies have a steady flow of funds coming in from premiums and interest on the financial assets they hold and a steady, but smaller, flow of funds paying claims. Their profit is the gap between the two flows. But in unusual times, when large and widespread losses are being incurred, insurance companies can run into difficulty in meeting their obligations. Such a situation arose in 2008 for one of the biggest insurers, AIG, and the firm was taken into public ownership.

## ECONOMICS IN ACTION

## The Financial Crisis and the Fix

Bear Stearns: absorbed by JPMorgan Chase with help from the Federal Reserve. Lehman Brothers: . (,gone. Fannie Mae and Freddie Mac: taken into government oversight. Merrill Lynch: absorbed by Bank of America. AIG: given an $\$ 85$ billion lifeline by the Federal Reserve and sold off in parcels to financial,: institutions around the world. Wachovia: taken oven by Wells Fargo. Washington Mutual: taken over by JPMorgan Chase. Morgan Stanley: 20 percent bought by Mitsubishi, a large Japanese bank. These are some of the events in the financial crisis of 2008. What was going on and how can a replay be avoided?

## What Was Going On?

Between 2002 and 2006, mortgage borrowing to buy a home exploded and home prices rocketed. You can see the rise in mortgage borrowing in Fig. 1. Mortgage increased from 65 percent of income in 2000 to more than 100 percent in 2006. And you can see the rocketing home prices in Fig. 2. Between 2000 and 2006, home prices doubled. Then, in 2007 they crashed. Banks and other financial institutions that had made mortgage loans to home buyers sold the loans to Fannie Mae, Freddie Mac, and other large banks that bundledthese loans into mortgage-backed securities and sold them to eager buyers around the world.
When home prices began to fall in 2007, many home owners found themselves with a mortgage that was bigger than the value of their home. The mortgage defaultrate jumped and the prices of mortgage-backed securities, and more widely of other assets, fell sharply. Financial institutions took big losses. Some losses were too big to bear and some big-name institutions failed.

Avoiding a Replay
In the hope of avoiding a replay, Congress enacted the Dodd-Frank WallStreet Reform and Consumer Protection Act of 2010. The main points of the Act are

- A Consumer Financial Protection Bureau to enforce consumer-oriented regulation, ensure that the fine print on financial services contracts is clear and accurate, and maintain a toll-free hotline for consumers to report alleged deception.
- A Financial Stability Oversight Council to anticipate financial market weakness.
- Authority for the Federal Deposit Insurance Corporation to seize, liquidate, and reconstruct troubled financial firms.
- Tight restrictions to stop banks gambling for their own profit and limit their risky investments.
- Mortgage reforms that require lenders to review the income and credit histories of applicants and ensure that they can afford payments.
- A requirement that the firms that create mortgage- backed securities keep at least 5 percent of them.
The 2010 Act does nothing to solve the problem that arises from government oversight of Fannie Mae and Freddie Mac. Many people believe that the measures are too timid and leave the financial system fragile.


Figure 1 Household Debt


Figure 2 Home Prices

The Federal Reserve The Federal Reserve System (usually called the Fed) is the central bank of the United States, a public authority whose main role is the regulation of banks and money. In recent years, in response to a financial crisis in 2007 and 2008, theFed has played a big role in the markets for bonds and mortgage-backed securities buying these items in large quantities. We study the Fed in detail in Chapter 25 but you need to keep in mind its presence in financial markets.

## Insolvency and Illiquidity

A financial institution's net worth is the market value of what it has lent minus the market value of what it has borrowed. If net worth is positive, the institution is solvent. But if net worth is negative, the institution is insolvent and must go out of business. The owners of an insolvent financial institution - usually its stockholders - bear the loss.
A financial institution both borrows and lends, so it is exposed to the risk that its net worth might become negative. To limit that risk, financial institutions are regulated and a minimum amount of their lending must be backed by their net worth.
Sometimes, a financial institution is solvent but illiquid. A firm is illiquid if it has made longterm loans with borrowed funds and is faced with a sudden demand to repay more of what it has borrowed than its available cash. In normal times, a financial institution that is illiquid can borrow from another institution. But if all the financial institutions are short of cash, the market for loans among financial institutions dries up.
Both insolvency and illiquidity were at the core of the financial meltdown of 2007-2008.

## Interest Rates and Asset Prices

Stocks, bonds, short-term securities, and loans are collectively called financial assets. The interest rate on a financial asset is the interest received expressed as a ; percentage of the price of the asset.
Because the interest rate is a percentage of the price of an asset, if the asset price rises, other things remaining the same, the interest rate falls. Conversely, if the asset price falls, other things remaining the same, the interest rate rises.
To see this inverse relationship between an asset price and the interest rate, let's look at an example. We'll consider a bond that promises to pay its holder \$5 a year forever. What is the rate of return - the interest rate - on this bond? The answer depends on the price of the bond. If you could buy this bond $\$ 50$, the interest rate would be 10 percent per year: Interest rate $=(\$ 5+\$ 50) \mathrm{X} \quad 100=10$ percent.
But if the price of this bond increased to $\$ 200$, its:; rate of return or interest rate would be only 2.5 percent per year. That is, Interest rate $=(\$ 5+\$ 200) \times 100=2.5$ percent.
This relationship means that the price of an asset and the interest rate on that asset are determined simultaneously - one implies the other.
This relationship also means that if the interest rate on the asset rises, the price of the asset falls, debt\$ become harder to pay, and the net worth of the financial institution falls. Insolvency can arise from a'previously unexpected large rise in the interest rate. '. In the next part of this chapter, we learn how interest rates and asset prices are determined in the financial markets.

## The Loanable Funds Market

In macroeconomics, we group all the financial markets that we described in the previous section into a single loanable funds market. The loanable funds market is the aggregate of all the individual financial markets.
The circular flow model of Chapter 21 (see p.531) can be extended to include flows in the loanable funds market that finance investment.

## Funds That Finance Investment

Figure 24.2 shows the flows of funds that finance investment. They come from three sources:

1. Household saving
2. Government budget surplus
3. Borrowing from the rest of the world

Households' income, $Y$, is spent on consumption goods and services, $C$, saved, $S$, or paid in net taxes, T. Net Taxes are the taxes paid to governments minus the cash transfers received from governments (such as Social Security and unemployment benefits). So income is equal to the sum of consumption expenditure, saving, and nettaxes:

$$
Y=C+S+T .
$$

You saw in Chapter 21 (p.532) that $Y$ also equals the sum of the items of aggregate expenditure: consumption expenditure, $C$, investment, $I$, government expenditure, $G$, and exports, $X$ minus imports, $M$. That is:

$$
Y=C+I+G+X-M .
$$

By using these two equations, you can see that

$$
I+G+X=M+S+T .
$$

Subtract $G$ and $X$ from both sides of the last equation to obtain

$$
I=S+(T-G)+(M-X) .
$$

This equation tells us that investment, $/$ is financed by household saving, $S$, the government budget surplus, ( $T-G$ ), and borrowing from the rest of the world, ( $M-X$ ). The sum of private saving, $S$, and government saving, ( $T-G$ ), is called national saving. National saving and foreign borrowing finance investment.

In 2014, U.S. investment was $\$ 2.8$ trillion. Governments (federal, state, and local combined) had a deficit of $\$ 0.8$ trillion. This total of $\$ 3.6$ trillion was financed by private saving of $\$ 3.0$ trillion and borrowing from the rest of the world (negative net exports) of $\$ 0.6$ trillion.
In the rest of this chapter, we focus on the influences on national saving and the effects of a government budget deficit (or surplus) in the loanable funds market. We broaden our view to examine the influences on and the effects of borrowing from the rest of the world in Chapter 26. You can think of this chapter as an account of the U.S. loanable funds market when exports equal imports, $(X=M)$ or as an account of the global loanable funds market.
You're going to see how investment and saving and the flows of loanable funds-all measured in constant 2009 dollars-are determined. The price inthe loanable funds market that achieves equilibrium is an interest rate, which we also measure in real terms as the real interest rate. In the loanable funds market, there is just one interest rate, which is an average of the interest rates on all the different types of financial securities that we described earlier. Let's see what we mean by the real interest rate.

FIGURE 24.2 Financial Flows and the Circular Flow of Expenditure and Income


Households use their income for consumption expenditure (CJ, saving (S), and net taxes(T). Firms borrow to finance their investment expenditure. Governments borrow to finance a budget deficit orrepay debt if they have a budget surplus. The rest of the world borrows to finance its deficit orlends its surplus.

## The Real Interest Rate

The nominal interest rate is the number of dollars that a borrower pays and a lender receives in interest in ayear expressed as a percentage of the number of dollars borrowed and lent. For example, if the annual interest paid on a $\$ 500$ loan is $\$ 25$, the nominal interest rate is 5 percent per year: $\$ 25 \div \$ 500 \times 100$ or 5 percent. The real interest rate is the nominal interest rate adjusted to remove the effects of inflation on the buying power of money. The real interest rate is approximately equal to the nominal interest rate minus the inflation rate. You can see why if you suppose that you have put $\$ 500$ in a savings account that earns 5 percent a year. At the end of a year, you have \$525 in your savings account. Suppose that the inflation rate is 2 percent per year-during the year, all prices increased by 2 percent. Now, at the end of the year, it costs $\$ 510$ to ' buy what $\$ 500$ would have bought one year ago. Your money in the bank has really only increased by $\$ 15$, from $\$ 510$ to $\$ 525$. That $\$ 15$ is equivalent to a/ real interest rate of 3 percent a year on your original $\$ 500$. So the real interest rate is the 5 percent nominal interest rate minus the 2 percent inflation rate The real interest rate is the opportunity cost of loanable funds. The real interest paid on borrowedfunds is the opportunity cost of borrowing. And the real interest rateforgone when funds are used either to buy consumption goods and services or to invest in new capital goods is the opportunity cost of not saving or not lending those funds.
We're now going to see how the loanable funds market determines the real interest rate, the quantity fof funds loaned, saving, and investment. In the rest of this section, we will ignore the government and the rest of the world and focus on households and firms in the loanable funds market.

We will study

1. The demand for loanable funds
2. The supply of loanable funds
3. Equilibrium in the loanable funds market

## ECONOMICS IN ACTION

## Nominal and Real Interest Rates

Nominal and real interest rates were extremely high during the 1970s and 1980s. They have trended downward for the past 30 years. Where will they go next? See Economics in the News on pp. 620-621.


## ECONOMICS IN ACTION

The Total Quantities Supplied and Demanded
Around $\$ 80$ trillion of loanable funds have been supplied and demanded. The figure shows who supplies the funds and who demands them. Almost one third of the funds are supplied to banks and similar financial institutions.

## The Demand for Loanable Funds

The quantity ofloanable funds demanded is the total quantity of funds demanded to finance investment, the government budget deficit, and international in- vestment or lending during a given period. Our focus here is on investment. We'll bring the government budget deficit into the picture later in this chapter.
What determines investment and the demand for loanable funds to finance it? Many details influence this decision, but we can summarize them in two factors:

1. The real interest rate
2. Expected profit

Firms invest in capital only if they expect to earn a profit and fewer projects are profitable at a high real interest rate than at a low real interest rate, so

Other things remaining the same, the higherthe real interest rate, the smaller is the quantity of loanable funds demanded; and the lower the real interest rate, the greater the quantity of loanable funds demanded.


The Demand for Loanable Funds Curve is the relationship between the quantity of loanable funds demanded and the real interest rate, when all other influences on borrowing plans remain the same. The demand curve DLF in Fig. 24.3 is a demand for loanable funds curve. To understand the demand for loanable funds, think about Amazon.com's decision to borrow $\$ 100$ million to build some new warehouses. If Amazon expects to get a return of $\$ 5$ million a year from thisinvestment before paying interest costs and the interest rate is less than 5 percent a year, Amazon wouldmake a profit, so it builds the warehouses. But if the interest rate is more than 5 percent a year, Amazon would incur a loss, so it doesn't build the warehouses. The quantity of loanable funds demanded is greater the lower is the real interest rate.

Changes in the Demand for Loanable Funds When the expected profit changes, the demand for loanablefunds changes. Other things remaining the same, thegreater the expected profit from new capital, the greater is the amount of investment and the greater the demand for loanable funds.
Expected profit rises during a business cycle expansion and falls during a recession; rises when technological change creates profitable new products; rises as a growing population brings increased demand for goods and services; and fluctuates with contagious swings of optimism and pessimism, called "animal spirits" by Keynes and "irrational exuberance" by Alan Greenspan. When expected profit changes, the demand for loanable funds curve shifts.
figure 24.3 The Demand for Loanable Funds


A change in the real interest rate changes the quantity of loanable funds demanded and brings a movement along the demand for loanable funds curve.

## The supply of Loanable Funds

The quantity of loanable funds supplied is the total funds available from private saving, the government budget surplus, and supplied by the Fed, during a given period. Our focus here is on saving. We'll bring the government budget and the Fed into the picture in the next part of the chapter.
How do you decide how much of your income to save and supply in the loanable funds market?

Your decision is influenced by many factors, but chief among them are

1. The real interest rate
2. Disposable income
3. Expected future income
4. Wealth
5. Default risk

We begin by focusing on the real interest rate.
Other things remaining the same, the higher the real interest rate, the greater is the quantity of loanable funds supplied; and the lower the real interest rate, the smaller is the quantity of loanable funds supplied.

The Supply of Loanable Funds Curve is the relationship between the quantity of loanable funds supplied and the real interest rate when all other influences on lending plans remain the same. The curve SLF in Fig. 24.4 is a supply of loanable funds curve.
Think about a student's decision to save some of what she earns from her summer job. With a real interest rate of 2 percent a year, she decides that it is not worth saving muchbetter to spend the incomeand take a student loan if funds run out during the semester. But if the real interest rate jumped to 10 percent a year, the payoff from saving would be high enough to encourage her to cut back on spending and increase the amount she saves.

FIGURE 24.4 The Supply of
Loanable Funds


A change in the real interest rate changes the quantity of loanable funds supplied and brings a movement along the supply of loanable funds curve.

Changes in the Supply of Loanable Funds A change in disposable income, expected future income, wealth, or default risk changes the supply of loanable funds.

Disposable Income A household's disposable income is the income earned minus net taxes. When disposableincome increases, other things remaining the same, consumption expenditure increases but by less than the increase in income. Some of the increase in income is saved. So the greater a household's disposable income, other things remaining the same, the greater is its saving.

Expected Future Income The higher a household's expected future income, other things remaining thesame, the smaller is its saving today.

Wealth The higher a household's wealth, other things remaining the same, the smaller is its saving. If a per-son's wealth increases because of a capital gain, the person sees less need to save. For example, from 2002 through 2006, when house prices were rising rapidly, wealth increased despite the fact that personal saving dropped close to zero.

Default Risk The risk that a loan will not be repaid is called default risk. The greater that risk, the higher is the interest rate needed to induce a person to lend and the smaller is the supply of loanable funds.

Shifts of the Supply of Loanable Funds Curve When any of the four influences on the supply of loanable funds changes, the supply of loanable funds changes and the supply curve shifts. Anincrease in disposable income, a decrease in expected future income, a decrease in wealth, or a fall in default risk increases saving and increases the supply of loanable funds.

## Equilibrium in the Loanable Funds Market

You've seen that other things remaining the same, the higher the real interest rate, the greater is the quantity of loanable funds supplied and the smaller is the quantity of loanable funds demanded. Thereis one real interest rate at which the quantities of loanable funds demanded and supplied are equal, and that interest rate is the equilibrium real interest rate.
Figure 24.5 shows how the demand for and supply of loanable funds determine the real interest rate. The DLF curve is the demand curve and the SLF curve is the supply curve. If the real interest rate exceeds 6 percent a year, the quantity of loanable funds sup-plied exceeds the quantity demanded-a surplus of funds. Borrowers find it easy to get funds, but lenders are unable to lend all the funds they have available.
The real interest rate falls and continues to fall until the quantity of funds supplied equals the quantity of funds demanded.
If the real interest rate is less than 6 percent a year, the quantity of loanable funds supplied is less than the quantity demanded-a shortage of funds. Borrowers can't get the funds they want, but lenders are able to lend all the funds they have. So the real interest rate rises and continues to rise until the quantity of funds supplied equals the quantity demanded. Regardless of whether there is a surplus or ashort- age of loanable funds, the real interest rate changes and is pulled toward an equilibrium level. In Fig. 24.5, the equilibrium real interest rate is 6 percent a year. At this interest rate, there is neither a surplus nor a shortage of loanable funds. Borrowers can get the funds they want, and lenders can lend all the funds they have available. The investment plans of borrowers and the saving plans of lenders are consistent with each other.

FIGURE 24.5 Equilibrium in the Loanable Funds Market


A surplus of funds lowers the real interest rate and a shortage of funds raises it. At an interest rate of 6 percent a year, the quantity of funds demanded equals the quantity supplied and the market is in equilibrium.

## Changes in Demand and Supply

Financial markets are highly volatile in the short run but remarkably stable in the long run. Volatility in themarket comes from fluctuations in either the demandfor loanable funds or the supply of loanable funds.
These fluctuations bring fluctuations in the real interest rate and in the equilibrium quantity of funds lent and borrowed. They also bring fluctuations in asset prices.
Here we'll illustrate the effects of increases in demand and supply in the loanable funds market.

## ECONOMICS IN ACTION

## Loanable Funds Fuel Home Price Bubble

The financial crisis that gripped the U.S. and global economies in 2007 and cascaded through the financial markets in 2008 had its origins much earlier in events taking place in the loanable funds market.

Between 2001 and 2005, a massive injection of loanable funds occurred. Some funds came from the rest of the world, but that source of supply has been stable. The Federal Reserve provided funds to keep interest rates low and that was a major source of the increase in the supply of funds. (The next chapter explains how the Fed does this.)

Figure 1 illustrates the loanable funds market starting in 2001. In that year, the demand for loanable funds was $D L F_{01}$ and the supply of loanable funds was $S L F_{01}$. The equilibrium real interest rate was 4 percent a year and the equilibrium quantity of loanable funds was $\$ 29$ trillion (in 2009 dollars).

During the ensuing four years, a massive increase in the supply of loanable funds shifted the supply curve rightward to $S L F_{05}$. A smaller increase in demand shifted the demand for loanable funds curve to $D L F_{05}$. The real interest rate fell to 1 percent a year and the quantity of loanable funds increased to $\$ 36$ trillion-a 24 percent increase in just four years.

With this large increase in available funds, much of it in the form of mortgage loans to home buyers, the demand for homes increased by more than the increase in the supply of homes. Home prices rose and the expectation of further increases fueled the demand for loanable funds.

By 2006, the expectation of continued rapidly rising home prices brought a very large increase in the demand for loanable funds. At the same time, the Federal Reserve began to tighten credit. (Again, you'll learn how this is done in the next chapter). The result of the Fed's tighter credit policy was a slowdown in the pace of increase in the supply of loanable funds.

Figure 2 illustrates these events. In 2006, the demand for loanable funds increased from $D L F_{05}$ to $D L F_{06}$ and the supply of loanable funds increased by a smaller amount from $S L F_{05}$ to $S L F_{06}$. The real interest rate increased to 3 percent a year.

The rise in the real interest rate (and a much higher rise in the nominal interest rate) put many homeowners in financial difficulty. Mortgage payments increased and some borrowers stopped repaying their loans.


Figure 1 The Foundation of the Crisis: 2001-2005


Figure 2 The Start of the Crisis: 2005-2006

By August 2007, the damage from mortgage default and foreclosure was so large that the credit market began to dry up. A large decrease in both demand and supply kept interest rates roughly constant but decreased the quantity of new business.

The total quantity of loanable funds didn't decrease, but the rate of increase slowed to a snail's pace and financial institutions most exposed to the bad mortgage debts and the securities that they backed (described on pp. 607-608) began to fail.

These events illustrate the crucial role played by the loanable funds market in our economy.
figure 24.6 Changes in Demand and Supply

(a) An increase in demand

(b) An increase in supply

In part (a), the demand for loanable funds increases and supply doesn't change. The real interest rate rises (financial asset prices fall) and the quantity of funds increases. In part (b), the supply of loanable funds increases and demand doesn't change. The real interest rate falls (financial asset prices rise) and the quantity of funds increases.

An Increase in Demand If the profits that firms expect to earn increase, they increase their planned investment and increase their demand for loanable funds to finance that investment. With an increase inthe demand for loanable funds, but no change in the supply of loanable funds, there is a shortage of funds. As borrowers compete for funds, the interest rate rises and lenders increase the quantity of funds supplied.
Figure 24.6(a) illustrates these changes. An increase in the demand for loanable funds shifts the demand curve rightward from $D L F_{0}$ to $D L F_{1}$, With no change in the supply of loanable funds, there is a shortage of funds at a real interest rate of 6 percent a year. The real interest rate rises until it is 7 percent a year. Equilibrium is restored and the equilibrium quantity of funds has increased.

An Increase in Supply If one of the influences on saving plans changes and increases saving, the supply of loanable funds increases. With no change in the demand for loanable funds, the market is flush with loanable funds. Borrowers find bargains and lenders find themselves accepting a lower interest rate. At the lower interest rate, borrowers find additional investment projects profitable and increase the quantity of loanable funds that they borrow.
Figure 24.6(b) illustrates these changes. An increase in supply shifts the supply curve rightwardfrom $S L F_{0}$ to $S L F_{1}$. With no change in demand, there is a surplus of funds at a real interest rate of 6 percent a year. The real interest rate falls until it is 5 percent a year. Equilibrium is restored and the equilibrium quantity of funds has increased.

Long-Run Growth of Demand and Supply Over time, both demand and supply in the loanable funds market fluctuate and the real interest rate rises and falls. Both the supply of loanable funds and the demand for loanable funds tend to increase over time. On the average, they increase at a similar pace, so although demand and supply trend upward, the real interest rate has no trend. It fluctuates around a constant average level.

## Government in the Loanable Funds Market

Government enters the loanable funds market when it has a budget surplus or budget deficit. A budget surplus increases the supply of loanable funds and contributes to financing investment; a budget deficit increases the demand for loanable funds and competes with businesses for funds.
Let's study the effects of government on the loan- able funds market.

## A Government in Budget Surplus

A government budget surplus increases the supply of loanable funds. The real interest rate falls, which decreases household saving and decreases thequantity of private funds supplied. The lower real interest rate increases the quantity of loanable funds demanded, and increases investment.
Figure 24.7 shows these effects of a government budget surplus. The private supply of loanable funds curve is PSLF. The supply of loanable funds curve, SLF, shows the sum of private supply and the government budget surplus. Here, the government budget surplus is $\$ 1$ trillion, so at each real interest rate the SLF curve lies $\$ 1$ trillion to the right of the

PSLF curve. That is, the horizontal distance between the PSLF curve and the SLF curve equals the government budget surplus.
With no government surplus, the real interest rate is 6 percent a year, the quantity of loanable funds is $\$ 2$ trillion a year, and investment is $\$ 2$ trillion a year. But with the government surplus of $\$ 1$ trillion a year, the equilibrium real interest rate falls to 5 percent a year and the equilibrium quantity of loanable funds increases to $\$ 2.5$ trillion a year.
The fall in the interest rate decreases private saving to $\$ 1.5$ trillion, but investment increases to $\$ 2.5$ trillion, which is financed by private saving plus the government budget surplus (government saving).

FIGURE 24.7 A Government Budget Surplus


A government budget surplus of $\$ 1$ trillion adds to private saving and the private supply of loanable funds curve, PSLF, to determine the supply of loanable funds curve, SLF. The real interest rate falls to 5 percent a year. Private saying decreases to $\$ 1.5$ trillion, but investment increases to \$2.5 trillion.

## A Government Budget Deficit

A government budget deficit increases the demand for loanable funds. The real interest rate rises, which increases household saving and increases the quantity of private funds supplied. But the higher real interest rate decreases investment and the quantity of loanable funds demanded by firms to finance investment.
Figure 24.8 shows these effects of a government budget deficit. The private demand for loanable funds curve is PDLF. The demand for loanable funds curve, $D L F$, shows the sum of
private demand and the government budget deficit. Here, the government budget deficit is $\$ 1$ trillion, so at each real interest rate the DLF curve lies $\$ 1$ trillion to the right of the PDLF curve. That is, the horizontal distance between the PDLF curve and the DLF curve equals the government budget deficit. With no government deficit, the real interest rate is 6 percent a year, the quantity of loanable funds is $\$ 2$ trillion a year and investment is $\$ 2$ trillion a year. But with the government budget deficit of $\$ 1$ trillion a year, the equilibrium real interest rate rises to 7 per- cent a year and the equilibrium quantity of loanable funds increases to $\$ 2.5$ trillion a year.
The rise in the real interest rate increases private saving to $\$ 2.5$ trillion, but investment decreases to $\$ 1.5$ trillion because $\$ 1$ trillion of private saving must finance the government budget deficit.

FIGURE 24.8 A Government Budget Deficit


A government budget deficit adds to the private demand for loanable funds curve, PDLF, to determine the demand for loanable funds curve, DLF. The real interest rate rises, saving increases, but investment decreases-a crowding-out effect.

FIGURE 24.9 The Ricardo-Barro Effect


A budget deficit increases the demand for loanable funds. Rational taxpayers increase saving, which shifts the supply of loanable funds curve from $S L F_{0}$ to $S L F_{1}$. Crowding out is avoided: Increased saving finances the budget deficit.

The Crowding-Out Effect The tendency for a government budget deficit to raise the real interest rate and decrease investment is called the crowding-out effect. The crowding-out effect does not decrease in- vestment by the full amount of the government budget deficit because a higher real interest rate induces an increase in private saving that partly contributes toward financing the deficit.

The Ricardo-Barro Effect First suggested by the English economist David Ricardo in the eighteenth century and refined by Robert J. Barro of Harvard University, the Ricardo-Barro effect holds that both of the effects we've just shown are wrong and the government budget has no effect on either the real interest rate or investment. Barro says that taxpayers are rational and can see that a budget deficit today means that future taxes must be higher and future disposable incomes smaller. With smaller expected future disposable incomes, saving increases today. The private supply of loanable funds
increases to match the quantity of loanable funds demanded by the government. So the budget deficit has no effect on either the real interest rate or investment. Figure 24.9 shows this outcome.
Most economists regard the Ricardo-Barro view as extreme. But there might be some change in private saving that goes in the direction suggested by the Ricardo- Barro effect that lessens the crowding-out effect.

## ECONOMIC ANALYSIS

- The news article reports that the interest rates on government bonds fell during 2014 from an already low level and bond prices increased.
- Government bonds, called Treasuries in the United States, gilts in the United Kingdom, and bunds in Germany, are the safest securities in the loanable funds market.
- They are also easily traded, so they can be sold at a moment's notice, which makes them highly liquid.
- Because they are safe and liquid, government bonds have a lower interest rate than corporate bonds-bonds issued by corporations.
- Although the level of an interest rate depends on the safety and liquidity of the security, interest rates, on average, move up and down together and are influenced by common forces that change the supply of and demand for loanable funds.
- Figure 1 shows the interest rate on U.S. government 10-year bonds from 2010 to mid-2014 (both the nominal rate and the real rate.)
- The striking feature of this graph is that although the interest rate was low in 2014, it was not as low as it had been in 2012, when the real interest rate was close to zero for two years and briefly negative at the end of 2012.
- The news article says that the falling rate in 2014 risks making the real rate negative again, but that would require a full 1 percentage point fall in the nominal interest rate or a 1 percentage point rise in the inflation rate.
- The news article says that the interest rate on U.S. government bonds fell from 3 percent to 2.5 percent per year. With inflation constant at 1.4 percent per year, these numbers translate to a fall in the real interest rate from 1.6 percent to 1.1 percent per year.
- Figure 2 illustrates why the real interest rate fell. In January 2014, the demand for loanable funds was $D L F_{\text {jan }}$ and the supply of loanable funds was $S L F_{\text {jan }}$. The equilibrium interest rate was 1.6 percent per year.
- During 2014, the factors described in the news article increased the supply of loanable funds to SLF Jun.
- A key influence on the interest rate is missing from the news article: In 2014, the German, U.S., and U.K. government budget deficits shrank, which decreased the demand for loanable funds to $D L F_{\text {Jun }}$.


Figure 1 Interest Rates 2010 to 2014


Figure 2 The Loanable Funds Market in 2014

- With an increase in supply and a decrease in demand, the equilibrium real interest rate fell from 1.6 percent to 1.1 percent per year.
- The news article speculates that bond prices will fall and interest rates will rise as economic expansion increases the demand for loanable funds.


## CHAPTER 25: Money, the Price Level, and Inflation

After studying this chapter, you will be able to:
Define money and describe its functions

- Explain the economic functions of banks
- Describe the structure and functions of the Federal Reserve System (the Fed)
- Explain how the banking system creates money
- Explain what determines the quantity of money and the nominal interest rate
- Explain how the quantity of money influences the price level and the inflation rate

Money, like fire and the wheel, has been around a long time, and it has taken many forms. It was beads made from shells for North American Indians and tobacco for early American colonists. Today, we use dollar bills or swipe a card or, in some places, tap a cell phone. Are all these things money?
In this chapter, we study money, its functions, how it gets created, how the Federal Reserve regulates its quantity, and what happens when its quantity changes.In Economics in the News at the end of the chapter, we look at the extraordinary increase in the quantity of money in recent years and the prospect it will begin to shrink.

## What is Money?

What do wampum, tobacco, and nickels and dimes have in common? They are all examples of money, which is defined as any commodity or token that is generally acceptable as a means of payment. A means of payment is a method of settling a debt. When a payment has been made, there is no remaining obligation between the parties to a transaction. So what wampum, tobacco, and nickels and dimes have in common is that they have served (or still do serve) as the means of payment. Money serves three other functions:

- Medium of exchange
- Unit of account
- Store of value


## Medium of Exchange

A medium of exchange is any object that is generally accepted in exchange for goods and services. Without a medium of exchange, goods and services must be exchanged directly for other goods and services-anexchange called barter. Barter requires a double coincidence of wants, a situation that rarely occurs. For example, if you want a hamburger, you might offer a CD in exchange for it. But you must find someone who is selling hamburgers and wants your CD.
A medium of exchange overcomes the need for a double coincidence of wants. Money acts as a medium of exchange because people with something to sell will always accept money in exchange forit. But money isn't the only medium of exchange. You can buy with a credit card, but a credit card isn't money. It doesn't make a final payment, and the debt it creates must eventually be settled by using money.

## Unit of Account

A unit of account is an agreed measure for stating the prices of goods and services. To get the most out ofyour budget, you have to figure out whether seeingone more movie is worth its opportunity cost. But that cost is not dollars and cents. It is the number of ice-cream cones, sodas, or cups of coffee that you must give up. It's easy to do such calculationswhen all these goods have prices in terms of dollars and cents (see Table 25.1). If the price of a movie is $\$ 8$ and the price of a cappuccino is $\$ 4$, you know right away that seeing one movie costs you 2 cappuccinos.

| TABLE 25.1Good | The Unit of Account Function of Money Simplifies Price Comparisons |  |
| :---: | :---: | :---: |
|  | Price in money units | Price in units of another good |
| Movie | \$8.00 each | 2 cappuccinos |
| Cappuccino | \$4.00 each | 2 icecream con |
| Ice cream | \$2.00 per cone | 2 packs of jelly beans |
| Jelly beans | 1.00 per pack | 2 st |
| Gum | \$0.50 per stick |  |
| Money as a unit of account. The price of a movie is $\$ 8$ and the price of a stick of gum is $50 \$$, so the opportunity cost of a movie is 16 sticks of gum ( $\$ 8.00 \div 50 \$=16$ ). |  |  |
| No unit of account. You go to a movie theater and learn that the cost of seeing a movie is 2 cappuccinos. You go to a grocery store and learn that a pack of jelly beans costs 2 sticks of gum. But how many sticks of gum does seeing a movie cost you? To answer that question, you go to the coffee shop and find that a cappuccino costs 2 ice-cream cones. Now you head for the ice-cream shop, where an ice-cream cone costs 2 packs of jelly beans. Now you get out your pocket calculator: 1 movie costs 2 cappuccinos, or 4 ice-cream cones, or 8 packs of jelly beans, or 16 sticks of gum! |  |  |
|  |  |  |
|  |  |  |

If jelly beans are \$1 a pack, one movie costs 8 packs of jelly beans. You need only one calculation to figure out the opportunity cost of any pair of goods and services. Imagine how troublesome it would be if your local movie theater posted its price as 2 cappuccinos, thecoffee shop posted the price of a cappuccino as 2 ice-cream cones, the icecream shop posted the price ofan ice-cream cone as 2 packs of jelly beans, and the grocery store priced a pack of jelly beans as 2 sticks of gum! Now how much running around and
calculating will you have to do to find out how much that movie is going to cost you in terms of the cappuccinos, ice-cream cones, jelly beans, or gum that you must give up to see it? You get the answer for cappuccinos right away from the sign posted on the movietheater. But for all the other goods, you're going to have to visit many different stores to establish the price of each good in terms of another and then calculate the prices in units that are relevant for your own decision. The hassle of doing all this researchmight be enough to make a person swear off movies! You can see how much simpler it is if all the pricesare expressed in dollars and cents.

## Store of Value

Money is a store of value in the sense that it can be held and exchanged later for goods and services. Ifmoney were not a store of value, it could not serve as a means of payment. Money is not alone in acting as a store of value. A house, a car, and a work of art are other examples.
The more stable the value of a commodity or token, the better it can act as a store of value and the more useful it is as money. No store of value has acompletely stable value. The value of a house, a car, or a work of art fluctuates over time. The values of the commodities and tokens that are used as money also fluctuate over time. Inflation lowers the value of money and the values of other commodities and tokens that are used as money. To make money as useful as possible as a store of value, a low inflation rate is needed.

## Money in the United States Today

- In the United States today, money consists of Currency
- Deposits at banks and other depository institutions

Currency The notes and coins held by individuals and businesses are known as currency. Notes are money because the government declares them so with the words "This note is legal tender for all debts, public and private." You can see these words on every dollar bill. Notes and coins inside banks are not counted as currency because they are not held byindividuals and businesses.

Deposits of individuals and businesses at banks and other depository institutions, such as savings and loan associations, are also counted as money. Deposits are money because the owners of the deposits can use them to make payments.

Official Measures of Money Two official measures of money in the United States today are known as M1 and M2. M1 consists of currency and traveler's checks plus checking deposits owned by individuals and businesses. M1 does not include currency held by banks, and it does not include currency and checking deposits owned by the U.S. government. M2 consists of MI plus time deposits, savings deposits, and money market mutual funds and other deposits.

Are M1 and M2 Really Money? Money is the means of payment. So the test of whether an asset is moneyis whether it serves as a means of payment. Currencypasses the test. But what about deposits? Checking deposits are money because they can be transferredfrom one
person to another by writing a check or using a debit card. Such a transfer of ownership is equivalent to handing over currency. Because M I consists of currency plus checking deposits and each of these is a means of payment, M1 is money. But what about M2? Some of the savings deposits in M 2 are just as much a means of payment as the checking deposits in M1. You can use an ATM to get funds from your savings account to pay for your purchase at the grocery store or the gas station. But some savings deposits are not means of payment. These deposits are known as liquid assets. Liquidity is the property of being easily convertible into a means of payment without loss in value. Because the deposits in M2 that are not means of payment are quickly and easily converted into a means of payment-currency or checking deposits-they are counted as money.

Deposits Are Money but Checks Are Not In defining money, we include, along with currency, deposits at banks and other depository institutions. But we do not count the checks that people write as money. Why are deposits money and checks not? To see why deposits are money but checks are not, think about what happens when Colleen buys some roller-blades for $\$ 100$ from Rocky's Rollers. When Colleen goes to Rocky's shop, she has $\$ 500$ in her deposit account at the Laser Bank. Rocky has $\$ 1,000$ in his deposit account-at the same bank, as it happens. The deposits of these two people total $\$ 1,500$. Colleen writes a check for $\$ 100$. Rocky takes the check to the bank right away and deposits it. Rocky's bank balance rises from $\$ 1,000$ to $\$ 1,100$, and Colleen's balance falls from $\$ 500$ to $\$ 400$. The deposits of Colleen and Rocky still total $\$ 1,500$. Rocky now has $\$ 100$ more than before, and Colleen has \$100 less.
This transaction has transferred money from Colleen to Rocky, but the check itself was never money. There wasn't an extra $\$ 100$ of money while the check was in circulation. The check instructs the bank to transfer money from Colleen to Rocky.
If Colleen and Rocky use different banks, there is an extra step. Rocky's bank credits \$100 to Rocky's account and then takes the check to a check-clearing center. The check is then sent to Colleen's bank, which pays Rocky's bank $\$ 100$ and then debits Colleen's account $\$ 100$. This process can take a fewdays, but the principles are the same as when two people use the same bank.

Credit Cards Are Not Money You've just seen that checks are not money, but what about credit cards? Isn't having a credit card in your wallet and presenting the card to pay for your roller-blades the samething as using money? Why aren't credit cards some- how valued and counted as part of the quantity of money?
When you pay by check, you are frequently asked to prove your identity by showing your driver's license. It would never occur to you to think of your driver's license as money. It's just an ID card. A credit card is also an ID card, but one that lets you take out a loan at the instant you buy something. When you sign a credit card sales slip, you are saying, "I agree to pay for these goods when the credit card company bills me." Once you get your statement from the credit card company, you must make at least the minimum payment due. To make that payment, you need money-you need to have currency or a checking deposit to pay the credit card company. So although you use a credit card when you buy some- thing, the credit card is not the means of payment and it is not money.

## ECONOMICS IN ACTION

## Official Measures of U.S. Money

The figure shows the relative magnitudes of the items that make up M1 and M2. Notice that M2 is almost five times as large as M1 and that currency is a small part of our money.


Two Measures of Money
M1 - Currency and traveler's checks

- Checking deposits at commercial banks, savings and loan associations, savings banks, and credit unions
M2 Ml
- Time deposits
- Savings deposits
- Money market mutual funds and other deposits

We've seen that the main component of money in the United States is deposits at banks and other depository institutions. Let's take a closer look at these institutions.

## Depository Institutions

A DepositoryInstitution is a financial firm that takes deposits from households and firms. These deposits are components of M 1 and M2. You will learn what these institutions are,
what they do, the economic benefits they bring, how they are regulated, and how they have innovated to create new financial products.

## Types of Depository Institutions

The deposits of three types of financial firms make up the nation's money. They are

- Commercial banks
- Thrift institutions
- Money market mutual funds

Commercial Banks A commercial bank is a firm that is licensed to receive deposits and make loans.In 2014, about 6,800 commercial banks operated in the United States but mergers make this numberfall each year as small banks disappear and bigbanks expand.
A few very large commercial banks offer a wide range of banking services and have extensive international operations. The largest of these banks are JPMorgan Chase, Bank of America, Wells Fargo, and Citigroup. Most commercial banks are small and serve their regional-and local communities.
The deposits of commercial banks represent 50 percent of M 1 and 71 percent of M 2 .
Thrift Institutions Savings and loan associations, savings banks, and credit unions are thrift institutions.

Savings and Loan Associations A savings and loan association (S\&L) is a depository institution that receives deposits and makes personal, commercial, and home-purchase loans.

Savings Bank A savings bank is a depository institution that accepts savings deposits and makes mostlyhome-purchase loans.

Credit Union A credit union is a depository institution owned by a social or economic group, such as a firm's employees, that accepts savings deposits and makes mostly personal loans. The deposits of the thrift institutions represent 8 percent of Ml and 13 percent of M 2 .

Money Market Mutual Funds A moneymarket mutual fund is a fund operated by a financial institution that sells shares in the fund and holds assets suchas U.S. Treasury bills and shortterm commercial bills. Money market mutual fund shares act like bank deposits. Shareholders can write checks on their money market mutual fund accounts, but there arerestrictions on most of these accounts. For example, the minimum deposit accepted might be $\$ 2,500$, and the smallest check a depositor is permitted to write might be $\$ 500$. Money market mutual funds do not feature in Ml and represent 6 percent of M 2 .

## What Depository Institutions Do

Depository institutions provide services such as check clearing, account management, and credit cards, all of which provide an income from servicefees.
But depository institutions earn most of their income by using the funds they receive from depositors to buy securities and make loans that earn a higher interest rate than that paid
to depositors. In this activity, a depository institution must perform a balancing act weighing return against risk. To see this balancing act, we'll focus on the commercial banks. A commercial bank puts the funds it receives from depositors and other funds that it borrows into threetypes of assets:

- Cash assets
- Securities
- Loans

Cash Assets A bank's cash assets consist of notes and coins in the bank's vault (called vault cash), a depositaccount at the Federal Reserve (the Fed), and loans
to other banks. The first two items, vault cash and deposits at the Fed, are the bank's
reserves. Loans to other banks earn interest and the interest rate on these loans is called the federal funds rate and the Fed sets a target for this interest rate to influence the economy. We explain how and why on pp. 798-799.
A bank holds cash assets as a first line of funds to ensure that it is always able to meet depositors' currency withdrawals and make payments to other banks. In normal times, a bank kept about a half of one percent of deposits as cash assets. But today, these assets earn interest and their quantity has swollen to 28 percent of total deposits.

Securities A bank holds U.S. government Treasury bills and commercial bills that earn a low but risk-free return, and U.S. government bonds and mortgage-backed securities that earn a higher but riskier return. Securities would be sold and converted into cash as- sets if a bank ran short of reserves.

Loans A loan is an advance of funds for a specified period of time to businesses to finance investment and to households to finance the purchase of homes, cars, and other durable goods. The outstanding balances on credit card accounts are also bank loans. Loans are a bank's riskiest and highest-earning assets: They can't be converted into cash assets until they are due to be repaid, and some borrowers default and never repay. To spread the risk on loans, some get converted to securities.
Table 25.2 provides a snapshot of the sources and uses of funds of the commercial banks in June 2014 and serves as a summary of what they do.

## Economic Benefits Provided by Depository Institutions

You've seen that a depository institution earns part of its profit because it pays a lower interest rate on deposits than what it earns on loans. What benefits do these institutions provide that make depositors willing to put up with a low interest rate and borrowers willing to pay a higher one? Depository institutions provide four benefits:

- Create liquidity
- Pool risk
- Lower the cost of borrowing
- Lower the cost of monitoring borrowers


## TABLE 25.2 Commercial Banks: Sources and Uses of Funds

|  | Funds <br> (billions of dollars) | Percentage <br> of deposits |
| :--- | :---: | :---: |
| Total funds | 14,662 | 144.3 |
| Sources |  |  |
| Deposits | 10,161 | 100.0 |
| Borrowing | 1,698 | 16.7 |
| Own capital and other sources (net) | 2,803 | 27.6 |
| Uses |  |  |
| Cash Assets |  |  |
| Securities | 2,850 | 28.0 |
| Loans | 2,809 | 27.6 |
| Other assets | 7,666 | 75.4 |
| Commercial banks get most of their funds from depositors |  |  |
| and use most of them to make loans. In normal times, banks |  |  |
| hold less than 1 percent of deposits as cash assets. But in |  |  |
| 2014, cash assets were 28 percent of deposits, most of |  |  |
| which were at the Fed earning a low interest rate. |  |  |
| Source of data: The Federal Reserve Board. The data are for June 2014. |  |  |

Create Liquidity Depository institutions create liquidity by borrowing short and lending longtaking deposits and standing ready to repay them on short notice or on demand and making loan commitmentsthat run for terms of many years.

Pool Risk A loan might not be repaid-a default. If you lend to one person who defaults, you lose the entire amount loaned. If you lend to 1,000 people (through a bank) and one person defaults, you lose almost nothing. Depository institutions pool risk.

Lower the Cost of Borrowing Imagine there are no depository institutions and a firm is looking for $\$ 1$ million to buy a new factory. It hunts around for several dozen people from whom to borrow the funds. Depository institutions lower the cost of this search. The firm gets its $\$ 1$ million from a single institution that gets deposits from a large number of people but spreads the cost of this activity over many borrowers.

Lower the Cost of Monitoring Borrowers By monitoring borrowers, a lender can encourage good decisions that prevent defaults. But this activity is costly. Imagine how costly it would be if each household that lent money to a firm incurred the costs of monitoring that firm directly. Depository institutions can perform this task at a much lower cost.

How Depository Institutions Are Regulated Depository institutions are engaged in a risky business, and a failure, especially of a large bank, would have damaging effects on the entire financial system and the economy. To make the risk of failure small, depository institutions are required to hold levels of reserves and owners' capital that equal or surpass ratios laid down by regulation. If a depository institution fails, its deposits are guaranteed up to $\$ 250,000$, per depositor per bank by the Federal Deposit Insurance Corporation or FDIC. The FDIC can take over management of a bank that appears to be heading toward failure.

## Financial Innovation

In the pursuit of larger profit, depository institutions are constantly seeking ways to improve theirproducts in a process calledfinancial innovation.
During the late 1970s, a high inflation rate sent the interest rate on home-purchase loans to 15 per- cent a year. Traditional fixed interest rate mortgages became unprofitable and variable interest rate mort- gages were introduced.
During the 2000s, when interest rates were low and depository institutions were flush with funds, sub-prime mortgages were developed. To avoid the risk of carrying these mortgages, mortgage-backed securities were developed. The original lending institution sold these securities, lowered their own expo- sure to risk, and obtained funds to make more loans.
The development of low-cost computing and communication brought financial innovations such as credit cards and daily interest deposit accounts.
Financial innovation has brought changes in the composition of money. Checking deposits at thrift institutions have become an increasing percentage of $M I$ while checking deposits at commercial banks have become a decreasing percentage. Savings deposits have decreased as a percentage of $M 2$, while time deposits and money market mutual funds have expanded. Surprisingly, the use of currency has not fallen much.

## The Federal Reserve System

The Federal Reserve System (usually called the Fed) is the central bank of the United States. A central bank is a bank's bank and a public authority that regulates a nation's depository institutions and conducts monetary policy, which means that it adjusts the quantity of money in circulation and influences interest rates. We begin by describing the structure of the Fed.

## The Structure of the Fed

Three key elements of the Fed's structure are

- The Board of Governors
- The regional Federal Reserve banks
- The Federal Open Market Committee


## ECONOMICS IN ACTION

## Commercial Banks Flush with Reserves

When Lehman Brothers (a New York investment bank) failed in October 2008, panic spread through financial markets. Banks that are normally happy to lend to each other overnight for an interest rate barely above the rate they can earn on safe Treasury bills lost confidence and the interest rate in this market shot up to 3 percentage points above the Treasury bill rate. Banks wanted to be safe and to hold cash and the Fed injected $\$ 1.5$ trillion or 17.5 percent of deposits into the banks.

From 2009 through 2014, bank reserves grew to $\$ 3$ trillion. The Fed pays interest on reserve balances, so the banks willingly hold these very large quantities of reserves.

The figure compares the commercial banks' sources and uses of funds (sources are liabilities and uses are assets) in 2008 with those in 2014.

(a) Sources of commercial bank funds

(b) Uses of commercial bank funds

Changes in the Sources and Uses of Commercial Bank Funds

The Board of Governors A seven-member board appointed by the President of the United States and confirmed by the Senate governs the Fed. Members have 14-year (staggered) terms and one seat on the board becomes vacant every two years. The President appoints one board member as chairman for a 4-year renewable term-currently Janet Yellen, a former economics professor at UCBerkeley.

The Federal Reserve Banks The nation is divided into 12 Federal Reserve districts (shown in Fig.25.1). Each district has a Federal Reserve Bank thatprovides check-clearing services to commercial banks and issues bank notes.
The Federal Reserve Bank of New York (known as the New York Fed) occupies a special place in the Federal Reserve System because it implements the Fed's policy decisions in the financial markets.

The Federal Open Market Committee (FOMC) is the main policy- making organ of the Federal Reserve System. The FOMC consists of the following voting members:

- The chairman and the other six members of the Board of Governors
- The president of the Federal Reserve Bank of New York
- The presidents of the other regional Federal Reserve banks (of whom, on a yearly rotating basis, only four vote)
The FOMC meets approximately every six weeks to review the state of the economy and to decide the actions to be carried out by the New York Fed.

FIGURE 25.1 The Federal Reserve System


The nation is divided into 12 FederalReserve districts, each having a Federal Reserve bank. (Some of the largerdistricts also have branch banks.) The Board of Governors of the Federal ReserveSystem is located in Washington, D.C.

## The Fed's Balance Sheet

The Fed influences the economy through the size and composition of its balance sheetthe assets that the Fed owns and the liabilities that it owes.

The Fed's Assets The Fed has two main assets:

1. U.S. government securities
2. Mortgage-backed securities
U.S. Government Securities The U.S. government securities held by the Fed are Treasury bonds. The Fed buys and sells these bonds in the loanable funds market (see pp. 611-612). The Fed does not buy bonds directly from the U.S. government.

Mortgage-Backed Securities Traditionally, the Fed held only U.S. government securities. But in recent years, the Fed has purchased large quantities of mortgage-backed securities to increase the supply of loanable funds (see pp. 614-617).

The Fed's Liabilities The Fed has two liabilities:

1. Currency
2. Reserves of depository institutions

Currency is the dollar bills that we use in our daily transactions. Some currency is in circulation and is a component of MI , and some is in banks andother depository institutions in their vaults and cash machines and is vault cash. (Coins are not a liability of the Fed. They are issued by the U.S. Mint.)
Reserves of Depository Institutions The Fed is the banker for the banks and the reserves that the banks deposit at the Fed are a liability of the Fed.

The Monetary Base The Fed's total liabilities make up the monetary base. That is, the monetary base is the sum of currency and the reserves of depository institutions. The Fed's assets are the sources of the monetary base. They are also called the backing for the monetary base. The Fed's liabilities are the uses of themonetary base as currency and bank reserves. Table 25.3 provides a snapshot of the sources and uses of the monetary base in June 2014.
When the Fed changes the monetary base, the quantity of money and interest rate change. You're going to see how these changes come about later in this chapter. First, we'll look at the Fed's tools that enable it to influence money and interest rates.

## The Fed's Policy Tools

The Fed influences the quantity of money and interest rates by adjusting the quantity of reserves available to the banks and the reserves the banks must hold. To do this, the Fed manipulates three tools:

- Open market operations
- Last resort loans
- Required reserve ratio


## TABLE 25.3 The Sources and Uses of the Monetary Base

| Sources <br> (billions of dollars) |  | Uses <br> (billions of dollars) |  |
| :--- | :--- | :--- | :--- |
| U.S. government <br> securities | 2,330 | Currency | 1,280 |
| Mortgage-backed <br> securities | $\underline{1,618}$ | Reserves of <br> depository <br> institutions | $\underline{2,668}$ |
| Monetary base | $\underline{\underline{3,948}}$ | Monetary base | $\underline{\underline{3,948}}$ |

Source of dato: The Federal Reserve Board. The data are for June 2014.

Open Market Operations An open market operation is \the purchase or sale of securities by the Fed in the loanable funds market. When the Fed buyssecurities, it pays for them with newly created bank reserves. When the Fed sells securities, the Fed is paid with reserves held by banks. So open market operations directly influence the reserves of banks. By changingthe quantity of bank reserves, the Fed changes the quantity of monetary base, which influences the quantity of money.

An Open Market Purchase To see how an open market operation changes bank reserves, suppose the Fed buys $\$ 100$ million of government securities from theBank of America. When the Fed makes this transaction, two things happen:

1. The Bank of America has $\$ 100$ million less securities, and the Fed has $\$ 100$ million more securities
2. The Fed pays for the securities by placing $\$ 100$ million in the Bank of America's deposit account at the Fed.
Figure 25.2 shows the effects of these actions on the balance sheets of the Fed and the Bank of America. Ownership of the securities passes from the Bank of America to the Fed, so the Bank of America's assets decrease by $\$ 100$ million and the Fed's assets increase by $\$ 100$ million, as shown by the blue arrow running from the Bank of America to the Fed. The Fed pays for the securities by placing $\$ 100$ million in the Bank of America's reserve account at the Fed, as shown by the green arrow running from the Fed to the Bank of America. The Fed's assets and liabilities increase by $\$ 100$ million. The Bank of America's total assets are unchanged: It sold securities to increase its reserves.

An Open Market Sale If the Fed sells $\$ 100$ million of government securities to the Bank of America in the open market,

1. The Bank of America has $\$ 100$ million more securities, and the Fed has $\$ 100$ million less securities.
2. The Bank of America pays for the securities by using $\$ 100$ million of its reserve deposit at the Fed.

You can follow the effects of these actions on the balance sheets of the Fed and the Bank of America by reversing the arrows and the plus and minus signs in Fig. 25.2. Ownership of the securities passes from the Fed to the Bank of America, so the Fed's assets decrease by \$100 million and the Bank of America's assets increase by $\$ 100$ million

FIGURE 25.2 The Fed Buys Securities in the Open Market

Federal Reserve Bank of New York


Bank of America


When the Fed buys securities in the open market, it creates bank reserves. The Fed's assets and liabilities increase, and the Bank of America exchanges securities for reserves.
The Bank of America uses $\$ 100$ million of its reserves to pay for the securities. Both the Fed's assets and liabilities decrease by \$100 million. The Bank of America's total assets are unchanged: It has used reserves to buy securities. The New York Fed conducts these open-market transactions on directions from the FOMC.

Last Resort Loans The Fed is the lender of last resort, which means that if a bank is short of reserves, it can borrow from the Fed. But the Fed sets the interest rate on last resort loans and this interest rate is calledthe discount rate.
During the period since August 2007 when the first effects of the financial crisis started to be felt, the Fed has been especially active as lender of last resort and, with the U.S. Treasury, has created a number ofnew lending facilities and initiatives to prevent banksfrom failing. Required Reserve Ratio The required reserve ratio is the minimum percentage of deposits that depository institutions are required to hold as reserves. In 2014, required reserves were 3 percent of checking deposits between $\$ 13.3$ million and $\$ 89$ million and 10 percent of checking deposits in excess of $\$ 89$ million. If the Fed requires the banks to hold more reserves, they must cut their lending.
Next, we're going to see how the banking system- the banks and the Fed-creates money and how thequantity of money changes when the Fed changes the monetary base.

## How Banks Create Money

Banks create money. But this doesn't mean that theyhave smoke-filled back rooms in which counterfeiters are busily working. Remember, money is both currency and bank deposits. What banks create is deposits, and they do so by making loans.

## Creating Deposits by Making Loans

The easiest way to see that banks create deposits is to think about what happens when Andy, who has a Visa card issued by Citibank, uses his card to buy a tank of gas from Chevron. When Andy swipes his card, two financial transactions occur. First, Andy takes a loan from Citibank and obligates himself to repay the loan at a later date. Second, a message is transmitted to Chevron's bank and the bank credits Chevron's account with the amount of Andy's purchase (minus the bank's commission).
For now, let's assume that Chevron, like Andy, banks at Citibank so that the two transactions we've just described both occur at the one bank.
You can see that these transactions have created a bank deposit and a loan. Andy has increased the size of his loan (his credit card balance), and Chevron has increased the size of its bank deposit. Because bank deposits are money, Citibank has created money. If, as we've just assumed, Andy and Chevron use the same bank, no further transactions take place. But if two banks are involved, there is another transaction. To see this additional transaction and its effects, assume that Chevron's bank is Bank of America.
Tofully settle the payment for Andy's gas purchase, Citibank must pay Bank of America. To make this payment, Citibank uses its reserves. Citibank's reserves decrease by the amount of its loan to Andy; Bank of America's reserves increase by the amount that Chevron's deposit increases. Payments like this one between the banks are made at the end of the business day. So, at the end of the business day the banking system as a whole has an increase in loans and deposits but no change in reserves. Three factors limit the quantity of loans and deposits that the banking system can create through transactions like Andy's. Theyare

- The monetary base
- Desired reserves
- Desired currency holding

The Monetary Base You've seen that the monetary base is the sum of Federal Reserve notes, coins, and banks' deposits at the Fed. The size of the monetary base limits the total quantity of money that the banking system can create. The reason is that banks have a desired level of reserves, households and firms have a desired holding of currency, and both of these desired holdings of the monetary base depend on the quantity of deposits.

Desired Reserves A bank's desired reserves are the reserves that it plans to hold. They contrast with a bank's required reserves, which is the minimum quantity of reserves that a bank must hold.

The quantity of desired reserves depends on the level of deposits and is determined by the desired reserve ratio - the ratio of reserves to deposits that the banks plan to hold. The desired reserve ratio exceeds the required reserve ratio by an amount that the banks determine to be prudent on the basis of their daily business requirements and in the light of the current outlook in financial markets.

Desired Currency Holding The proportions of money held as currency and bank deposits-the ratio of currency to deposits-depend on how households and firms choose to make payments: Whether they plan to use currency or debit cards and checks. Choices about how to make payments change slowly so the ratio of desired currency to deposits also changes slowly, and at any given time this ratio is fixed. If bank deposits increase, desired currency holding also increases. For this reason, when banks make loans that increase deposits, some currency leaves the banks-the banking system leaks reserves. We call the leakage of bank reserves into currency the currency drain, and we call the ratio of currency to deposits the currency drain ratio.

We've sketched the way that a loan creates a deposit and described the three factors that limit the amount of loans and deposits that can be created. We're nowgoing to examine the money creation process moreclosely and discover a money multiplier.

## The Monetary Creation Process

The money creation process begins with an increase in the monetary base, which occurs if the Fed con- ducts an open market operation in which it buys securities from banks and other institutions. The Fed pays for the securities it buys with newly created bank reserves. When the Fed buys securities from a bank, the bank's reserves increase but its deposits don't change.So the bank has excess reserves. A bank's excess reserves are its actual reserves minus its desired reserves.
When a bank has excess reserves, it makes loans and creates deposits. When the entire banking system hasexcess reserves, total loans and deposits increase and the quantity of money increases. One bank can make a loan and get rid of excess reserves. But the banking system as a whole can'tget rid of excess reserves so easily. When the banks make loans and create deposits, the extra deposits lower excess reserves for two reasons. First, the increase in deposits increases desired reserves. Second, a currency drain decreases total reserves. But excess reserves don't completely disappear. So the banks lend some more and the process repeats.
As the process of making loans and increasing deposits repeats, desired reserves increase, total reserves decrease through the currency drain, and eventually enough new deposits have been created to use all thenew monetary base. Figure 25.3 summarizes one round in the process we've just described. The sequence has the following eight steps:

1. Banks have excess reserves.
2. Banks lend excess reserves.
3. The quantity of money increases.
4. New money is used to make payments.
5. Some of the new money remains on deposit.
6. Some of the new money is a currency drain.
7. Desired reserves increase because deposits have increased.
8. Excess reserves decrease.

If the Fed sells securities in an open market operation, then banks have negative excess reserves- they are short of reserves. When the banks are short of reserves, loans and deposits decrease and the process we've described above works in a downward direction until desired reserves plus desired currency holding has decreased by an amount equal to the decrease in monetarybase. A money multiplier determines the change in the quantity of money that results from a change in the monetary base.

## ECONOMICS IN ACTION

## The Variable Money Multipliers

We can measure the money multiplier, other things remaining the same, as the ratio of the quantity of money ( M I or M 2 ) to the monetary base. Innormal times, these ratios (and the money multipliers) change slowly.
In the early 1990s, the M 1 multiplier - the ratio of M 1 to the monetary base - was about 3 and the M2 multiplier - the ratio of M2 to the monetary base - was about 12. Through the 1990s and 2000s, the currency drain ratio gradually increased and the money multipliers decreased. By 2007, the M I multiplier was 2 and the M 2 multiplier was 9. Since 2008, the unprecedented increase in the monetary base has been willingly held by banks as reserves. In an environment of uncertainty, desired reserves increased by a similar amount to the in- crease in actual reserves. The quantity of money changed by much less than the change in monetary base.

The Money Multiplier is the ratio of the change in the quantity of money to the change in monetary base. For example, if a $\$ 1$ million increase in the monetary base increases the quantity of money by $\$ 2.5$ million, then the money multiplier is 2.5 . The smaller the banks' desired reserve ratio and the smaller the currency drain ratio, the larger is the money multiplier. (See the Mathematical Note on pp. 650-651 for details on the money multiplier).

## The Money Market

There is no limit to the amount of money we would like to receive in payment for our labor or as interest on our savings. But there is a limit to how big an inventory of money we would like to hold and neither spend nor use to buy assets that generate an income. The quantity of money demanded is the inventory of money that people plan to hold on any given day. It is the quantity of money in our wallets and in our deposit accounts at banks. The quantity of money held must equal the quantity supplied, and the forces that bring about this equality in the money market have powerful effects on the economy, as you will see in the rest of this chapter. But first, we need to explain what determines the amount of money that people plan to hold.

## The Influences on Money Holding

The quantity of money that people plan to hold depends on four main factors:

- The price level
- The nominal interest rate
- Real GDP
- Financial innovation

The Price Level The quantity of money measured in dollars is nominal money. The quantity of nominal money demanded is proportional to the price level, other things remaining the same. If the price level rises by 10 percent, people hold 10 percent more nominal money than before, other things remaining the same. If you hold $\$ 20$ to buy your weekly movies and soda, you will increase your money holding to $\$ 22$ if the prices of movies and soda-and your wage rate-increase by 10 percent.
The quantity of money measured in constant dollars (for example, in 2009 dollars) is real money. Real money is equal to nominal money divided by theprice level and is the quantity of money measured in terms of what it will buy. Inthe above example, when the price level rises by 10 percent and you increase your money holding by 10 percent, your real money holding is constant. Your $\$ 22$ at the new price level buys the same quantity of goods and is the same quantity of real money as your $\$ 20$ at the original price level. The quantity of real money demanded is independent of the price level.

The Nominal Interest rate A fundamental principle of economics is that as the opportunity cost of some- thing increases, people try to find substitutes for it. Money is no exception. The higher the opportunity cost of holding money, other things remaining the same, the smaller is the quantity of real money demanded. The nominal interest rate on other assets minus the nominal interest rate on money is the opportunity cost of holding money. The interest rate that you earn on currency and checking deposits is zero. So the opportunity cost of holding these items is the nominal interest rate on other assets such as a savings bond or Treasury bill.
By holding money instead, you forgo the interest that you otherwise would have received. Money loses value because of inflation, so why isn't the inflation rate part of the cost of holding money? It is. Other things remaining the same, the higher the expected inflation rate, the higher is the nominal interest rate.

Real GDP The quantity of money that households and firms plan to hold depends on the amount they are spending. The quantity of money demanded in the economy as a whole depends on aggregate expenditure-real GDP.
Again, suppose that you hold an average of $\$ 20$ to finance your weekly purchases of movies and soda. Now imagine that the prices of these goods and of all other goods remain constant but that your income increases. As a consequence, you now buy more goods and services and you also keep a larger amount of money on hand to finance your higher volume of expenditure.

Financial Innovation Technological change and the arrival of new financial products influence the quantity of money held. Financial innovations include

1. Daily interest checking deposits
2. Automatic transfers between checking and saving deposits
3. Automatic teller machines
4. Credit cards and debit cards
5. Internet banking and bill paying

These innovations have occurred because of the development of computing power that has lowered the cost of calculations and record keeping.
We summarize the effects of the influences on money holding by using a demand for money curve.

The Demand for Money is the relationship between the quantity of real money demanded and the nominal interest rate when all other influences on the amount of money that people wish to hold remain the same.
Figure 25.4 shows a demand for money curve, MD. When the interest rate rises, other things remaining the same, the opportunity cost of holding money rises and the quantity of real money demanded decreases-there is a movement up alongthe demand for money curve. Similarly, when the interest rate falls, the opportunity cost of holding money falls, and the quantity of real money demanded increases-there is a movement down along the demand for money curve.
When any influence on money holding other than the interest rate changes, there is a change in the demand for money and the demand for money curve shifts. Let's study these shifts.

## Shifts in the Demand for Money Curve

A change in real GDP or financial innovation changes the demand for money and shifts the demand for money curve.
Figure 25.5 illustrates the change in the demand for money. A decrease in real GDP decreases the demand for money and shifts the demand for money curve leftward from $M D_{0}$ to $M D_{1}$. An increase in real GDP has the opposite effect: It increases the demand for money and shifts the demand for money curve rightward from $M D_{0}$ to $M D_{2}$, The influence of financial innovation on the demand for money curve is more complicated. It decreases the demand for currency and might increase the demand for some types of deposits and decrease the demand for others. But generally, financial innovation decreases the demand for money.
Changes in real GDP and financial innovation have brought large shifts in the demand for money in the United States.

FIGURE 25.4 The Demand for Money


The demand for money curve, MD, shows the relationship between the quantity of real money that people plan to hold and the nominal interest rate, other things remaining the same. The interest rate is the opportunity cost of holding money. A change in the interest rate brings a movement along the demand for money curve.
figure 25.5 Changes in the Demand for Money


A decrease in real GDP decreases the demand for money. The demand for money curve shifts leffward from $M D_{0}$ to $M D_{1}$. An increase in real GDP increases the demand for money. The demand for money curve shifts rightward from $M D_{0}$ to $M D_{2}$. Financial innovation generally decreases the demand for money.

## Money Market Equilibrium

You now know what determines the demand for money, and you've seen how the banking system creates money. Let's now see how the money market reaches an equilibrium. Money market equilibrium occurs when the quantity of money demanded equals the quantity of money supplied. The adjustments that occur tobring money market equilibrium are fundamentally different in the short run and the long run.

Short-Run Equilibrium Thequantity of moneysup- plied is determined by the actions of the banks and the Fed. As the Fed adjusts the quantity of money, the interest rate changes. In Fig. 25.6, the Fed uses open market operations to make the quantity of real money supplied equal to $\$ 3.0$ trillion and the supply of money curve MS.
With demand for money curve MD, the equilibrium interest rate is 5 percent a year. If the interest rate were 4 percent a year, people would want to hold more money than is available.
figure 25.6 Money Market Equilibrium


Money market equilibrium occurs when the quantity of money demanded equals the quantity supplied. In the short run, real GDP determines the demand for money curve, MD, and the Fed determines the quantity of real money supplied and the supply of money curve, MS. The interest rate adjusts to achieve equilibrium, here 5 percent a year.
They would sell bonds, bid down their price, and the interest rate would rise. If the interest rate were 6 per-cent a year, people would want to hold less money than is available. They would buy bonds, bid up their price, and the interest rate would fall.

The Short-Run Effect of a Change in the Quantity of Money Starting from a short-run equilibrium, if the Fed increases the quantity of money, people find themselves holding more money than the quantity demanded. With a surplus of money holding, people enter the loanable funds market and buy bonds. The increase in demand for bonds raises the price of a bond and lowers the interest rate (refresh your memory by looking at Chapter 24, p. 610). If the Fed decreases the quantity of money, people find themselves holding less money than the quantity demanded. They now enter the loanable funds market to sell bonds. The decrease in the demand for bonds lowers their price and raises the interest rate.
Figure 25.7 illustrates the effects of the changes in the quantity of money that we've just described.
When the supply of money curve shifts rightward from MSO to MS 1, the interest rate falls to 4 percent a year; when the supply of money curve shifts leftward • to MS2, the interest rate rises to 6 percent a year.

Long-Run Equilibrium You've just seen how the nominal interest rate is determined in the money marketat the level that makes the quantity of money demanded equal the quantity supplied by the Fed.
You learned in Chapter 24 (on p. 615) that the real interest rate is determined in the loanable funds market at the level that makes the quantity of loanable funds demanded equal the quantity of loanable funds supplied. You also learned in Chapter 24 (on p. 612) that the real interest rate equals the nominal interest rate minus the inflation rate.
When the inflation rate equals the expected (or forecasted) inflation rate and when real GDP equals, potential GDP, the money market, the loanable funds market, the goods market, and the labor market are in long-run equilibrium - the economy is in long-run equilibrium.
If in long-run equilibrium, the Fed increases the quantity of money, eventually a new long-run equilibrium is reached in which nothing real has changed Real GDP, employment, the real quantity of money, and the real interest rate all return to their original levels. But something does change: the price level. The price level rises by the same percentage as the rise in the quantity of money. Why does this outcome occur in the long run?

FIGURE 25.7 A Change in the Quantity of Money


An increase in the quantity of money increases the supply of money. The supply of money curve shifts from $M S_{0}$ to $M S_{1}$ and the interest rate falls. A decrease in the quantity of money decreases the supply of money. The supply of money curve shifts from $M S_{0}$ to $M S_{2}$ and the interest rate rises.
The reason is that real GDP and employment are determined by the demand for labor, the supply of labor, and the production function - the real forces described in Chapter

23 (pp.584-586); and the real interest rate is determined by the demand for and supply of (real) loanable funds-the real forces described in Chapter 24 (pp. 613-615). The only variable that is free to respond to a change in the supply of money in the long run is the price level. The price level adjusts to make the quantity of real money supplied equal to the quantity demanded.
So when the Fed changes the nominal quantity of money, in the long run the price level changes by a percentage equal to the percentage change in the quantity of nominal money. In the long run, the change in the price level is proportional to the change in the quantity of money.

The Transition from the Short Run to the Long Run How does the economy move from the first short-run response to an increase in the quantity of money to the long-run response? The adjustment process is lengthy and complex. Here, we'll only provide a sketch of the process. A more thorough account must wait until you've studied Chapter 27.
We start out in long-run equilibrium and the Fed increases the quantity of money by 10 percent. Here are the steps in what happens next. First, the nominal interest rate falls (just like you saw on p. 644 and in Fig. 25.6). The real interest rate falls too, as people try to get rid of their excess money holdings and buy bonds. With a lower real interest rate, people want to borrow and spend more. Firms want to borrow to invest and households want to borrow to invest in bigger homes or to buy more consumer goods.
The increase in the demand for goods cannot be met by an increase in supply because the economy is already at full employment. So there is a general shortage of all kinds of goods and services.
The shortage of goods and services forces the price level torise. As the price level rises, the real quantity of money decreases. The decrease in the quantity of real money raises the nominal interest rate and the real interest rate. As the interest rate rises, spending plans are cut back, and eventually the original full-employment equilibrium is restored. At the new long-run equilibrium, the price level has risen by 10 percent and nothing real has changed.

## The Quantity Theory of Money

In the long run, the price level adjusts to make the quantity of real money demanded equal the quantity supplied. Aspecial theory of the price level and inflation - the quantity theory of money - explains this long-run adjustment of the price level. The quantity theory of money is the proposition that in the long run, an increase in the quantity of money brings an equal percentage increase in the price level. To explain the quantity theory of money, we first need to define the velocity of circulation.

The velocity of circulation is the average number of times a dollar of money is used annually to buy thegoods and services that make up GDP. But GDP equals the price level $(P)$ multiplied by real GDP (Y).

That is, $G D P=P Y$.
Call the quantity of money $M$. The velocity of circulation, $V$, is determined by the equation

$$
V=P Y / M .
$$

For example, if GDP is $\$ 1,000$ billion ( $P Y=\$ 1,000$ billion) and the quantity of money is $\$ 250$ billion, then the velocity of circulation is 4 .
From the definition of the velocity of circulation, the equation of exchange tells us how $M$, $V, P$, and $Y$ are connected. This equation is

$$
M V=P Y .
$$

Given the definition of the velocity of circulation, the equation of exchange is always trueit is true by definition. It becomes the quantity theory of money if the quantity of money does not influence the velocity of circulation or real GDP. In this case, the equation of exchange tells us that in the long run, the price level is determined by the quantity of money. That is,

$$
P=M(V / Y),
$$

where (WY) is independent of $M$. So a change in $M$ brings a proportional change in $P$. We can also express the equation of exchange in growth rates, in which form it states that

Solving this equation for the inflation rate gives

$$
\underset{\text { rate }}{\text { Inflation }}=\underset{\text { growth rate }}{\text { Money }}+\begin{gathered}
\text { Rate of } \\
\text { velocity } \\
\text { change }
\end{gathered}-\underset{\text { Real GDP }}{\text { growth rate }}
$$

In the long run, the rate of velocity change is not influenced by the money growth rate. More strongly, in the long run, the rate of velocity change is approximately zero. With this assumption, the inflation rate in the long run is determined as

$$
\underset{\text { rate }}{\text { Inflation }}=\frac{\text { Money }}{\text { growth rate }}-\underset{\text { Real GDP }}{\text { growth rate }} .
$$

In the long run, fluctuations in the money growth rate minus the real GDP growth rate bring equal fluctuations in the inflation rate.
Also, in the long run, with the economy at full employment, real GDP equals potential GDP, so the real GDP growth rate equals the potential GDPgrowth rate. This growth rate might be influenced by inflation, but the influence is most likely small and the quantity theory assumes that it is zero. So the real GDP growth rate is given and doesn't change when the money growth rate changes-inflation is correlated with money growth.
You now know what money is, how the banks create it, and how the quantity of money influences the nominal interest rate in the short run and the price level in the long run. Economics in the News on pp. 648-649 looks at what will happen to the quantity of money when interest rates begin to rise.

## ECONOMICS IN ACTION

## Does the Quantity Theory Work?

On average, as predicted by the quantity theory of money, the inflation rate fluctuates in line with fluctuations in the money growth rate minus the real GDP growth rate. Figure 1 shows the relationship between money growth (M2 definition) and inflation in the United States. You can see a clear relationship between the two variables.


Figure 1 U.S. Money Growth and Inflation
Sources of data: Federal Reserve and Bureau of Lobor Statistics.

International data also support the quantity theory. Figure 2 shows a scatter diagram of the inflation rate and the money growth rate in 134 countries and Fig. 3 shows the inflation rate and money growth rate in countries with inflation rates below 20 percent a year. You can see a general tendency for money growth and inflation to be correlated, but the quantity theory (the red line) does not predict inflation precisely.


Figure 2134 Countries: 1990-2005

The correlation between money growth and inflation isn't perfect, and the correlation does not tell us that money growth causes inflation. Money growth might cause inflation; inflation might cause money growth; or some third variable might cause both inflation and money growth. Other evidence does confirm, though, that causation runs from money growth to inflation


Figure 3 Lower-Inflation Countries: 1990-2005

## ECONOMIC ANALYSIS

- A global financial crisis started quietly in the summer of 2007.
- In 2006, on the eve of the crisis, the interest rate on U.S. Treasury bills was 4.7 percent and the quantity of M2 money was a bit more than $\$ 7$ trillion.
- Between 2006 and 2009, the interest rate fell to almost zero and it remained near zero through 2014 (see Fig. 1).
- Between 2006 and 2014, the quantity of M2 money increased by $\$ 2.8$ trillion, an annual average growth rate of 6.6 percent (see Fig. 2).
- Of the $\$ 2.8$ trillion increase in M2, bank deposits increased by $\$ 2.4$ trillion and currency in circulation increased by $\$ 0.4$ trillion.
- You've learned in this chapter that the quantity of money demanded depends inversely on the interest rate. So it is to be expected that a falling interest rate would bring an increasing quantity of money.
- Figure 3 shows this inverse relationship in the past 20 years (1993-2013). The demand for money is influenced by GDP, so the graph removes the influence of GDP by measuring the quantity of real M2 as a percentage of GDP.
- In Fig. 3, each red dot represents the quantity of money and the interest rate in a given year and the blue curve is the demand for M2 curve.
- Figure 3 highlights the bankers' concerns discussed in the news article, which reports that the banks expect deposits to decrease by $\$ 1$ trillion if interest rates return to normal.
- The news article does not tell us what the bankers regard as normal. We can use the demand for M2 curve in Fig. 3 to find the quantity of money that will be held at different possible normal interest rates.
- At an interest rate of 1 percent per year the quantity of M2 demanded in Fig. 3 is 52.5 percent of GDP or $\$ 8.2$ trillion.
- At an interest rate of zero, the quantity of M2 demanded is 64 percent of GDP, which is $\$ 10$ trillion.
- So, based on the demand for M2 curve in Fig. 3, the quantity of $M 2$ demanded will decrease by $\$ 1.8$ trillion if the interest rate rises to 1 percent per year.
- Most of the decrease in the quantity of M2 demanded will be a decrease in bank deposits.
- The outcome for which the bankers are bracing themselves looks optimistic!


Figure 1 The Falling Interest Rate


Figure 2 The Increasing Quantity of M2


Figure 3 The Demand for M2

## CHAPTER 26: The Exchange Rate and the Balance of Payments

After studying this chapter, you will be able to:

- Explain how the exchange rate is determined
- Explain interest rate parity and purchasing power parity
- Describe the alternative exchange rate policies and explain their effects
- Describe the balance of payments accounts and explain what causes an international deficit.

The dollar $(\$)$, the euro $(€)$, and the yen $(¥)$ are of the world's monies and most international paymentsare made using one of them. But the world has morethan IOO different monies. What determines the value of the dollar in terms of other kinds of money?
For almost thirty years, foreign entrepreneurs have roamed the United States with giant, virtual shopping carts buying U.S. businesses. Why?
In this chapter, you're going to discover the answers to these questions. In Economics in the News at the end of the chapter, we'll look at the rising dollar in the summer of 2014.

## The Foreign Exchange Market

When Wal-Mart imports Blu-ray players from Japan, it pays for them using Japanese yen. And when Japan Airlines buys an airplane from Boeing, it pays using U.S. dollars. When you take a European holiday, you pay for the holiday with euros. Whenever people buy things from another country, they use the currency of that country to make the transaction.
It doesn't make any difference what the item is that is being traded internationally. It might be a Blu-ray player, an airplane, an international holiday, insurance or banking services, real estate, the stocks and bonds of a government or corporation, or even an entire business.
Foreign money is just like U.S. money. It consists of notes and coins issued by a central bank and mint and deposits in banks and other depository institutions. When we described U.S. money in Chapter 25, we distinguished between currency (notes and coins) and deposits. But when we talk about foreign money, we refer to it as foreign currency. Foreign currency is the money of other countries regardless of whether that money is in the form of notes, coins, or bank deposits.
We buy these foreign currencies and foreigners buy U.S. dollars in the foreign exchange market.

## Trading Currencies

The currency of one country is exchanged for the currency of another in the foreign exchangemarket. The foreign exchange market is not a place like a downtown flea market or a fruit and vegetable market. The foreign exchange market is made up of thousands of people-importers and exporters, banks, international investors and speculators, international travelers, and specialist traders called foreign exchange brokers.

The foreign exchange market opens on Monday morning in Sydney, Australia, and Hong Kong, which is still Sunday evening in New York. As the day advances, markets open in Singapore, Tokyo, Bahrain, Frankfurt, London, New York, Chicago, and San Francisco. As the West Coast markets close, Sydney is only an hour away from opening for the next day of business. The sun barely sets in the foreign exchange market. Dealers around the world are in continual Internet contact, and on a typical day in 2014, $\$ 5.3$ trillion (of all currencies) were traded in the foreign exchange market - that's $\$ 6$ million every second.

## Exchange Rates

An exchange rate is the price at which one currency exchanges for another currency in the foreign exchange market. For example, on August 25, 2014, $\$ 1$ would buy 104 Japanese yen or 76 euro cents. So the exchange rate was 104 yen per dollar or, equivalently, 76 euro cents per dollar.
The exchange rate fluctuates. Sometimes it rises and sometimes it falls. A rise in the exchange rate is called an appreciation of the dollar, and a fall in the exchange rate is called a depreciation of the dollar. For) example, when the exchange rate rises from 104 yen to 110 yen per dollar, the dollar appreciates against the yen; when the exchange rate falls from 110 yen to 104 yen per dollar, the dollar depreciates against the yen.
Economics in Action on p. 659 shows the fluctuations of the U.S. dollar against three currencies from 2000 to 2014.

## Questions About the U.S. Dollar Exchange Rate

The performance of the U.S. dollar in the foreign exchange market raises a number of questions that we address in this chapter.
First, how is the exchange rate determined? Why does the U.S. dollar sometimes appreciate and at other times depreciate?
Second, how do the Fed and other central banks operate in the foreign exchange market? In particular, •how was the exchange rate between the U.S. dollar and the Chinese yuan fixed and why did it remain constant for many years?
Third, how do exchange rate fluctuations influence our international trade and international payments? In particular, could we eliminate, or at least decrease, our international deficit by changing the exchange rate? Would an appreciation of the yuan change the balance of trade and payments between the United States and China? We begin by learning how trading in the foreign exchange market determines the exchange rate.

## An Exchange Rate is a Price

An exchange rate is a price - the price of one currency in terms of another. And like all prices, an exchange rate is determined in a market-the foreign $\cdot ;$ exchange market."
The U.S. dollar trades in the foreign exchange market and is supplied and demanded by tens of thousands of traders every hour of every business day. Because it has many traders and no restrictions on who may trade, the foreign exchange market is a competitive market.

Ina competitive market, demand and supply determine the price. So to understand the forces that determine the exchange rate, we need to study the factors that influence demand and supply in the foreign exchange market. But there is a feature of the foreign exchange market that makes it special.

## The Demand for One Money Is the Supply of Another Money

When people who are holding the money of some other country want to exchange it for U.S. dollars, they demand U.S. dollars and supply that other country's money. And when people who are holding U.S. dollars want to exchange them for the money of some other country, they supply U.S. dollars and demand that other country's money.
So the factors that influence the demand for U.S. dollars also influence the supply of European euros, or Japanese yen, or Chinese yuan. And the factors that influence the demand for that other country's money also influence the supply of U.S. dollars. We'll first look at the influences on the demand for U.S. dollars in the foreign exchange market.

## Demand in the Foreign Exchange Market

People buy U.S. dollars in the foreign exchange market so that they can buy U.S.-produced goods and services-U.S. exports. They also buy U.S. dollars so that they can buy U.S. assets such as bonds, stocks, businesses, and real estate or so that they can keep part of their money holding in a U.S. dollar bank account.
The quantity of U.S. dollars demanded in the foreign exchange market is the amount that traders plan to buy during a given time period at a given exchange rate. This quantity depends on many factors, but the main ones are

1. The exchange rate
2. World demand for U.S. exports
3. Interest rates in the United States and other countries
4. The expected future exchange rate

We look first at the relationship between the quantity of U.S. dollars demanded in the foreign exchange market and the exchange rate when the other three influences remain the same.

The Law of Demand for Foreign Exchange The law of demand applies to U.S. dollars just as it does to anything else that people value. Other things remaining the same, the higher the exchange rate, the smaller is the quantity of U.S. dollars demanded in the foreign exchange market. For example, if the market price of the U.S. dollar rises from 100 yen to 120 yen but nothing else changes, the quantity of U.S. dollars that people plan to buy in the foreign exchange market decreases. The exchange rate influences thequantity of U.S. dollars demanded for two reasons:

- Exports effect
- Expected profit effect


## ECONOMICS IN ACTION

## The U.S. Dollar: More Down than Up

The figure shows the U.S. dollar exchange rate against the three currencies that feature prominently in U.S. imports- the Chinese yuan, the European euro, and the Japanese yen-between 2000 and 2014.

Against the yuan, the dollar was constant before 2005 and since then it has depreciated. Against the yen and the euro, the dollar appreciated before 2002. Since then, the dollar depreciated against the yen through 2012 and then appreciated. Against the euro, the dollar depreciated from 2002 through 2008 and then appreciated.

Notice the high-frequency fluctuations (rapid brief up and down movements) of the dollar against the euro and the yen compared to the smooth changes against the yuan. Think about why that might be, and we'll check your answer later in this chapter.


The U.S. Dollar Against Three Currencies
Exports Effect The larger the value of U.S. exports, the larger is the quantity of U.S. dollars demanded by the buyers of U.S. exports in the foreign exchange market. But the value of U.S. exports depends on the prices of U.S.-produced goods and services ex-pressed in the currency of the foreign buyer. And these prices depend on the exchange rate. The lower the exchange rate, other things remaining the same, the lower are the prices of U.S.-produced goods and services to foreigners and the greater is the volume of U.S. exports. So if the
exchange rate falls (and other influences remain the same), the quantity of U.S. dollars demanded in the foreign exchange market increases. To see the exports effect at work, think about orders for Boeing's new 787 Dreamliner. If the price of this airplane is $\$ 100$ million and the exchange rate is 90 euro cents per U.S. dollar, its price to KLM, a European airline, is $€ 90$ million. KLM decides that this price is too high, so it doesn't buy a Dreamliner. If the exchange rate falls to 80 euro cents per U.S. dollar and other things remain the same, the price of a Dreamliner falls to $€ 80$ million, so KLM decides to buy one and enters the foreign exchange market to buy 100 million U.S. dollars.

Expected Profit Effect The larger the expected profit from holding U.S. dollars, the greater is the quantity of U.S. dollars demanded in the foreign exchange market. But expected profit depends on the exchangerate. For a given expected future exchange rate, the lower the exchange rate today, the larger is the expected profit from buying U.S. dollars today and holding them, so the greater is the quantity of U.S. dollars demanded in the foreign exchange market today. Let's look at an example.
U.S. Suppose that Mitsubishi Bank, a Japanese bank, expects the exchange rate to be 120 yen per U.S. dollar at the end of the year. If today's exchange rate is also 120 yen per U.S. dollar, Mitsubishi Bank expects no profit from buying U.S. dollars and holding them until the end of the year. But if today's exchange rate is 100 yen per U.S. dollar and Mitsubishi Bank buys U.S. dollars, it expects to sell those dollars at the end of the year for 120 yen per dollar and make a profitof 20 yen on each U.S. dollar bought.
The lower the exchange rate today, other things remaining the same, the greater is the expected profit from holding U.S. dollars, so the greater is the quantity of U.S. dollars demanded in the foreign exchange market today.

## Demand Curve for U.S. Dollars

Figure 26.1 shows the demand curve for U.S. dollars in the foreign exchange market. A change in the exchange rate, other things remaining the same, brings a change in the quantity of U.S. dollars demanded and a movement along the demand curve. The arrows show such movements.
We will look at the factors that change demandin the next section of this chapter. Before doing that, let's see what determines the supply of U.S. dollars.

Supply in the Foreign Exchange Market People and businesses sell U.S. dollars and buy other currencies so that they can buy foreign-produced goods and services-U.S. imports. They also sell U.S. dollars and buy foreign currencies so that they can buy foreign assets such as bonds, stocks, businesses, and real estate or so that they can hold part of their money in bankdeposits denominated in a foreign currency.

FIGURE 26.1 The Demand for U.S. Dollars


The quantity of U.S. dollars demanded depends on the exchange rate. Other things remaining the same, if the exchange rate rises, the quantity of U.S. dollars demanded decreases and there is a movement up along the demand curve for U.S. dollars. If the exchange rate falls, the quantily of U.S. dollars demanded increases and there is a movement down along the demand curve for U.S. dollars.

The quantity of U.S. dollars supplied in the foreign exchange market is the amount that traders plan to sell during a given time period at a given exchange rate. This quantity depends on many factors, but the main ones are

- The exchange rate
- U.S. demand for imports
- Interest rates in the United States and other countries
- The expected future exchange rate

Let's look at the law of supply in the foreign exchange market-the relationship between the quantity of U.S. dollars supplied in the foreign exchange market and the exchange rate when the other three influences remain the same.

The Law of Supply of Foreign Exchange Other things remaining the same, the higher the exchange rate, the greater is the quantity of U.S. dollars sup- plied in the foreign exchange market. For example, if the exchange rate rises from 100 yen to 120 yen per U.S. dollar and other things remain the same, the quantity of U.S. dollars that people plan to sell in the foreign exchange market increases.

The exchange rate influences the quantity of dollars supplied for two reasons:

- Imports effect
- Expected profit effect

FIGURE 26.2 The Supply of U.S. Dollars


The quantity of U.S. dollars supplied depends on the exchange rate. Other things remaining the same, if the exchange rate rises, the quantity of U.S. dollars supplied increases and there is a movement up along the supply curve of U.S. dollars. If the exchange rate falls, the quantity of U.S. dollars supplied decreases and there is a movement down along the supply curve of U.S. dollars.

Imports Effect The larger the value of U.S. imports, the larger is the quantity of U.S. dollars supplied in the foreign exchange market. But the value of U.S. imports depends on the prices of foreign-produced goods and services expressed in U.S. dollars. These prices depend on the exchange rate. The higher the exchange rate, other things remaining the same, the lower are the prices of foreign-produced goods and services to Americans and the greater is the volume of U.S. imports. So if the exchange rate rises (and other influences remain the same), the quantity of U.S. dollars supplied in the foreign exchange marketincreases.

Expected Profit Effect This effect works just like that on the demand for the U.S. dollar but in the oppositedirection. The higher the exchange rate today, other things remaining the same, the larger is the expected profit from selling U.S. dollars today and holding foreign currencies, so the greater is the quantity of U.S. dollars supplied in the foreign exchange market.

## Supply Curve for U.S. Dollars

Figure 26.2 shows the supply curve of U.S. dollars in the foreign exchange market. A change in theexchange rate, other things remaining the same, brings a change in the quantity of U.S. dollars sup-plied and a movement along the supply curve. The

## Market Equilibrium

Equilibrium in the foreign exchange market depends on how the Federal Reserve and other central banks operate. Here, we will study equilibrium whencentral banks keep out of the foreign exchange market and examine the effects of alternative central bank actions later (on pp. 669-671).
Figure 26.3 shows the demand curve for U.S. dollars, $D$, from Fig. 26.1, the supply curve of U.S. dollars, S, from Fig. 26.2, and the equilibrium exchange rate. The exchange rate acts as a regulator of the quantities demanded and supplied. If the exchange rate is too high, there is a surplus of dollars. For ex- ample, in Fig. 26.3, if the exchange rate is 150 yen per U.S. dollar, there is a surplus of U.S. dollars. If the exchange rate is too low, there is a shortage of dollars. For example, if the exchange rate is 50 yen per U.S. dollar, there is a shortage of U.S. dollars.
At the equilibrium exchange rate, there is neither a shortage nor a surplus - the quantity supplied equals the quantity demanded. In Fig. 26.3, the equilibrium exchange rate is 100 yen per U.S. dollar. At this exchange rate, the quantity demanded and the quantity supplied are each $\$ 1.5$ trillion a day.
The foreign exchange market is constantly pulled to its equilibrium by foreign exchange traders who are constantly looking for the best price they can get. If they are selling, they want the highest price avail- able. If they are buying, they want the lowest price available. Information flows from trader to trader through a worldwide computer network, and the price adjusts minute by minute to keep the exchange rate at its equilibrium.
But as you've seen (in Economics inAction on p. 659), the U.S. dollar fluctuates a lot against other currencies. Changes in the demand for U.S. dollars or the supply of U.S. dollars bring these exchange rate fluctuations. We'll now look at the factors that make demand and supply change, starting with the demand side of the market.

Changes in the Demand for U.S. Dollars The demand for U.S. dollars in the foreign exchange market changes when there is a change in

- World demand for U.S. exports
- U.S. interest rate relative to the foreign interest rate
- The expected future exchange rate

World Demand for U.S. Exports An increase in world demand for U.S. exports increases the demand for U.S. dollars. To see this effect, think about Boeing's airplane sales. An increase in demand for air travel in Australia sends that country's airlines on a global shopping spree. They decide that the 787 is the ideal product, so they order 50 airplanes from Boeing. The demand for U.S. dollars now increases.

FIGURE 26.3 Equilibrium Exchange Rate


The demand curve for U.S. dollars is $D$, and the supply curve of U.S. dollars is $S$. If the exchange rate is 150 yen per U.S. dollar, there is a surplus of U.S. dollars and the exchange rate falls. If the exchange rate is 50 yen per U.S. dollar, there is a shortage of U.S. dollars and the exchange rate rises. If the exchange rate is 100 yen per U.S. dollar, there is neither a shortage nor a surplus of U.S. dollars and the exchange rate remains constant. The foreign exchange market is in equilibrium.
U.S. Interest Rate Relative to the Foreign Interest Rate People and businesses buy financial assets to make a return. The higher the interest rate that people can, make on U.S. assets compared with foreign assets, the• more U.S. assets they buy. What matters is not the level of the U.S. interest rate, but the U.S. interest rate relative to the foreign interest rate - the U.S. interest rate minus the foreign interest rate, which is called the U.S. interest rate differential. If the U.S. interest rate rises and the foreigninterest rate remains constant, the U.S. interest rate differential increases. The larger the U.S. interest ratedifferential, the greater is the demand for U.S. assets and the greater is the demand for U.S. dollars in theforeign exchange market.

The Expected Future Exchange Rate For a given current exchange rate, other things remaining the same, arise in the expected future exchange rate increases theprofit that people expect to make by holding U.S. dollars and the demand for U.S. dollars increases today.
Figure 26.4 summarizes the influences on the demand for U.S. dollars. An increase in the demand for U.S. exports, a rise in the U.S. interest rate differential, or a rise in the expected future exchange rate increases the demand for U.S. dollars today and shifts the demand curve rightward from $D_{0}$ to $D_{1}$. A decrease in the demand for U.S. exports a fall in
the U.S. interest rate differential, or a fall in the expected future exchange rate decreases the demand for U.S. dollars today and shifts the demand curve leftward from $D_{0}$ to $D_{2}$.

## Changes in the Supply of U.S. Dollars

The supply of U.S. dollars in the foreign exchange market changes when there is a change in

- U.S. demand for imports
- U.S. interest rate relative to the foreign interest rate
- The expected future exchange rate

FIGURE 26.4 Changes in the Demand for U.S. Dollars


A change in any influence on the quantity of U.S. dollars that people plan to buy, other than the exchange rate, brings a change in the demand for U.S. dollars.

## The demand for U.S. dollars

## Increases if:

- World demand for U.S. exports increases
- The U.S. interest rate differential rises
- The expected future exchange rate rises


## Decreases if:

- World demand for U.S. exports decreases
- The U.S. interest rate differential falls
- The expected future exchange rate falls

FIGURE 26.5 Changes in the Supply of U.S. Dollars


A change in any influence on the quantity of U.S. dollars that people plan to sell, other than the exchange rate, brings a change in the supply of dollars.

The supply of U.S. dollars

Increases if:

- U.S. import demand increases
- The U.S. interest rate differential falls
- The expected future exchange rate falls

Decreases if:

- U.S. import demand decreases
- The U.S. interest rate differential rises
- The expected future exchange rate rises
U.S. Demand for Imports An increase in the U.S. demand for imports increases the supply of U.S. dollars in the foreign exchange market. To see why, think about Wal-Mart's purchase of Blu-ray players. An increase in the demand for Blu-ray players sends Wal-Mart out on a global shopping spree. Wal-Mart decides that Panasonic Blu-ray players produced in Japan are the best buy, so Wal-Mart increases its purchases of these players. The supply of U.S. dollars now increases as Wal-Mart goes to the foreign exchange market for Japanese yen to pay Panasonic.
U.S. Interest Rate Relative to Foreign Interest Rate The effect of the U.S. interest rate differential on the supply of U.S. dollars is the opposite of its effect on the demand for U.S.
dollars. The larger the U.S. interest rate differential, the smaller is the supply of U.S. dollars in the foreign exchange market.
With a higher U.S. interest rate differential, people decide to keep more of their funds in U.S. dollarassets and less in foreign currency assets. They buy asmaller quantity of foreign currency and sell a smallerquantity of dollars in the foreign exchange market.
So, a rise in the U.S. interest rate, other things remaining the same, decreases the supply of U.S. dollars in the foreign exchange market.

The Expected Future Exchange Rate For agivencurrent exchange rate, other things remaining the same, afall in the expected future exchange rate decreases theprofit that can be made by holding U.S. dollars and decreases the quantity of U.S. dollars that people and businesses want to hold. To reduce their holdings of dollar assets, people and businesses must sell U.S. dollars. When they do so, the supply of U.S. dollars in the foreign exchange market increases. Figure 26.5 summarizes the influences on the supply of U.S. dollars. If the supply of U.S. dollars increases, the supply curve shifts rightward from $\mathrm{S}_{0}$ to $\mathrm{S}_{1}$. And if the supply of U.S. dollars decreases, the supply curve shifts leftward from $S_{0}$ to $S_{2}$.

## Changes in the Exchange Rate

The exchange rate changes when either the demandfor dollars or the supply of dollars changes.
If the demand for U.S. dollars increases and the supply does not change, the exchange rate rises. If the demand for U.S. dollars decreases and the supply does not change, the exchange rate falls.
Similarly, if the supply of U.S. dollars decreases and the demand does not change, the exchange rate rises. If the supply of U.S. dollars increases and the demand does not change, the exchange rate falls.
These predictions are exactly the same as thosefor any other market. Two episodes in the life of the U.S. dollar (next page) illustrate these predictions.
Two of the influences on demand and supply - the U.S. interest rate differential and the expected future exchange rate - change both sides of the foreign exchange market simultaneously. A rise in the U.S. interest rate differential or a rise in the expected future exchange rate increases demand, decreases supply, and raises the exchange rate. Similarly, a fall in the U.S. interest rate differential or a fall in the expected future exchange rate decreases demand, increases supply, and lowers the exchange rate.
We take a closer look at the interest rate differential and expectations in the next section.

## Arbitrage, Speculation, and Market Fundamentals

You've just seen how an exchange rate is determined. In our example, we used the U.S. dollar-Japanese yen exchange rate, but exchange rates between the U.S. dollar and all other currencies are determined in a similar way. So are the exchange rates among the other currencies such as that of the European euro and U.K. pound. Exchange rates are kept in alignment with each other by a process called arbitrage.

## ECONOMICS IN ACTION

## The Dollar on a Roller Coaster

The foreign exchange market is a striking example of a competitive market. The expectations of thousands of traders around the world influence this market minute-by-minute throughout the 24 -hour global trading day.

Demand and supply rarely stand still and their fluctuations bring a fluctuating exchange rate. Two episodes in the life of the dollar illustrate these fluctuations: 2007-2012, when the dollar depreciated and 2012-2014, when the dollar appreciated.

A Depreciating U.S. Dollar: 2007-2012 Between July 2007 and August 2012, the U.S. dollar depreciated against the yen. It fell from 120 yen to 77 yen per U.S. dollar. Part (a) of the figure provides a possible explanation for this depreciation:

In 2007, the demand and supply curves were those labeled $D_{07}$ and $S_{07}$. The exchange rate was 120 yen per U.S. dollar.

During the last quarter of 2007 and the first three quarters of 2008, the U.S. economy entered a severe credit crisis. The Federal Reserve cut the interest rate in the United States, but the Bank of Japan kept the interest rate unchanged in Japan. With a narrowing of the U.S. interest rate differential, funds flowed out of the United States. Also, currency traders expected the U.S.

(a) 2007-2012

The Falling and Rising U.S. Dollar
dollar to depreciate against the yen. The demand for U.S. dollars decreased and the supply of U.S. dollars increased.

In part (a) of the figure, the demand curve shifted leftward from $D_{07}$ to $D_{12}$, the supply curve shifted rightward from $S_{07}$ to $S_{12}$, and the exchange rate fell to 77 yen per U.S. dollar.

An Appreciating U.S. Dollar: 2012-2014 Between January 2012 and June 2014, the U.S. dollar appreciated against the yen. It rose from 77 yen to 102 yen per U.S. dollar. Part (b) of the figure provides an explanation for this appreciation. The demand and supply curves labeled $D_{12}$ and $S_{12}$ are the same as in part (a).

During 2013 and 2014, the Federal Reserve kept the U.S. interest rate low, but traders began to expect a future interest rate rise. Interest rates in Japan were even lower than in the United States, and the Bank of Japan, the central bank, embarked on a policy of expanding the Japanese money supply. With an expected future increase in U.S. interest rates and a lessened prospect of a rise in Japanese interest rates, the U.S. interest rate differential was expected to increase, so the dollar was expected to appreciate. The demand for U.S. dollars increased, and the supply of U.S. dollars decreased.

In the figure, the demand curve shifted rightward from $D_{12}$ to $D_{14}$, the supply curve shifted leftward from $S_{12}$ to $S_{14}$, and the exchange rate rose to 102 yen per U.S. dollar

(b) 2012-2014

Arbitrage is the practice of seeking to profit by buying in one market and selling for a higher price in another related market. Arbitrage in the foreign ex-change market and international loans markets and goods markets achieves threeoutcomes:

- The law of one price
- No round-trip profit
- Interest rate parity
- Purchasing power parity

The Law of One Price The law ofone price states that if an item is traded in more than one place, the pricewill be the same in all locations. An example of thislaw is that the exchange rate between the U.S. dollarand the U.K. pound is the same in New York as it is in London. You can see why arbitrage brings about this out- come by imagining that the exchange rate in London is 0.60 U.K. pounds per dollar and the price in New York is 0.61 U.K. pounds per dollar. In this imaginary situation, a trader who buys dollars in London and sells them in New York makes a profit of 0.01 U.K. pounds on every dollar traded. A trade of 1 mil- lion dollars brings a profit of 10,000 U.K. pounds.
Within a few seconds, the demand for U.K. pounds increases in London and the supply of U.K. pounds increases in New York. These changes in demand and supply raise the exchange rate in London and lower it in New York and make it equal in both marketsremoving the profit opportunity.

No Round-Trip Profit A round trip is using currency $A$ to buy currency $B$, and then using $B$ to buy $A$. A round trip might involve more stages, using $B$ to buy $C$ and then using $C$ to buy $A$. Arbitrage removes profit from all transactions of this type. Any fleeting profit is taken, and the changes in supply and demand induced by the momentarily available profit snap the exchange rates back to levels that remove the profit.

Interest Rate Parity Borrowers and lenders must choose the currency in which to denominate their as-sets and debts. Interest rate parity, which means equal rates of return across currencies, means that for risk-free transactions, there is no gain from choosing onecurrency over another.
To see why interest rate parity always prevails, sup- pose a Brazilian real bank deposit in Rio de Janeiroearns 10 percent a year and a U.S. dollar bank deposit in New York earns 1 percent a year. Why wouldn't people move their funds from New York to Rio?
The answer begins with the fact that to earn 10 percent in Rio, funds must be converted from U.S. dollar; to reals at the beginning of the year and from reals back to dollars at the end of the year. This transaction can be done without risk by selling reals for U.S. dollars today for delivery one year from today at an, exchange rate agreed today. Such a transaction is called" a future or forward transaction and it takes place at the: one-year forward exchange rate. Suppose that today's exchange rate is 2.30 re als per dollar, and you convert $\$ 100$ to 230 reals. In one year, you will have 253 reals - your deposit of 230 reals plus interest of 23 reals. If the one-year forward, exchange rate is 2.50 reals perU.S. dollar, you can contract today to sell 253 reals for $\$ 101$ for delivery one year. But that is exactly the amount you can earn by putting your $\$ 100$ in the New York bank and earning 1 percent a year. If for a few seconds, interest rate parity did not hold and it was possible to profit from buying and holding\}' Brazilian reals, traders would flock to the profit opportunity, supply dollars and demand reals, and drive the\}exchange rate to its interest rate parity level.

Purchasing Power Parity Suppose a camera costs 10,000 yen in Tokyo and $\$ 100$ in New York. If the exchange rate is 100 yen per dollar, the two monies have the same value. You can buy the camera in either Tokyo or New York for the same price. You can express that price as either 10,000 yen or $\$ 100$, but the price is the same in the two currencies.
The situation we've just described is called purchasing power parity (or PPP), which means equal value of money. PPP is an example of the law of one price, and if it does not prevail, arbitrage forces go to work. To see these forces, suppose that the price of the camera in New York is $\$ 120$, but in Tokyo it remains at 10,000 yen and the exchange rate remains at 100 yen per dollar. In this case, the camera in Tokyo still costs 10,000 yen or $\$ 100$, but in New York, it costs $\$ 120$ or 12,000 yen. Money buys more in Japan than in the United States. Money is not of equal value in the two countries.
Arbitrage now kicks in. With the camera cheaper in Tokyo than in New York, the demand for cam-eras increases in Tokyo and the supply of cameras increases in New York. The New York price falls and the Tokyo price rises to eliminate the price difference and restore purchasing power parity.
If most goods and services cost more in one country than another, the currency of the first country is said to be overvalued: a depreciation of the currency would restore PPP.
Similarly, the currency of the country with the lower prices is said to be undervalued: an appreciation of that currency would restore PPP. When goods and services cost the same in two countries, their currencies are said to be at their PPP levels.
Determining whether a currency is overvalued or undervalued based on PPP is not easy, and testing PPP by looking at individual prices requires care to ensure that the goods compared are identical. What is identical isn't always immediately obvious (see Economics in Action below).

## Speculation

Speculation is trading on the expectation of making a profit. Speculation contrasts with arbitrage, which is trading on the certainty of making a profit. Most foreign exchange transactions are based on speculation, which explains why the expected future exchange rate plays such a central role in the foreign exchange market.
The expected future exchange rate influences both supply and demand, so it influences the current equilibrium exchange rate. But what determines the expected future exchange rate?

The Expected Future Exchange Rate An expectation is a forecast. Exchange rate forecasts, like weatherforecasts, are made over horizons that run from a fewhours to many months and perhaps years. Also, like weather forecasters, exchange rate forecasters use scientific models and data to make their predictions.
But exchange rate forecasting differs from weather forecasting in three ways. First, exchange rate forecasts are hedged with a lot of uncertainty; second, there are many divergent forecasts; and third, the forecasts influence the outcome.
The dependence of today's exchange rate on forecasts of tomorrow's exchange rate can give rise to exchange rate volatility in the short run.

## ECONOMICS IN ACTION

## A Big Mac Index

Because a Big Mac is the same in Chicago as in Beijing, The Economist magazine wondered if its price in these cities might tell us how far China's yuan is from its PPP level. In July 2014, the price of a Big Mac was $\$ 4.80$ in America and 16.93 yuan or $\$ 2.73$ in China. Does this dollar price difference mean that the yuan is undervalued?

The Big Mac price comparison doesn't answer this question. A Big Mac looks the same in all places but most of its value is in its service, not its appearance.

The figure shows the price of a Big Mac as a percentage of the U.S. price averaged over 2000, 2007, and 2014. It shows that the price is persistently above the U.S. price in a few rich countries and persistently below the U.S. price in lower-income countries.

The persistent differences arise from different relative prices of services, not from over- or under-valued currencies.


The U.S. Doilar Price of a Big Mac in $\mathbf{1 0}$ Countries
Source of dala: The Economist, April 2000, June 2007, and July 2014.

Exchange Rate Volatility An exchange rate might rise one day and fall the next, as news about the influences on the exchange rate change the expected future exchange rate. For example, news that the Fed isgoing to start to raise U.S. interest rates next month brings an immediate increase in the demand for U.S. dollars, decrease in the supply of U.S. dollars, and appreciation of the U.S. dollar. As the news is digestedand its expected consequences revised, expectations are revised, sometimes upward and sometimes down- ward, bringing further changes in the exchange rate.
The influences of expectations and the constant arrival of news about the influences on supply and demand, make day-to-day and week-to-week changes in the exchange rate impossible to predict. But trends around which the exchange rate fluctuates are predictable and depend on market fundamentals.

## Market Fundamentals

The demand for U.S. dollars depends on world demand for U.S. exports, the supply of U.S. dollars depends on U.S. demand for imports, and both demand and supply depend on the U.S. interest rate differential. These are the market fundamentals that influence the exchange rate. But how they influence the exchange rate is different in the short run and the long run. The short-run influences are those described in the previous section of this chapter. To understand the long-run, we need to define and understand the role played by the real exchange rate.

The Real Exchange Rate The real exchange rate is the relative price of U.S.-produced goods and services toforeign-produced goods and services. It is a measureof the quantity of the real GDP of other countries that a unit of U.S. real GDP buys. For example, thereal Japanese yen exchange rate, $R E R$, is

$$
R E R=(E \times P) \div P
$$

where $E$ is the exchange rate (yen per U.S. dollar), $P$ is the U.S. price level, and $P^{*}$ is the Japanese price level. To understand the real exchange rate, suppose that the exchange rate $E$ is 100 yen per dollar. The United States produces only computer chips priced at $\$ 150$ each, so $P$ equals $\$ 150$ and $E X P$ equals 15,000 yen. Japan produces only iPods priced at 5,000 yen each, so $P^{*}$ equals 5,000 yen. Then the real Japanese yen exchange rate is

$$
R E R=(100 \times 150) \div 5,000=3 \text { iPods per chip }
$$

If Japan and the United States produced identical goods, the real exchange rate would equal 1 unit of U.S. real GDP per unit of Japanese real GDP.
In reality, U.S. real GDP is a different bundle of goods and services from Japanese real GDP. So the real exchange rate is not 1 and it changes over time. The forces of demand and supply in the markets for the millions of goods and services that make up real GDP determine the relative price of Japanese and U.S. real GDP and the real exchange rate.

Price Levels and Money We can turn the real exchange rate equation around and determine the ex- change rate as

$$
E=R E R \times P \div P .
$$

This equation says that the exchange rate equals the real exchange rate multiplied by the foreign price level, divided by the domestic price level.
In the long run, the quantity of money determines the price level. But the quantity theory of money applies to all countries, so the quantity of money in Japan determines the price level in Japan, and the quantity of money in the United States determines the price level in the United States.
For a given real exchange rate, a change in the quantity of money brings a change in the price leveland a change in the exchange rate.
The market fundamentals that determine the ex- change rate in the long run are the real exchange rate and the quantities of money in each economy.

## Exchange Rate policy

Because the exchange rate is the price of a country's money in terms of another country's money, governments and central banks must have a policy toward the exchange rate. Three possible exchange rate policies are

- Flexible exchange rate
- Fixed exchange rate
- Crawling peg


## Flexible Exchange Rate

A flexible exchange rate is an exchange rate that is determined by demand and supply in the foreign exchange market with no direct intervention by the central bank.
Most countries, including the United States, operate a flexible exchange rate, and the foreign exchange market that we have studied so far in this chapter is an example of a flexible exchange rate regime.

But even a flexible exchange rate is influenced by central bank actions. If the Fed raises the U.S. interest rate and other countries keep their interest rates unchanged, the demand for U.S. dollars increases, the supply of U.S. dollars decreases, and the exchange rate rises. (Similarly, if the Fed lowers the U.S. interest rate, the demand for U.S. dollars decreases, the supply increases, and the exchangerate falls.)
In a flexible exchange rate regime, when the central bank changes the interest rate, its purpose is not usually to influence the exchange rate, but to achieve some other monetary policy objective. (We return to this topic at length in Chapter 31.)

## Fixed Exchanged Rate

A fixed exchange rate is an exchange rate that is determined by a decision of the government or the central bank and is achieved by central bank intervention in the foreign exchange market to block the unregulated forces of demand and supply.
The world economy operated a fixed exchange rate regime from the end of World War II to the early1970s. China had a fixed exchange rate until recently. Hong Kong has had a fixed exchange rate for many years and continues with that policy today.
Active intervention in the foreign exchange market is required to achieve a fixed exchange rate.
If the Fed wanted to fix the U.S. dollar exchange rate against the Japanese yen, the Fed would have tosell U.S. dollars to prevent the exchange rate from rising above the target value and buy U.S. dollars to prevent the exchange rate from falling below the target value.
There is no limit to the quantity of U.S. dollars that the Fed can sell. The Fed creates U.S. dollars and can create any quantity it chooses. But there is a limit to the quantity of U.S. dollars the Fed can buy. That limit is set by U.S. official foreign currency reserves because to buy U.S. dollars the Fed must sell foreign currency. Intervention to buy U.S. dollars stops when U.S. official foreign currency reserves run out. Let's look at the foreign exchange interventions that the Fed can make.
Suppose the Fed wants the exchange rate to be steady at 100 yen per U.S. dollar. If the exchange raterises above 100 yen, the Fed sells dollars. If the exchange rate falls below 100 yen, the Fed buys dollars. By these actions, the Fed keeps the exchange rate close to its target rate of 100 yen per U.S. dollar.
Figure 26.6 shows the Fed's intervention in the foreign exchange market. The supply of dollars is $S$ and initially the demand for dollars is $D_{0}$. The equilibrium exchange rate is 100 yen per dollar. This exchange rate is also the Fed's target exchange rate, shown by the horizontal red line.
When the demand for U.S. dollars increases and the demand curve shifts rightward to $D_{1}$, the Fed sells $\$ 100$ billion. This action prevents the exchange rate from rising. When the demand for U.S. dollars de-creases and the demand curve shifts leftward to $D_{2}$, the Fed buys $\$ 100$ billion. This action prevents theexchange rate from falling.
If the demand for U.S. dollars fluctuates between $D_{1}$ and $D_{2}$ and on average is $D_{0}$, the Fed can repeatedly intervene in the way we've just seen. Sometimes the Fed buys and sometimes it sells but, on average, it neither buys nor sells.
But suppose the demand for U.S. dollars increases permanently from $D_{0}$ to $D_{1}$. To maintain the exchange rate at 100 yen per U.S. dollar, the Fed must sell dollars and buy foreign
currency, so U.S. official foreign currency reserves would be increasing. At some point, the Fed would abandon the exchange rate of 100 yen per U.S. dollar and stop piling up foreign currency reserves.
Now suppose the demand for U.S. dollars decreases permanently from Do to D2. In this situation, the Fed cannot maintain the exchange rate at 100 yen per U.S. dollar indefinitely. To hold the exchange rate at 100 yen, the Fed must buy U.S. dollars. When the Fed buys U.S. dollars in the foreign exchange market, it uses U.S. official foreign currency reserves. So the Fed's action decreases its foreign currencyreserves. Eventually, the Fed would run out of foreign currency and would then have to abandon the target exchange rate of 100 yen per U.S. dollar.

## Crawling Peg

A crawling peg is an exchange rate that follows a path determined by a decision of the government or the central bank and is achieved in a similar way to a fixed exchange rate by central bank intervention in the foreign exchange market. A crawling peg works like a fixed exchange rate except that the target value changes. The target might change at fixed intervals (daily, weekly, and monthly) or at random intervals.
The Fed has never operated a crawling peg, but some prominent countries do use this system. When China abandoned its fixed exchange rate, it replaced it with a crawling peg. Developing countries might use a crawling peg as a method of trying to control inflation - of keeping the inflation rate close to target.

The ideal crawling peg sets a target for the exchange rate equal to the equilibrium exchange rate The bottom line is that in the long run; exchange rate policy is monetary policy, not foreign trade pol-icy. To change its exports and imports, a country must change its comparative advantage (Chapter 2). on average. The peg seeks only to prevent large swings in the expected future exchange rate that change demand and supply and make the exchange rate fluctuate too wildly.
A crawling peg departs from the ideal if, as often happens with a fixed exchange rate, the target rate departs from the equilibrium exchange rate for too long. When this happens, the country either runs out of reserves or piles up reserves.
In the final part of this chapter, we explain how the balance of international payments is determined.

## FIGURE 26.6 Foreign Exchange

 Market Intervention

Initially, the demand for U.S. dollars is $D_{0}$, the supply of U.S. dollars is $S$, and the exchange rate is 100 yen per U.S. dollar. The Fed can intervene in the foreign exchange market to keep the exchange rate close to its target rate ( 100 yen in this example]. If the demand for U.S. dollars increases and the demand curve shifts from $D_{0}$ to $D_{1}$, the Fed sells dollars. If the demand for U.S. dollars decreases and the demand curve shifts from $D_{0}$ to $D_{2}$, the Fed buys dollars. Persistent intervention on one side of the market cannot be sustained.

## Financing International Trade

You now know how the exchange rate is determined, but what is the effect of the exchange rate? How does currency depreciation or currency appreciation influence our international trade and payments? We're going to lay the foundation for addressing these questions by looking at the scale of international trading, borrowing, and lending and at the way in which we keep our records of international transactions. These records are called the balance of payments accounts.

## ECONOMICS IN ACTION

## The People's Bank of China in the Foreign Exchange Market

You saw in the figure on p. 659 that the exchange rate between the U.S. dollar and the Chinese yuan was constant for several years. The reason for this constant exchange rate is that China's central bank, the People's Bank of China, intervened to operate a fixed exchange rate policy. From 1997 until 2005, the yuan was pegged at 8.28 yuan per U.S. dollar. Since 2005, the yuan
has appreciated slightly, but it has not been permitted to fluctuate freely. Since 2005, the yuan has been on a crawling peg Why Does China Manage Its Exchange Rate? The popular story is that China manages its exchange rate to keep its export prices low and to make it easier to compete in world markets. You've seen that this story is correct only in the short run. With prices in China unchanged, a lower yuan-U.S. dollar exchange rate brings lower U.S. dollar prices for China's exports. But the yuan-U.S. dollar exchange rate was fixed foralmost 10 years and has been managed for five moreyears. This long period of a fixed exchange rate has long-run, not short-run, effects. In the long run, the exchange rate has no effect on competitiveness. Thereason is that prices adjust to reflect the exchange rate and the real exchange rate is unaffected by the nominal exchange rate.

(a) Increase in U.S. dollar reserves

(b) Pegging the yuan

China's Foreign Exchange Market Intervention

So why does China fix its exchange rate? The most convincing answer is that China sees a fixed exchange rate as a way of controlling its inflation rate. By making the yuan crawl against the U.S. dollar, China's inflation rate is anchored to the U.S. inflation rate and will depart from U.S. inflation by an amount determined by the speed of the crawl.

## Balance of Payments Accounts

A country's balance of payments accounts records its international trading, borrowing, and lending in three accounts:

1. Current account
2. Capital and financial account
3. Official settlements account

The current account records receipts from exports of goods and services sold abroad, payments for importsof goods and services from abroad, net interest income paid abroad and net transfers abroad (such as foreign aid payments). The current account balance equals the sum of exports minus imports, net interest income, and net transfers.

The capital and financial account recordsforeign investment in the United States minus U.S. investment abroad. (This account also has a statistical discrepancy that arises from errors and omissions in measuring international capital transactions.)

The official settlements account records the change in U.S. official reserves, which are the government's holdings of foreign currency. If U.S. official reserves increase, the official settlements account balance is negative. The reason is that holding foreign money is like investing abroad. U.S. investment abroad is a minus item in the capital and financial account and in the official settlements account.
The sum of the balances on the three accounts always equals zero. That is, to pay for our current account deficit, we must either borrow more from abroad than we lend abroad or use our official reserves to cover the shortfall.
Table 26.1 shows the U.S. balance of payments accounts in 2013. Items in the current account and the capital and financial account that provide foreign currency to the United States have a plus sign; items that cost the United States foreign currency have a minus sign. The table shows that in 2013, U.S.imports exceeded U.S. exports and the current account had a deficit of $\$ 400$ billion. How do we pay for imports that exceed the value of our exports? That is, how do we pay for our current account deficit? We pay by borrowing from the rest of the world. The capital account tells us by how much. We borrowed $\$ 1,017$ billion (foreign investment in the United States) but made loans of $\$ 650$ billion (U.S. investment abroad). Our net foreign borrowing was $\$ 1,017$ billion minus $\$ 650$ billion, which equals $\$ 367$ billion. There is almost always a statistical discrepancy between capital and financial account and current account transactions, and in 2013, the discrepancy was $\$ 30$ billion. Combining the discrepancy with the measured net foreign borrowing gives a capital and financial account balance of $\$ 397$ billion.

The capital and financial account balance plus the current account balance equals the change in U.S. official reserves. In2013, the capital and financial accountbalance of $\$ 397$ billion plus the current account balance of - $\$ 400$ billion equaled - $\$ 3$ billion. Official reserves decreased in 2013 by $\$ 3$ billion. Holding less foreign reserves is like borrowing from the rest of the world, so this amount appears in the official settlements account in Table 26.1 as $+\$ 3$ billion. The sum of the balances on the three balance of paymentsaccounts equals zero.
To see more clearly what the nation's balance of payments accounts mean, think about your own balance of payments accounts. They are similar to the nation's accounts.

An Individual's Balance of Payments Accounts An individual's current account records the income from supplying the services of factors of production and the expenditure on goods and services.
Consider Jackie, for example. She worked in 2014 and earned an income of $\$ 25,000$. Jackie has $\$ 10,000$ worth of investments that earned her an interest income of \$1,000. Jackie's current account shows an income of $\$ 26,000$. Jackie spent $\$ 18,000$ buying consumption goods and services. She also bought a new house, which cost her \$60,000. So Jackie's total expenditure was $\$ 78,000$. Jackie's expenditure minus her income is $\$ 52,000(\$ 78,000$ minus $\$ 26,000$ ). This amount is Jackie's current account deficit.

To pay for expenditure of $\$ 52,000$ in excess of her income, Jackie must either use the money that she has in the bank or take out a loan. Suppose that Jackie took out a loan of $\$ 50,000$ to help buy her house and that this loan was the only borrowing that she did. Borrowing is an inflow in the capital account, soJackie's capital account surpluswas $\$ 50,000$. With a current account deficit of $\$ 52,000$ and a capital account surplus of $\$ 50,000$, Jackie was still $\$ 2,000$ short. She got that $\$ 2,000$ from her own bank account. Her cash holdings decreased by $\$ 2,000$.
Jackie's income from her work is like a country's income from its exports. Her income from her investments is like a country's interest income from foreigners. Her purchases of goods and services, including her purchase of a house, are like a country's imports. Jackie's loanborrowing from some-one else-is like a country's borrowing from the rest of the world. The change in Jackie's bank account is like the change in the country's official reserves.

## Borrowers and Lenders

A country that is borrowing more from the rest of the world than it is lending to the rest of the world is called a net borrower. Similarly, a net lender is a country that is lending more to the rest of the world than it is borrowing from the rest of the world.
The United States is a net borrower, but it has not always been in this situation.
Throughout the 1960s and most of the 1970s, the United States was a net lender to the rest of the world-the United States had a current account surplus and a capitalaccount deficit. But from the early 1980s, with the exception of only a single year, 1991, the United States has been a net borrower from the rest of the world. And during the years since 1992, the scale of U.S. borrowing hasmushroomed.
Most countries are net borrowers like the United States. But a few countries, including China, Japan, and oil-rich Saudi Arabia, are net lenders. In 2014, when the United States
borrowed more than $\$ 397$ billion from the rest of the world, China alone lent more than \$200 billion.
International borrowing and lending takes place in the global market for loanable funds. You studied the loanable funds market in Chapter 24, but there, we didn't take explicit account of the effects of the balance of payments and international borrowing and lending on the market. That's what we will now do.

TABLE 26.1 U.S. Balance of Payments Accounts in 2013

## Current account

Exports of goods and services $\quad+2,280$
Imports of goods and services
Net interest income +209
Net transfers -132
Current account balance $\quad \underline{\underline{-400}}$

## Capital and financial account

Foreign investment in the United States
U.S. investment abroad -650

| Statistical discrepancy | $\underline{+30}$ |
| :--- | ---: |
| Capital and financial account balance |  |
| $\underline{+397}$ |  |

## Official settlements account

Official settlements account balance 3

Source of data: Bureau of Economic Analysis.

## The Global Loanable Funds Market

Figure 26.7(a) illustrates the demand for loanable funds, $D L F_{W}$, and the supply of loanable funds, $S L F_{W}$, in the global loanable funds market. The world equilibrium real interest rate makes the quantity of funds supplied in the world as a whole equal to the quantity demanded. In this example, the equilibrium real interest rate is 5 percent a year and the quantity of funds is $\$ 10$ trillion.

## ECONOMICS IN ACTION

## Three Decades of Deficits

The numbers that you reviewed in Table 26.1 give a snapshot of the balance of payments accounts in 2013. The figure below puts that snapshot into perspective by showing the balance of payments between 1980 and the first half of 2014.

Because the economy grows and the price level rises, changes in the dollar value of the balance of payments do not convey much information. To remove the influences of economic growth and inflation, the figure
shows the balance of payments expressed as a percentage of nominal GDP.

As you can see, a large current account deficit emerged during the 1980s but declined from 1987 to 1991. The current account deficit then increased through 2006, decreased again through 2009, and then remained steady.

The capital and financial account balance is almost a mirror image of the current account balance. The official settlements balance is very small in comparison with the balances on the other two accounts.


The U.S. Balance of Payments

An International Borrower Figure 26.7(b) shows the loanable funds market in a country that borrows from the rest of the world. The country's demand for loanable funds, $D L F$, is part of the world demand inFig. 26.7(a). The country's supply of loanable funds, SLFD, is part of the world supply.
If this country were isolated from the global market, the real interest rate would be 6 percent a year (where the DLF and SLFD curves intersect). But if the country is integrated into the global economy, with an interest rate of 6 percent a year, funds would flood into it. With a real interest rate of 5 percent a year in the global market, suppliers of loanable funds would seek the higher return in this country. In effect, the country faces the supply of loanable funds curve, $S L F$, which is horizontal at the world equilibrium real interest rate. The country's demand for loanable funds and the world interest rate determine the equilibrium quantity of loanable funds-\$2.5 billion in Fig. 26.7(b).

An International Lender Figure 26.7(c) shows the situation in a country that lends to the rest of the world. As before, the country's demand for loanable funds, $D L F$, is part of the world demand and the country's supply of loanable funds, SLFD, is part of the world supply in Fig. 26.7(a).

If this country were isolated from the global market, the real interest rate would be 4 percent a year (where the $D L F$ and $S L F_{D}$ curves intersect). But if this country is integrated into the global economy, with an interest rate of 4 percent a year, funds would quickly flow out of it. With a real interest rate of 5 percent a year in the rest of the world, domestic suppliers of loanable funds would seek the higher return in other countries. Again, the country faces the supply of loanable funds curve, SLF, which is horizontal at the world equilibrium real interest rate.
The country's demand for loanable funds and the world interest rate determine the equilibrium quantity of loanable funds-\$1.5 billion in Fig. 26.7(c).

FIGURE 26.7 Borrowing and Lending in the Global Loanable Funds Market


In the global loanable funds market in part (a), the demand for loanable funds curve, $D L F_{W}$, and the supply of funds curve, SLF ${ }_{W}$, determine the world real interest rate. Each country can get funds at the world real interest rate and faces the (horizontal) supply curve SLF in parts (b) and (c).

At the world real interest rate, borrowers in part (b)
want more funds than the quantity supplied by domestic lenders, $\$ 1.5$ million on the domestic supply curve $S L F_{D}$. The shortage is made up by net foreign borrowing. Domestic suppliers of funds in part (c) want to lend more than domestic borrowers demand. The excess quantity supplied goes to foreign borrowers.

## Debtors and Creditors

A net borrower might be decreasing its net assets held in the rest of the world, or it might be going deeper into debt. A nation's total stock of foreign investment determines whether it is a debtor or a creditor. A debtor nation is a country that during its entire history has borrowed more from the rest of the world than it has lent to it. It has a stock of outstanding debt to the rest of the world that exceeds the stock of its own claims on the rest of the world. A creditor nation is a country that during its entire history has invested more in the rest of the world than other countries have invested in it.
The United States was a debtor nation through the nineteenth century as we borrowed from Europe to finance our westward expansion, railroads, and industrialization. We paid off our debt and became a creditor nation for most of the twentieth century. But following a string of current account deficits, we became a debtor nation again in 1986.
Since 1986, the total stock of U.S. borrowing from the rest of the world has exceeded U.S. lending to the rest of the world. The largest debtor nations are the capital-hungry developing countries (such as the United States was during the nineteenth century).

The international debt of these countries grew fromless than a third to more than a half of their gross domestic product during the 1980s and created what was called the "Third World debt crisis."
Should we be concerned that the United States is a net borrower and a debtor? The answer depends on whether the borrowing is financing investment that in turn is generating economic growth and higher income, or financing consumption expenditure. If the borrowed money is used to finance consumption, it will eventually have to be reduced, and the longer it goes on, the greater is the reduction in consumption that will eventually be necessary.

## Is U.S. Borrowing for Consumption?

In 2014, the United States borrowed $\$ 397$ billionfrom abroad. In that year, private investment in buildings, plant, and equipment was $\$ 2,829$ billion and government investment in defense equipment and social projects was $\$ 588$ billion. All this investment added to the nation's capital and increased productivity. Government also spends on education and healthcare services, which increase human capital. U.S. international borrowing is financing private and public investment, not consumption.

## Current Account Balance

What determines a country's current account balanceand net foreign borrowing? You've seen that net exports ( $N X$ ) is the main item in the current account. We can define the current account balance (CAB) as

$$
C A B=N X+\text { Net interest income }+ \text { Net transfers. }
$$

We can study the current account balance by looking at what determines net exports because the other two items are small and do not fluctuate much.

## Net Exports

Net exports are determined by the government budget and private saving and investment. To see how net exports are determined, we need to recall some of thethings that we learned in Chapter 24 about the flows of funds that finance investment. Table 26.2 refreshesyour memory and summarizes some calculations.
Part (a) lists the national income variables that are needed, with their symbols. Part (b) defines three balances: net exports, the government sector balance, and the private sector balance.
Net exports is exports of goods and services minus imports of goods and services.
The government sector balance is equal to net taxes minus government expenditure on goods and services. If that number is positive, a government sector surplus is lent to other sectors; if that number is negative, a government deficit must be financed by borrowing from other sectors. The government sector deficit is the sum of the deficits of the federal, state, and local governments.
The private sector balance is saving minus invest-ment. If saving exceeds investment, a private sector surplus is lent to other sectors. If investment exceeds saving, a private sector deficit is financed by borrowing from other sectors.

Part (b) also shows the values of these balances for the United States in 2014. As you can see, net exports were - $\$ 564$ billion, a deficit of $\$ 564$ billion. The government sector's revenue from net taxes was $\$ 2,362$ billion and its expenditure was $\$ 3,162$ billion, so the government sector balance was $-\$ 800$ billion-a deficit of $\$ 800$ billion. The private sector saved $\$ 3,065$ billion and invested $\$ 2,829$ billion, so its balance was $\$ 236$ billion - a surplus of $\$ 236$ billion. Part (c) shows the relationship among the three balances. From the National Income and Product Accounts, we know that real GDP, $Y$, is the sum of consumption expenditure ( $C$ ), investment, government expenditure, and net exports. Real GDP also equals the sum of consumption expenditure, saving, and net taxes. Rearranging these equations tells us that net exports is the sum of the government sector balance and the private sector balance. In the United States in 2014, the government sector balance was $-\$ 800$ billion and the private sector balance was $\$ 236$ billion. The government sector balance plus the private sector balance equaled net exports of $-\$ 564$ billion.

## Where is the Exchange Rate?

We haven't mentioned the exchange rate while dis- cussing the balance of payments. Doesn't it play a role? The answer is that in the short run it does but in the long run it doesn't. In the short run, a fall in the dollar lowers the real exchange rate, which makes U.S. imports more costly and U.S. exports more competitive. A higher price of imported consumption goods and services might induce a decrease in consumption expenditure and an increase in saving. A higher price of imported capital goods might induce a decrease in investment. Other things remaining the same, an increase in saving or a decrease in investment decreases the private sector deficit and decreases the current account deficit. But in the long run, a change in the nominal ex- change rate leaves the real exchange rate unchanged and plays no role in influencing the current account balance.

## ECONOMICS IN ACTION

## The Three Sector Balances

You've seen that net exports equal the sum of the government sector balance and the private sector balance. How do these three sector balances fluctuate over time?
The figure answers this question. It shows the government sector balance (the red line), net exports (the blue line), and the private sector balance (the green line).

The private sector balance and the government sector balance move in opposite directions. When the government sector deficit increased during the late 1980s and early 1990s, the private sector surplus increased. And when the government sector deficit decreased and became a surplus during the late 1990s and early 2000 s, the private sector's surplus decreased and became a deficit. And when the government deficit increased yet again from 2007 to 2009, the private sector deficit shrank and became a surplus.
Sometimes, when the government sector deficit increases, as it did during the first half of the 1980s, net exports become more negative. But after the early 1990s, net exports did not follow the government sector balance closely. Rather, net exports respond to the sum of the government sector and private sector
balances. When both the private sector and the government sector have a deficit, net exports are negative and the combined private and government deficit is financed by borrowing from the rest of the world. But the dominant trend in net exports is negative.


The Three Sector Balances
Source of data: Bureau of Economic Analysis.

TABLE 26.2 Net Exports, the Government Budget, Saving, and Investment

| Symbols and | United States <br> in 2014 <br> equations <br> (billions of dollars) |
| :---: | :---: |

(a) Variables

| Exports* | X | 2,335 |
| :--- | :---: | :---: |
| Imports $^{*}$ | $M$ | 2,899 |
| Government expenditure | $G$ | 3,162 |
| Net taxes | $T$ | 2,362 |
| Investment | 1 | 2,829 |
| Saving | $S$ | 3,065 |

(b) Balances

| Net exports | $X-M$ | $2,335-2,899=-564$ |
| :--- | :---: | :--- |
| Government sector | $T-G$ | $2,362-3,162=-800$ |
| Private sector | $S-1$ | $3,065-2,829=+236$ |

## (c) Relationship among balances

| National accounts | , $Y=C+1+G+X-M$ |  |
| :---: | :---: | :---: |
|  | $=C+S$ |  |
| Rearranging: $X$ | $X-M=S-1+T-G$ |  |
| Net exports | $X-M$ | -564 |
| equals: |  |  |
| Government sector | or $\quad T-G$ | -800 |
| plus |  |  |
| Private sector | S-1 | +236 |

Source of data: Bureau of Economic Analysis. The data are for 2014, second quarter, seasonally adjusted at annual rate.
*The National Income and Product Accounts measures of exports and imports are slightly different from the balance of payments accounts measures in Table 26.1 on p. 634.

## ECONOMIC ANAIYSIS

- The news article says the dollar rose against "a basket" of other currencies when the Fed published the minutes of the July meeting of the FOMC on August 20, 2014.
- The main currencies in the "basket" are the Japanese yen, the European euro, and the U.K. pound.
- Figure 1 shows how the dollar exchange rate changed against these three currencies from July 2 to August 27, 2014.
- You can see that the dollar increased against all three currencies. It increased most against the European euro and the U.K. pound and least against the Japanese yen.
- The Fed's July meeting minutes indicated that the Fed was moving closer to being ready to raise interest rates, and the news article aitributes the stronger dollar to this news.
- But as you can see in Fig. 1, although the foreign exchange value of the dollar did rise immediately after the July minutes were released, it had been rising for almost two months.
- Also, after August 21, the dollar rose only against the euro. Against the yen and the pound, the dollar remained approximately constant.
- These facts about the timing of changes in the exchange rate and the different behavior of the dollar against the euro from the other two currencies suggests that other forces are at work.
- The summer of 2014 was a time of global tension arising from the poliltical situation in Ukraine. Currency traders sold the Russian ruble and the Ukrainian hryvinia and bought the U.S. dollar rather than the euro or pound.
- The Russia-Ukraine situation and the added effect of the expectation of a U.S. interest rate rise changed the demand for and supply of U.S. dollars in the foreign exchange market.
- The political tensions and predicted future rise in the U.S. interest rate increased the expected future exchange rate.
- With a higher expected future exchange rate, the demand for dollars increases and the supply of dollars decreases, and these changes in demand and supply bring an immediate appreciation of the dollar.
- Figure 2 shows these changes in supply and demand and their effects on the U.S. dollar-euro exchange rate.
- On July 2, demand was $D_{0}$ and supply was $S_{0}$. The equilibrium exchange rate was 0.732 euros per dollar. (The equilibrium quantity of dollars traded is an assumption.)


Figure 1 The U.S. Dollar Exchange Rate in
July and August 2014


Figure 2 The U.S. Dollar Foreign Exchange Market

- On August 22, demand had increased to $D_{1}$ and supply had decreased to $S_{1}$. The equilibrium exchange rate had risen to 0.755 euros per dollar-an appreciation of the dollar (and a depreciation of the euro).
- Although we can explain past changes in the exchange rate, we can't predict the future exchange rate because we can't predict the course of the factors that influence it.


## PART NINE: MACROECONOMIC FLUCTUATIONS

## CHAPTER 27: Aggregate Supply and Aggregate demand

After studying this chapter, you will be able to:

- Explain what determines aggregate supply in the long run and in the short run
- Explain what determines aggregate demand
- Explain how real GDP and the price level aredetermined and what causes growth, inflation, and cycles
- Describe the main schools of thought in macroeconomics today

Real GDP grew by 4.2 percent in the second quarteof 2014 after shrinking in the first quarter. The inflation rate also edged upward in 2014. Why do realGDP and inflation fluctuate? This chapter explains the economic fluctuations that we call the business cycle. You will study the aggregate supply-aggregate demand model or AS-AD model - a model of real GDP and the pricelevel. And in Economics in the News at the end of the chapter, you will use that model to interpret and explain the state of the U.S. economy in 2014.

## Aggregate Supply

The purpose of the aggregate supply-aggregate demand model that you study in this chapter is to explain how real GDP and the price level are deter-mined and how they interact. The model uses similarideas to those that you encountered in Chapter 3 when you learned how the quantity and price in a competitive market are determined. But the aggregate supplyaggregate demand model (AS-AD model) isn't just an application of the competitive market model. Some differences arise because the $A S-A D$ model is amodel of an imaginary market for the total of all the final goods and services that make up real GDP. Thequantity in this "market" is real GDP and the price isthe price level measured by the GDP deflator. One thing that the $A S-A D$ model shares with the competitive market model is that both distinguish between supplyand the quantitysupplied. Webegin by explaining what we mean by the quantity of real GDP supplied.

## Quantity Supplied and Supply

The quantity of real GDPsupplied is the total quantity of goods and services, valued in constant base- year (2009) dollars, that firms plan to produce during a given period. This quantity depends on the quantity of labor employed, the quantity of physical andhuman capital, and the state of technology.
At any given time, the quantity of capital and the state of technology are fixed. They depend on decisions that were made in the past. The population is also fixed. But the quantity of labor is not fixed. It depends on decisions made by households and firms about the supply of and demand for labor.
The labor market can be in any one of three states: at full employment, above full employment, or belowfull employment. At full employment, the quantity ofreal GDP supplied
is potential GDP, which depends on the full-employment quantity of labor (see Chapter23, pp. 546-548). Over the business cycle, employment fluctuates around full employment and the quantity of real GDP supplied fluctuates around potential GDP.
Aggregate supply is the relationship between the quantity of real GDP supplied and the price level. This relationship is different in the long run than in the short run and to study aggregate supply, we distinguish between two time frames:

- Long-run aggregate supply
- Short-run aggregate supply

Long-Run Aggregate Supply is the relationship between the quantity of real GDP supplied and the price level when the money wage rate changes in step with the price level to maintain full employment. The quantity of real GDP supplied at full employment equals . potential GDP and this quantity is the same regard- less of the price level.
The long-run aggregate supply curve in Fig. 27.1 illustrates long-run aggregate supply as the vertical line at potential GDP labeled LAS. Along the long- run aggregate supply curve, as the price level changes, the money wage rate also changes so the real wage rate remains at the full-employment equilibrium level and real GDP remains at potential GDP. The long-run aggregate supply curve is always vertical and is always located at potential GDP.
The long-run aggregate supply curve is vertical because potential GDP is independent of the pricelevel. The reason for this independence is that a movement along the LAS curve is accompanied by a change, in two sets of prices: the prices of goods and services-\} the price level-and the prices of the factors of production, most notably, the money wage rate. A 10 percent $\}$ increase inthe prices of goods and services is matched by a 10 percent increase in the money wage rate. Because the price level and the money wage ratechange' by the same percentage, the real wage rate remains unchanged at its full-employment equilibrium level. So when the price level changes and the real wage rate, remains constant, employment remains constant andreal GDP remains constant at potential GDP.

Production at a Pepsi Plant You can see more clearly why real GDP is unchanged when all prices change by the same percentage by thinking aboutproduction decisions at a Pepsi bottling plant. How does the quantity of Pepsi supplied change if the price of Pepsi changes and the wage rate of the workers and prices of all the other resources used vary by the same percentage? The answer is that the.: quantity supplied doesn't change. The firm produces the quantity that maximizes profit. That quantity depends on the price of Pepsi relative to the cost of producing it. With no change in price relative to cost, production doesn't change.

## Short-Run Aggregate Supply

Short-run aggregate supply is the relationship between the quantity of real GDP supplied and the price level when the money wage rate, the prices of other resources, and potential GDP remain constant. Figure 27.1 illustrates this relationship as the short-run aggregate sup-
ply curve SAS and the short-run aggregate supply schedule. Each point on the SAS curve corresponds to a row of the short-run aggregate supply schedule. For example, point $A$ on the SAS curve and row $A$ of the schedule tell us that if the price level is 100 , the quantity of real GDP supplied is $\$ 15$ trillion. In the short run, a rise in the price level brings an increase in the quantity of real GDP supplied. The short-run aggregate supply curve slopes upward. With a given money wage rate, there is one price level at which the real wage rate is at its fullemployment equilibrium level. At this price level, the quantity of real GDP supplied equals potential GDP and the SAS curve intersects the LAS curve. In this example, that price level is 110. If the price level rises above 110 , the quantity of real GDP supplied increases along the SAS curve and exceeds potential GDP; if the price level falls below 110, the quantity of real GDP sup- plied decreases along the SAS curve and is less than potential GDP.

Back at the Pepsi Plant You can see why the short- run aggregate supply curve slopes upward by returning to the Pepsi bottling plant. If production increases, marginal cost rises and if production decreases, marginal cost falls (see Chapter 2, p. 73). If the price of Pepsi rises with no change in the money wage rate and other costs, Pepsi can increase profit by increasing production. Pepsi is in businessto maximize its profit, so it increases production. Similarly, if the price of Pepsi falls while the money wage rate and other costs remain constant, Pepsi can avoid a loss by decreasing production. Thelower price weakens the incentive to produce, so Pepsi decreases production.
What's true for Pepsi bottlers is true for the producers of all goods and services. When all prices rise, theprice level rises. If the price level rises and the money wage rate and other factor prices remain constant, all firms increase production and the quantity of realGDP supplied increases. A fall in the price level hasthe opposite effect and decreases the quantity of real GDP supplied.

## Changes in Aggregate Supply

A change in the price level changes the quantity of real GDP supplied, which is illustrated by a movement along the short-run aggregate supply curve. It does not change aggregate supply. Aggregate supply changes when an influence on production plans other than the price level changes. These other influences include changes in potential GDP and changes in the money wage rate. Let's begin by looking at a change in potential GDP. Changes in Potential GDP When potential GDP changes, aggregate supply changes. An increase in potential GDP increases both long-run aggregate supply and short-run aggregate supply. Figure 27.2 shows the effects of an increase in potential GDP. Initially, the long-run aggregate supply curve is $L A S_{0}$ and the short-run aggregate supply curve is $S A S_{0}$. If potential GDP increases to $\$ 17$ trillion, long-run aggregate supply increases and the long-run aggregate supply curve shifts rightward to $L A S_{1}$. Short- run aggregate supply also increases, and the shortrun aggregate supply curve shifts rightward to SAS 1. Thetwo supply curves shift by the same amount only if thefull-employment price level remains constant, which we will assume to be the case.

FIGURE 27.1 Long-Run and Short-Run Aggregate Supply


In the long run, the quantity of real GDP supplied is potential GDP and the LAS curve is vertical at potential GDP. In the short run, the quantity of real GDP supplied increases if the price level rises, while all other influences on supply plans remain the same.

The short-run aggregate supply curve, SAS, slopes upward. The short-run aggregate supply curve is based on the aggregate supply schedule in the table. Each point $A$ through $E$ on the curve corresponds to the row in the table identified by the same letter.

When the price level is 110 , the quantity of real GDP supplied is $\$ 16$ trillion, which is potential GDP. If the price level rises above 110 , the quantity of real GDP supplied increases and exceeds potential GDP; if the price level falls below 110 , the quantity of real GDP supplied decreases below potential GDP.

Potential GDP can increase for any of three reasons:

- An increase in the full-employment quantity of labor
- An increase in the quantity of capital
- An advance in technology

Let's look at these influences on potential GDP and the aggregate supply curves.
An Increase in the Full-Employment Quantity of Labor A Pepsi bottling plant that employs 100 workers bottles more Pepsi than does an otherwise identical plant that employs 10 workers. The same is true for the economy as a whole. The larger the quantity of laboremployed, the greater is real GDP.
Over time, potential GDP increases because the labor force increases. But (with constant capital and technology) potential GDP increases only if the full- employment quantity of labor increases. Fluctuations in employment over the business cycle bring fluctuations in real GDP. But these changes in real GDP are fluctuations around potential GDP. They are not changes in potential GDP and long-run aggregate supply.

FIGURE 27.2 A Change in Potential GDP


An increase in potential GDP increases both long-run aggregate supply and short-run aggregate supply. The long-run aggregate supply curve shifts rightward from $L A S_{0}$ to $L A S_{1}$ and the short-run aggregate supply curve shifts from $S A S_{0}$ to SAS 1 .
An Increase in the Quantity of Capital A Pepsi bottling plant with two production lines bottles more Pepsi than does an otherwise identical plant that has only one pro-duction line. For the economy, the larger the quantity of capital, the more productive is the labor force and the greater is its potential GDP. Potential GDP per person in the capital-rich United States is vastly greater than that in capital-poor China or Russia.

Capital includes human capital. One Pepsi plant is managed by an economics major with an MBA and has a labor force with an average of 10 years of experience. This plant produces a larger output than does an otherwise identical plant that is managed by someone with no business training or experience and that has a young labor force that is new to bottling. The first plant has a greater amount of human capital than the second. For the economy as a whole, the larger the quantity of human capital-the skills that people have acquired in school and through on-the- job training-the greater is potential GDP.

An Advance in Technology A Pepsi plant that has pre- computer age machines produces less than one that uses the latest robot technology. Technological change enables firms to produce more from any given amount of factors of production. So even with fixed quantities of labor and capital, improvements in technology increase potential GDP Technological advances are by far the most important source of increased production over the past two centuries. As a result of technological advances, one farmer in the United States today can feed 100 people and in a year one autoworker can produce almost 14 cars and trucks.
Let's now look at the effects of changes in the money wage rate.
Changes in the Money Wage Rate When the money wage rate (or the money price of any other factor of production such as oil) changes, short-run aggregatesupply changes but long-run aggregate supply does not change.
Figure 27.3 shows the effect of an increase in the money wage rate. Initially, the short-run aggregate supply curve is SASo, A rise in the money wage ratedecreases short-run aggregate supply and shifts the short-run aggregate supply curve leftward to SAS2, A rise in the money wage rate decreases short-run aggregate supply because it increases firms' costs.
With increased costs, the quantity that firms are willing to supply at each price level decreases, which isshown by a leftward shift of the SAS curve.
A change in the money wage rate does not change long-run aggregate supply because on the LAS curve, the change in the money wage rate is accompanied byan equal percentage change in the price level. With nochange in relative prices, firms have no incentive to change production and real GDP remains constant at potential GDP. With no change in potential GDP, thelong-run aggregate supply curve LAS does not shift.

What Makes the Money Wage Rate Change? The money wage rate can change for two reasons: departures from full employment and expectations about inflation. Unemployment above the natural rate putsdownward pressure on the money wage rate, and unemployment below the natural rate puts upward pressure on it. An expected rise in the inflation ratemakes the money wage rate rise faster, and an expected fall in the inflation rate slows the rate at which the money wage rate rises.

## Aggregate Demand

The quantity of real GDP demanded $(Y)$ is the sum of real consumption expenditure (C), investment ( $I$ ), government expenditure ( $G$ ), and exports $(X)$ minus imports ( $M$ ). That is,

$$
Y=C+I+G+X-M .
$$

FIGURE 27.3 A Change in the Money Wage Rate


A rise in the money wage rate decreases short-run aggregate supply and shifts the short-run aggregate supply curve leftward from $S A S_{0}$ to $S A S_{2}$. A rise in the money wage rate does not change potential GDP, so the long-run aggregate supply curve does not shift.

## The Aggregate Demand

The quantity of real GDP demanded is the total amount of final goods and services produced in the United States that people, businesses, governments, and foreigners plan to buy.
These buying plans depend on many factors.
Some of the main ones are

1. The price level
2. Expectations
3. Fiscal policy and monetary policy
4. The world economy

We first focus on the relationship between the quantity of real GDP demanded and the price level. To study this relationship, we keep all other influences on buying plans the same and ask: How does the quantity of real GDP demanded vary as the price level varies?

## The Aggregate Demand Curve

Other things remaining the same, the higher the price level, the smaller is the quantity of real GDP demanded. This relationship between the quantity ofreal GDP demanded and the
price level is called aggregate demand. Aggregate demand is described byan aggregate demand schedule and an aggregate demand curve.
Figure 27.4 shows an aggregate demand curve (AD) and an aggregate demand schedule. Each point on the $A D$ curve corresponds to a row of the schedule. For example, point $\mathrm{C}^{\prime}$ on the $A D$ curve and row $C^{\prime}$ of the schedule tell us that if the price level is 110 , the quantity of real GDP demanded is $\$ 16$ trillion.
The aggregate demand curve slopes downward for two reasons:

- Wealth effect
- Substitution effects

Wealth Effect When the price level rises but other things remain the same, real wealth decreases. Real wealth is the amount of money in the bank, bonds,stocks, and other assets that people own, measured not in dollars but in terms of the goods and services that the money, bonds, and stocks will buy.
People save and hold money, bonds, and stocks for many reasons. One reason is to build up funds for education expenses. Another reason is to build up enough funds to meet possible medical expenses or other big bills. But the biggest reason is to build up enough funds to provide a retirement income.
If the price level rises, real wealth decreases. People then try to restore their wealth. To do so, they must increase saving and, equivalently, decrease current consumption. Such a decrease in consumption is a decrease in aggregate demand.

Maria's Wealth Effect You can see how the wealth effect works by thinking about Maria's buying plans. Maria lives in Moscow, Russia. She has worked hardall summer and saved 20,000 rubles (the ruble is thecurrency of Russia), which she plans to spend attending graduate school when she has finished her economics degree. So Maria's wealth is 20,000 rubles.
Maria has a part-time job, and her income from this job pays her current expenses. The price level in Russia rises by 100 percent, and now Maria needs 40,000 rubles to buy what 20,000 once bought. Totry to make up some of the fall in value of her savings, Maria saves even more and cuts her current spending to the bare minimum.

Substitution Effects When the price level rises and other things remain the same, interest rates rise. Thereason is related to the wealth effect that you've juststudied. A rise in the price level decreases the realvalue of the money in people's pockets and bank accounts. With a smaller amount of real money around, banks and other lenders can get a higher interest rate on loans. But faced with a higher interest rate, people and businesses delay plans to buy new capital and consumer durable goods and cut back on spending.
This substitution effect involves changing the timing of purchases of capital and consumer durable goods and is called an intertemporal substitution effect - a substitution across time. Saving increases to increase future consumption. To see this intertemporal substitution effect more clearly, think about your own plan to buy a new computer. At an interest rate of 5 percent a year, you might borrow $\$ 1,000$ and buy the new computer.

FIGURE 27.4 Aggregate Demand


|  | Price level <br> (GDP deflator) | Real GDP demanded <br> (trillions of 2009 dollars) |
| :--- | :---: | :---: |
| $A^{\prime}$ | 90 | 17.0 |
| $B^{\prime}$ | 100 | 16.5 |
| $C^{\prime}$ | 110 | $\mathbf{1 6 . 0}$ |
| $D^{\prime}$ | 120 | 15.5 |
| $E^{\prime}$ | 130 | 15.0 |

The aggregate demand curve $(A D)$ shows the relationship between the quantity of real GDP demanded and the price level. The aggregate demand curve is based on the aggregate demand schedule in the table. Each point $A^{\prime}$ through $E^{\prime}$ on the curve corresponds to the row in the table identified by the same letter. When the price level is 110 , the quantity of real GDP demanded is $\$ 16$ trillion, as shown by point $C^{\prime}$ in the figure. A change in the price level, when all other influences on aggregate buying plans remain the same, brings a change in the quantity of real GDP demanded and a movement along the $A D$ curve.

But at an interest rate of 10 percent a year, you might decide that the payments would be too high. You don't abandon your plan to buy the computer, butyou decide to delay your purchase. A second substitution effect works through inter- national prices. When the U.S. price level rises and other things remain the same, U.S.-made goods and services become
more expensive relative to foreign- made goods and services. This change in relative prices encourages people to spend less on U.S.-made items and more on foreign-made items. For example, if the U.S. price level rises relative to the Japanese price level, Japanese buy fewer U.S.-made cars (U.S. exports decrease) and Americans buy more Japanese- made cars (U.S. imports increase). U.S. GDP decreases.

Maria'sSubstitution Effects In Moscow, Russia, Maria makes some substitutions. She was planning to trade in her old motor scooter and get a new one. But with a higher price level and a higher interest rate, she decides to make her old scooter last one more year. So, with the prices of Russian goods sharply increasing, Maria substitutes a low-cost dress made in Malaysia for the Russian-made dress she had originally planned to buy.

Changes inthe Quantity of RealGDP Demanded When the price level rises and other things remain thesame, the quantity of real GDP demanded decreases - a movement up along the $A D$ curve as shown by the arrow in Fig. 27.4. When the price level falls and otherthings remain the same, the quantity of real GDP demanded increases - a movement down along the $A D$ curve.
We've now seen how the quantity of real GDP demanded changes when the price level changes. How do other influences on buying plans affect aggregate demand?

## Changes in Aggregate Demand

A change in any factor that influences buying plans other than the price level brings a change in aggregate demand. The main factors are

- Expectations
- Fiscal policy and monetary policy
- The world economy

Expectations An increase in expected future income increases the amount of consumption goods (especially big-ticket items such as cars) that people plan to buy today and increases aggregate demand.
An increase in the expected future inflation rate increases aggregate demand today because people decide to buy more goods and services at today's relatively lower prices. An increase in expected future profits increases the investment that firms plan to undertake today and increases aggregate demand.

Fiscal Policy and Monetary Policy The government's attempt to influence the economy by setting and changing taxes, making transfer payments, and purchasing goods and services is called fiscal policy. A tax cut or an increase in transfer payments-for example, unemployment benefits or welfare payments- increases aggregate demand. Both of these influences operate by increasing households' disposable income.
Disposable income is aggregate income minus taxes plus transfer payments. The greater the disposable income, the greater is the quantity of consumption goods and services that households plan to buy and the greater is aggregate demand.

Government expenditure on goods and services is one component of aggregate demand. So if the government spends more on spy satellites, schools, and highways, aggregate demand increases The Federal Reserve's (Fed's) attempt to influence the economy by changing interest rates and the quantity of money is called monetary policy. The Fed influences the quantity of money and interest rates by using the tools and methods described in Chapter 25.
An increase in the quantity of money increases aggregate demand through two main channels: It lowers interest rates and makes it easier to get aloan.
With lower interest rates, businesses plan a greater level of investment in new capital and households plan greater expenditure on new homes, on home improvements, on automobiles, and a host of other consumer durable goods. Banks and others eager to lend lower their standards for making loans and more people are able to get home loans and other consumer loans.
A decrease in the quantity of money has the opposite effects and lowers aggregate demand.

## ECONOMICS IN ACTION

## World Economy Headwinels

As the U.S. economy expanded in 2014, it faced strong headwinds from the global economy.

Although the U.S. economy is the world's largest economy and it can generate much of its growth from consumer and business investment spending, the United States gets help from the other large economies, China, the European Union, and Japan.


European Central Bank, Frankfurt, Germany

In 2014, real GDP growth in the euro area ground to a halt, real GDP shrank in Japan, and growth slowed in China.

Additionally, in 2014, the U.S. dollar appreciated making U.S. exporters and producers of importcompeting goods and services less competitive.

A weak world economy and a strong dollar slows U.S. exports growth and leaves a recessionary gap with spare capacity at ports like that at Long Beach, California.


Port of Long Beach, California

The World Economy Two main influences that the world economy has on aggregate demand are theexchange rate and foreign income. The exchange rate is the amount of a foreign currency that you can buywith a U.S. dollar. Other things remaining the same, a rise in the exchange rate decreases aggregate demand. To see how the exchange rate influences aggregate demand, suppose that the exchange rate is 1.20 euros per U.S. dollar. An Airbus plane made in France costs 120 million euros and an equivalent Boeing airplane made in the United States costs $\$ 110$ million. In U.S. dollars, the Airbus plane costs $\$ 100$ million, so airlines both in the United States and around the world buy the cheaper airplane from France. Now suppose the exchange rate falls to 1 euro per U.S. dollar. The Airbus airplane now costs $\$ 120$ million and is more expensive than the Boeing air- plane. Airlines will
switch from Airbus to Boeing. U.S. exports will increase and U.S. imports will decrease, so U.S. aggregate demand will increase.

## FIGURE 27.5 Changes in Aggregate Demand



Aggregate demand

Decreases if.

- Expected future income, inflation, or profit decreases
- Fiscal policy decreases government expenditure, increases taxes, or decreases transfer payments
- Monetary policy decreases the quantity of money and increas. es interest rates
- The exchange rate increases or foreign income decreases

Increases if.

- Expected future income, inflation, or profit increases
- Fiscal policy increases government expenditure, decreases taxes, or increases transfer payments
- Monetary policy increases the quantity of money and decreases interest rates
- The exchange rate decreases or foreign income increases

An increase in foreign income increases .S. exports and increases U.S. aggregate demand. For example, an increase in income in Japan and Germany increases Japanese and German consumers' and producers' planned expenditures on U.S.-produced goods and services.

Shifts of the Aggregate Demand Curve When aggregate demand changes, the aggregate demand curve shifts. Figure 27.5 shows two changes in aggregate demand and summarizes the factors that bring aboutsuch changes.
Aggregate demand increases and the $A D$ curve shifts rightward from $A D_{0}$ to $A D_{1}$ when expected future income, inflation, or profit increases; governmentexpenditure on goods and services increases; taxes are cut; transfer payments increase; the quantity of money increases and the interest rate falls; the exchange rate falls; or foreign income increases.
Aggregate demand decreases and the $A D$ curve shifts leftward from $A D_{0}$ to $A D_{2}$ when expected futureincome, inflation, or profit decreases; government expenditure on goods and services decreases; taxes increase; transfer payments decrease; the quantity of money decreases and the interest rate rises; the exchange rate rises; or foreign income decreases.

## Explaining Macroeconomic Trends and Fluctuations

The purpose of the $A S-A D$ model is to explain changes in real GDP and the price level. The model's main purpose is to explain business cycle fluctuations in these variables. But the model also aids our understanding of economic growth and inflation trends. We begin by combining aggregate supply and aggregate demand to determine real GDP and the price level in equilibrium. Just as there are two time frames for aggregate supply, there are two time frames for macroeconomic equilibrium: a long-run equilibrium and a short-run equilibrium. We'll first look at short- run equilibrium.

## Short-Run Macroeconomic Equilibrium

The aggregate demand curve tells us the quantity of real GDP demanded at each price level, and the short-run aggregate supply curve tells us the quantity of real GDP supplied at each price level. Short-run macroeconomic equilibrium occurs when the quantity of real GDP demanded equals the quantity of real GDP supplied. That is, short-run macroeconomic equilibrium occurs at the point of intersection of the $A D$ curve and the SAS curve.
Figure 27.6 shows such an equilibrium at a price level of 110 and real GDP of $\$ 16$ trillion (points C and $\mathrm{C}^{\prime}$ ).
To see why this position is the equilibrium, think about what happens if the price level is some- thing other than 110 . Suppose, for example, that the price level is 120 and that real GDP is $\$ 17$ trillion (at point $E$ on the SAS curve). The quantity ofreal GDP demanded is less than $\$ 17$ trillion, so firms are unable to sell all their output. Unwantedinventories pile up, and firms cut both productionand prices. Production and prices are cut until firms can sell all their output. This situation occurs only when real GDP is $\$ 16$ trillion and the price level is 110.
Now suppose the price level is 100 and real GDP is $\$ 15$ trillion (at point $A$ on the SAS curve). Thequantity of real GDP demanded exceeds $\$ 15$ trillion, so firms are unable to meet the demand for their output. Inventories decrease, and customers clamor for goods and services, so firms increase production and raise prices. Production and prices increase until firms can meet the demand for their output. This situation occurs only when real GDP is $\$ 16$ trillion and the price level is 110 .

FIGURE 27.6 Short-Run Equilibrium


Short-run macroeconomic equilibrium occurs when real GDP demanded equals real GDP supplied-at the intersection of the aggregate demand curve $(A D)$ and the short-run aggregate supply curve (SAS).
In the short run, the money wage rate is fixed. It does not adjust to move the economy to full employment. So in the short run, real GDP can be greater than or less than potential GDP. But in the long run, the money wage rate adjusts and real GDP moves toward potential GDP. Let's look at the long-run equilibrium and see how we get there.

Long-Run Macroeconomic Equilibrium occurs when real GDP equals potential GDP-equivalently, when the economy is on its SAS curve.
When the economy is away from long-run equilibrium, the money wage rate adjusts. If the money wage rate is too high, short-run equilibrium is below;: potential GDP and the unemployment rate is above the natural rate. With an excess supply of labor, the money wage rate falls. If the money wage rate is too low, short-run equilibrium is above potential GDP and the unemployment rate is below the natural rate With an excess demand for labor, the money wage rate rises.
Figure 27.7 shows the long-run equilibrium and how it comes about. If the short-run aggregate supply curve is $S A S_{1}$, the money wage rate is too high to achieve full employment. A fall in the money wage rate shifts the SAS curve to SAS and brings full employment. If the short-run aggregate supply curve is $S A S_{2}$, the money wage rate is too low to achieve full employment. Now, a rise in the money wage rate shifts the SAS curve to SAS and brings full employment. In long-run equilibrium, potential GDP deter- mines real GDP, and potential GDP and aggregate demand together determine the price level. The money wage rate adjusts until the SAS curve passes through the long-run equilibrium point.

Let's now see how the AS-AD model helps us to understand economic growth and inflation.
FIGURE 27.7 Long-Run Equilibrium


In long-run macroeconomic equilibrium, real GDP equals potential GDP. So long-run equilibrium occurs where the aggregate demand curve, $A D$, intersects the long-run aggregate supply curve, LAS. In the long run, aggregate demand determines the price level and has no effect on real GDP. The money wage rate adjusts in the long run, so that the SAS curve intersects the LAS curve at the long-run equilibrium price level.

## Economic Growth and Inflation in the AS-AD Model

Economic growth results from a growing labor force and increasing labor productivity, which together make potential GDP grow (Chapter 23, pp.
548-551). Inflation results from a growing quantity of money that outpaces the growth of potential GDP (Chapter 25, pp. 608-609).
The $A S-A D$ model explains and illustrates economic growth and inflation. It explains economic growth as increasing long-run aggregate supply and itexplains inflation as a persistent increase in aggregatedemand at a faster pace than that of the increase in potential GDP.

Figure 27.8 illustrates this explanation in terms of the shifting $S A S$ and $A D$ curves. When the $L A S$ curve shifts rightward from $S A S_{0}$ to $L A S_{1}$, potential GDP grows from $\$ 16$ trillion to $\$ 17$ trillion. And in long-run equilibrium, real GDP also grows to $\$ 17$ trillion. When the $A D$ curve shifts rightward from ADo to AD 1 and the growth of aggregate demand outpaces the growth of potential GDP, the price level rises from 110 to 120.

If aggregate demand were to increase at the same pace as long-run aggregate supply, real GDP would grow with no inflation.

## ECONOMICS IN ACTION

## U.S. Economic Growth and Inflation

The figure is a scatter diagram of U.S. real GDP and the price level. The graph has the same axes as those of the $A S-A D$ model. Each dot represents a year between 1960 and 2014. The red dots are recession years. The pattern formed by the dots shows the combination of economic growth and inflation. Economic growth was fastest during the 1960 s; inflation was fastest during the 1970 s.

The $A S$ - $A D$ model interprets each dot as being at the intersection of the $S A S$ and $A D$ curves.


The Path of Real GDP and the Price Level

Our economy experiences periods of growth and inflation, like those shown in Fig. 27.8, but it does not experience steady growth and steady inflation.
Real GDP fluctuates around potential GDP in a business cycle. When we study the business cycle, we ignore economic growth and focus on the fluctuations around the trend growth rate. By doing so, we see the business cycle more clearly. Let's now see how the $A S-A D$ model explains the business cycle.

FIGURE 27.8 Economic Growth and Inflation


Economic growth results from a persistent increase in potential GDP-a rightward shift of the LAS curve. Inflation results from persistent growth in the quantity of money that shifts the $A D$ curve rightward at a faster pace than the real GDP growth rate.

## The Business Cycle intheAS-AD Model

The business cycle occurs because aggregate demand and short-run aggregate supply fluctuate but the money wage rate does not adjust quickly enough to keep real GDP at potential GDP. Figure 27.9 shows three types of short-run equilibrium.
Figure 27.9(a) shows an above full-employment equilibrium. An abovefull-employment equilibrium is an equilibrium in which real GDP exceeds potential GDP. The gap between real GDP and potential GDP is the output gap. When real GDP exceeds potential GDP, the output gap is called an inflationary gap.
The above full-employment equilibrium shown in Fig. 27.9(a) occurs where the aggregate demand curve $A D_{0}$ intersects the short-run aggregate supply curve SASO at a real GDP of $\$ 16.2$ trillion. There is an inflationary gap of $\$ 0.2$ trillion.

In Fig. 27.9(b), real GDP equals potential GDP and there is full-employment equilibrium. Inthis ex- ample, the equilibrium occurs where the aggregate demand curve $A D_{1}$ intersects the short-run aggregatesupply curve $S A S_{1}$ at a real GDP and potential GDP of $\$ 16$ trillion. In part (c), there is a below full-employment equilibrium. A below full-employment equilibrium is an equilibrium in which potential GDP exceeds real GDP. When potential GDP exceeds real GDP, the output gap is called a recessionarygap.

## ECONOMICS IN ACTION

## The U.S. Business Cycle

The U.S. economy had an inflationary gap in 2000 (at $A$ in the figure), full employment in 2007 (at $B$ ), and a recessionary gap in 2009 (at $C$ ). The fluctuating output gap in the figure is the real-world version of Fig. 27.9 (d) and is generated by fluctuations in aggregate demand and short-run aggregate supply.


## The U.S. Output Gap

The below full-employment equilibrium shown in Fig. 27.9(c) occurs where the aggregate demand curve $A D_{2}$ intersects the short-run aggregate supply curve $S A S_{2}$ at a real GDP of $\$ 15.8$ trillion. Potential GDP is $\$ 16$ trillion, so the recessionary gap is $\$ 0.2$ trillion.
The economy moves from one type of macroeconomic equilibrium to another as a result of fluctuations in aggregate demand and in short-run aggregate supply. These fluctuations produce fluctuations in real GDP. Figure 27.9(d) shows how real GDP fluctuates around potential GDP.
Let's now look at some of the sources of these fluctuations around potential GDP.
figure 27.9 The Business Cycle



Part (a) shows an above full-employment equilibrium in year 1 ; part (b) shows a full-employment equilibrium in year 2; and part (c) shows a below full-employment equilibrium in year 3. Part (d) shows how real GDP fluctuates around potential GDP in a business cycle during year 1, year 2, and year 3 .

In year 1, an inflationary gap exists and the economy is at point $A$ in parts (a) and (d). In year 2 , the economy is at full employment and the economy is at point $B$ in parts (b) and (d). In year 3, a recessionary gap exists and the economy is at point $C$ in parts (c) and (d).

## (d) Fluctuations in real GDP

## Fluctuations in Aggregate Demand

One reason real GDP fluctuates around potential GDP is that aggregate demand fluctuates. Let's see what happens when aggregate demandincreases.
Figure 27.10(a) shows an economy at full employment. The aggregate demand curve is $A D O$, the short-run aggregate supply curve is $S A S_{0}$, and the long-run aggregate supply curve is LAS. Real GDP equals potential GDP at $\$ 16$ trillion, and the price level is 110. Now suppose that the world economy expands and that the demand for U.S.-produced goods increases in Asia and Europe. The increase in U.S. exports increases aggregate demand in the United States, and the aggregate demand curve shifts right- ward from $A D_{0}$ to $A D_{1}$ in Fig. 27.10(a).
Faced with an increase in demand, firms increase production and raise prices. Real GDP increases to $\$ 16.5$ trillion and the price level rises to 115 . The economy is now in an above full-employment equilibrium. Real GDP exceeds potential GDP, and there is an inflationary gap. The increase in aggregate demand has increased the prices of 11 goods and services. Faced with higher prices, firms increased their output rates. At thisstage, prices of goods and
services have increased but the money wage rate has not changed. (Recall that as we move along the SAS curve, the money wage rate is constant.)
The economy cannot produce in excess of potential GDP forever. Why not? What are the forces at work that bring real GDP back to potential GDP?
Because the price level has increased and the money wage rate is unchanged, workers have experienced a fall in the buying power of their wages and firms' profits have increased. Under these circumstances, workers demand higher wages and firms, anxious to maintain their employment and output levels, meet those demands. If firms do not raise the money wage rate, they will either lose workers or have to hire less productive ones.
As the money wage rate rises, the short-run aggregate supply begins to decrease. In Fig. 27.10(b), the short-run aggregate supply curve begins to shift from $S A S_{0}$ toward $S A S_{1}$. The rise in the money wage rate and the shift in the SAS curve produce a sequence of new equilibrium positions. Along the adjustment path, real GDP decreases and the price level rises. The economy moves up along its aggregate demand curve as shown by the arrows in the figure.
FIGURE 27.10 An Increase in Aggregate Demand

(a) Short-run effect

An increase in aggregate demand shifts the aggregate demand curve from $A D_{0}$ to $A D_{1}$. In shortrun equilibrium, real GDP increases to $\$ 16.5$ trillion and the price level rises to 115. In this situation, an inflationary gap exists. In the long run in part (b), the money wage rate starts to rise and short-run

(b) Long-run effect
aggregate supply starts to decrease. The SAS curve gradually shifts from $S A S_{0}$ toward $S A S_{1}$, intersecting the aggregate demand curve $A D_{1}$ at higher price levels and real GDP decreases. Eventually, the price level has risen to 125 and real GDP has decreased to $\$ 16$ trillion-potential GDP.

Eventually, the money wage rate rises by the same percentage as the price level. At this time, the aggregate demand curve $A D_{1}$ intersects $S A S_{1}$ at new full- employment equilibrium. The price level has risen to 125 , and real GDP is back where it started, at potential GDP. A decrease in aggregate demand has effects similar but opposite to those of an increase in aggregate demand. That is, a decrease in aggregate demand shifts the aggregate demand curve leftward. Real GDP decreases to less than potential GDP, and a recessionary gap emerges. Firms cut prices. The lower price level increases the purchasing power of wages and increases firms' costs relative to their output prices because the money wage rate is
unchanged. Eventually, the money wage rate falls and the short-run aggregate supply increases.
Let's now work out how real GDP and the price level change when aggregate supply changes.
Fluctuations in Aggregate Supply Fluctuations in short-run aggregate supply can bring fluctuations in real GDP around potential GDP.
Suppose that initially real GDP equals potential GDP.
Then there is a large but temporary rise in the price of oil. What happens to real GDP and the price level?
Figure 27.11 answers this question. The aggregate demand curve is $A D_{0}$, the short-run aggregate supply curve is $S A S_{0}$, and the long-run aggregate supply curve is LAS. Real GDP is $\$ 16$ trillion, which equals potential GDP, and the price level is 110 . Then the price of oil rises. Faced with higher energy and transportation costs, firms decrease production. Short-run aggregate supply decreases, and the short-run aggregate supply curve shifts leftward to $S A S_{1}$. The price level rises to 120, and real GDP decreases to $\$ 15.5$ trillion. Because real GDP decreases, the economyexperiences recession. Because the price level increases, the economy experiences inflation. A combination of recession and inflation is called stagflation. The United States experienced stagflation in the mid-1970s and early 1980s, but events like this are not common.
When the price of oil returns to its original level, the economy returns to full employment.

## Macroeconomic Schools of Thought

Macroeconomics is an active field of research, and much remains to be learned about the forces that make our economy grow and fluctuate. There is a greater degree of consensus and certainty about economic growth and inflation-the longer-term trendsin real GDP and the price level-than there is aboutthe business cycle-the short-term fluctuations in these variables. Here, we'll look only at differences of view about short-term fluctuations. The $A S-A D$ model that you've studied in this chapter provides a good foundation for understanding the range of views that macroeconomists hold about this topic. But what you will learn here is just a first glimpse at the scientific controversy and debate. We'll return to these issues at various points later in the text and deepen your appreciation of the alternative views.
Classification usually requires simplification, and classifying macroeconomists is no exception to this general rule. The classification that we'll use here is simple, but it is not misleading. We're going to divide macroeconomists into three broad schools of thought and examine the views of each group in turn. The groups are

- Classical
- Keynesian
- Monetarist

FIGURE 27.11 A Decrease in Aggregate Supply


An increase in the price of oil decreases short-run aggregate supply and shifts the short-run aggregate supply curve from $S A S_{0}$ to $S A S_{1}$. Real GDP decreases from $\$ 16$ trillion to $\$ 15.5$ trillion, and the price level rises from 110 to 120 . The economy experiences stagflation.

## The Classical View

A classical macroeconomist believes that the economy is self-regulating and always at full employment. The term "classical" derives from the name of the founding school of economics that includes Adam Smith, David Ricardo, and John Stuart Mill.
A new classical view is that business cycle fluctuations are the efficient responses of a wellfunctioning market economy that is bombarded by shocks that arise from the uneven pace of technological change.
The classical view can be understood in terms of beliefs about aggregate demand and aggregate supply.

Aggregate Demand Fluctuations In the classical view, technological change is the most significant influence on both aggregate demand and aggregate supply. For this reason, classical macroeconomists don't use the AS-AD framework. But their views can be interpreted in this framework. A technological change that increases the productivity of capital brings an increase in aggregate demand because firms increase their expenditure on new plant and equipment. A technological change that lengthens the useful life of existing capital decreases the demand for new capital, which decreases aggregate demand.

Aggregate Supply Response In the classical view, the money wage rate that lies behind the short-run aggregate supply curve is instantly and completely flexible. The money wage rate adjusts so quickly to maintain equilibrium in the labor market that real GDP always adjusts to equal potential GDP.
Potential GDP itself fluctuates for the same reasons that aggregate demand fluctuates: technological change. When the pace of technological change is rapid, potential GDP increases quickly and so does real GDP. And when the pace of technological change slows, so does the growth rate of potential GDP.

Classical Policy The classical view of policy emphasizes the potential for taxes to stunt incentives andcreate inefficiency. Minimizing the disincentive effects of taxes will allow employment, investment, and technological advance to be at their efficient levels and the economy to expand at an appropriate and rapid pace.

## The Keynesian View

A Keynesian macroeconomist believes that left alone, the economy would rarely operate at full employment and that to achieve and maintain full employment, active help from fiscal policy and monetary policy is required.
The term "Keynesian" derives from the name of one of the twentieth century's most famous econmists, John Maynard Keynes (see p. 765).
The Keynesian view is based on beliefs about the forces that determine aggregate demand and short- run aggregate supply.

Aggregate Demand Fluctuations In the Keynesian view, expectations are the most significant influenceon aggregate demand. Those expectations are based on herd instinct, or what Keynes himself called "animal spirits." A wave of pessimism about future profitprospects can lead to a fall in aggregate demand and plunge the economy into recession.

Aggregate Supply Response Inthe Keynesian view, the money wage rate that lies behind the short-runaggregate supply curve is extremely sticky in the downward direction. Basically, the money wage rate doesn't fall. So if there is a recessionary gap, there is no automatic mechanism for getting rid of it. If it were to happen, a fall in the money wage rate would increase short-run aggregate supply and restore full employment. But the money wage rate doesn't fall, so the economy remains stuck in recession.
A modern version of the Keynesian view, known as the new Keynesian view, holds not only that the moneywage rate is sticky but also that prices of goods and services are sticky. With a sticky price level, the short-run aggregate supply curve is horizontal at a fixed price level.

Policy Response Needed The Keynesian view calls for fiscal policy and monetary policy to actively offset changes in aggregate demand that bring recession.
By stimulating aggregate demand in a recession, full employment can be restored.

## The Monetarist View

A monetarist is a macroeconomist who believes that the economy is self-regulating and that it will normally operate at full employment, provided that monetary policy is not erratic and that the pace of money growth is kept steady.
The term "monetarist" was coined by an outstanding twentieth-century economist, Karl Brunner, to describe his own views and those of Milton Friedman (see p. 821).
The monetarist view can be interpreted in terms of beliefs about the forces that determine aggregate demand and short-run aggregate supply.

Aggregate Demand Fluctuations In the monetarist view, the quantity of money is the most significant influence on aggregate demand. The quantity of money is determined by the Federal Reserve (the Fed). If the Fed keeps money growing at a steadypace, aggregate demand fluctuations will be minimized and the economy will operate close to full employment. But if the Fed decreases the quantity of money or even just slows its growth rate too abruptly, the economy will go into recession. In the monetaristview, all recessions result from inappropriate monetary policy.

Aggregate Supply Response The monetarist view of short-run aggregate supply is the same as the Keynesian view: the money wage rate is sticky. If the economy is in recession, it will take an unnecessarily long time for it to return unaided to full employment.

Monetarist Policy The monetarist view of policy is the same as the classical view on fiscal policy. Taxes should be kept low to avoid disincentive effects that decreasepotential GDP. Provided that the quantity of money iskept on a steady growth path, no active stabilization isneeded to offset changes in aggregate demand.

## The Way Ahead

In the chapters that follow, you're going to encounterKeynesian, classical, and monetarist views again. In the next chapter, we study the original Keynesian model
of aggregate demand. This model remains useful today because it explains how expenditure fluctuations aremagnified and bring changes in aggregate demand thatare larger than the changes in expenditure. We then goon to apply the $A S-A D$ model to a deeper look at the business cycle, inflation, and deflation.
Our attention then turns to macroeconomic policy - the fiscal policy of the Administration andCongress and the monetary policy of the Fed.

## ECONOMIC ANALYSIS

- U.S. real GDP grew at an annual rate of 4.2 percent during the second quarter of 2014-a faster than average growth rate and slightly higher than the original estimate a month earlier.
- In the second quarter of 2014, real GDP was estimated to be $\$ 16.0$ trillion. The price level was 108 (up 8 percent since 2009).
- A year earlier, in the second quarter of 2013, real GDP was $\$ 15.6$ trillion and the price level was 106.
- Figure 1 illustrates the situation in the second quarter of 2013. The aggregate demand curve was $A D_{13}$ and the short-run aggregate supply curve was $S A S_{13}$. Real GDP ( $\$ 15.6$ trillion) and the price level (106) are at the intersection of these curves.
- The Congressional Budget Office (CBO) estimated that potential GDP in the second quarter of 2013 was $\$ 16.5$ trillion, so the long-run aggregate supply curve in 2013 was LAS 13 in Fig. 1.
- Figure 1 shows the output gap in 2013, which was a recessionary gap of $\$ 0.9$ trillion-about 5.5 percent of potential GDP.
- During the year from June 2013 to June 2014, the labor force increased, the capital stock increased, and labor productivity increased. Potential GDP increased to an estimated \$16.7 trillion.
- In Fig. 2, the LAS curve shifted rightward to $L A S_{14}$.
- Also during the year from June 2013 to June 2014, a combination of fiscal and monetary policy stimulus and an increase in demand from a slowly expanding world economy increased aggregate demand.
- The increase in aggregate demand exceeded the increase in long-run aggregate supply, and the $A D$ curve shifted rightward to $A D_{14}$.
- Two forces act on short-run aggregate supply: The increase in potential GDP shifts the SAS curve rightward and a rise in the money wage rate and other factor prices shifts the SAS curve leftward.
- Because potential GDP didn't increase by much, shortrun aggregate supply probably didn't change by much, and here we assume it didn't change at all. The SAS curve in 2014, $S A S_{14}$, was the same as $S A S_{13}$.
- Real GDP increased to $\$ 16.0$ trillion and the price level increased to 108.


Figure 1 AS-AD in Second Quarter of 2013


Figure 2 AS-AD in Second Quarter of 2014

- The output gap narrowed slightly to $\$ 0.7$ trillion or 4.4 percent of potential GDP.
- If real GDP continues to grow at 4 percent a year for the rest of 2014, the output gap will shrink further.
- If real GDP maintains an annual growth rate of 4 percent, and if the CBO estimates of potential GDP are correct, the output gap will be zero and full employment will be restored in the fourth quarter of 2016.


## CHAPTER 28: Expenditure Multipliers

After studying this chapter, you will be able to:

- Explain how expenditure plans are determined when the price level is fixed
- Explain how real GDP is determined when the price level is fixed
- Explain the expenditure multiplier
- Explain the relationship between aggregate expenditure and aggregate demand

Investment and inventories fluctuate like the volume of a rock singer's voice and the uneven surface of New york City Street. How does the economy react to those fluctuations? Does it behave like an amplifier, blowing up the fluctuations and spreading them out to affect the many millions of participants in an economic rock concert? Or does it react like a limousine, absorbing the shocks and providing a smooth ride for the economy's passengers?
You will explore these questions in this chapter and in Economics in the News at the end of the chapter you will see the role played by inventoryinvestment during 2014 as the economy expanded.

## Fixed Prices and Expenditure Plans

In the model that we study in this chapter, all the firms are like your grocery store: They set their prices and sell the quantities their customers are willing to buy. If they persistently sell more than they plan to and keep running out of inventory, they eventually raise their prices. And if they persistently sell less than they plan to and have inventories piling up, they eventually cut their prices. But on any given day, their prices are fixed and the quantities they sell depend on demand, not supply.
Because each firm's prices are fixed, for the economy as a whole

1. The price level is fixed, and
2. Aggregate demand determines real GDP.

We call this model the Keynesian model because it was first suggested by John Maynard Keynes (see p.765) as a model of persistent depression.
We begin by identifying the forces that determine expenditure plans.

## Expenditure Plans

Aggregate expenditure has four components: consumption expenditure, investment, government expenditure on goods and services, and net exports (exports minus imports). These four components sum to real GDP (see Chapter 21, pp. 531-532).

Aggregate planned expenditure is equal to the sum of the planned levels of consumption expenditure, investment, government expenditure on goods and services, and exports minus imports. Two of these components of planned expenditure, consumption expenditure and imports, change when income changes and so they depend on real GDP.

A Two-Way Link Between Aggregate Expenditure and real GDP There is a two-way link between aggregate expenditure and real GDP. Other thingsremaining the same,

- An increase in real GDP increases aggregate expenditure, and
- An increase in aggregate expenditure increases real GDP.

Youare now going to study this two-way link.

## Consumption and Saving Plans

Several factors influence consumption expenditure and saving plans. The more important ones are

- Disposable income
- Real interest rate
- Wealth
- Expected future income

Disposable income is aggregate income minus taxes.t plus transfer payments. Aggregate income equals real;' GDP, so disposable income depends on real GDP. To:(explore the two-way link between real GDP and planned consumption expenditure, we focus on the ;:relationship between consumption expenditure and disposable income when the other three factors listedabove are constant.

Consumption Expenditure and Saving The table in Fig. 28.1 lists the consumption expenditure and the saving that people plan at each level of disposable income. Households can only spend their disposable income on consumption or save it, so planned consumption expenditure plus planned saving always equals disposable income.
The relationship between consumption expenditure and disposable income, other things remaining the same, is called the consumption function. The relationship between saving and disposable income, other things remaining the same, is called the saving function.

Consumption Function Figure 28.1(a) shows a consumption function. The $y$-axis measures consumption expenditure, and the $x$-axis measures disposable income. Along the consumption function, the point labeled $A$ through $F$ correspond to the rows of the table. For example, point $E$ shows that when disposable income is $\$ 8$ trillion, consumption expenditure is $\$ 7.5$ trillion. As disposable income increases, consumption expenditure also increases. At point $A$ on the consumption function, consumption expenditure is $\$ 1.5$ trillion even though . . disposable income is zero. This consumption expenditure is called autonomous consumption, and it is this amount of consumption expenditure that would take place in the short run even if people had no current income. Consumption expenditure in excess of this amount is called induced consumption, which is the consumption expenditure that is induced by an increase in disposable income.
$45^{\circ}$ Line Figure 28.1(a) also contains a $45^{\circ}$ line, the height of which measures disposable income. At each point on this line, consumption expenditure equals disposable income. Between points $A$ and $D$, consumption expenditure exceeds disposable income, between points $D$ and $F$ consumption expenditure
is less than disposable income, and at point $D$, consumption expenditure equals disposable income.
FIGURE 28.1 Consumption Function and Saving Function

(b) Saving function

|  | Disposable <br> income | Planned <br> consumption <br> expenditure | Planned <br> saving |
| :---: | :---: | :---: | :---: |
|  | (trillions of 2009 dollars) |  |  |
| A | 0 | 1.5 | -1.5 |
| B | 2 | 3.0 | -1.0 |
| C | 4 | 4.5 | -0.5 |
| D | 6 | 6.0 | 0 |
| E | 8 | 7.5 | 0.5 |
| F | 10 | 9.0 | 1.0 |

The table shows consumption expenditure and saving plans at various levels of disposable income. Part (a) of the figure shows the relationship between consumption expenditure and disposable income (the consumption function). The height of the consumption function measures consumption expenditure at each level of disposable income. Part (b) shows the relationship between saving and disposable income (the saving function). The height of the saving function measures saving at each level of disposable income. Points $A$ through $F$ on the consumption and saving functions correspond to the rows in the table.

The height of the $45^{\circ}$ line in part (a) measures disposable income. So along the $45^{\circ}$ line, consumption expenditure equals disposable income. Consumption expenditure plus saving equals disposable income. When the consumption function is above the $45^{\circ}$ line, saving is negative (dissaving occurs). When the consumption function is below the $45^{\circ}$ line, saving is positive. At the point where the consumption function intersects the $45^{\circ}$ line, all disposable income is spent on consumption and saving is zero.

Saving Function Figure 28 .1(b) shows a saving function. Again, the points $A$ through $F$ correspond to the rows of the table. For example, point $E$ shows that when disposable income is $\$ 8$ trillion, saving is $\$ 0.5$ trillion. As disposable income increases, saving increases. Notice that when consumption expenditure exceeds disposable income in part (a), saving is negative, called dissaving, in part (b).

## Marginal Propensities to Consume and Save

The marginal propensity to consume (MPC) is the fraction of a change in disposable income that is spent on consumption. It is calculated as the change in consumption expenditure $(\Delta C)$ divided by the change $m$ disposable income ( $\Delta \mathrm{YD}$ ). The formula is

$$
M P C=\frac{\Delta C}{\Delta Y D}
$$

In the table in Fig. 28.1, when disposable income increases by $\$ 2$ trillion, consumption expenditure increases by $\$ 1.5$ trillion. The MPC is $\$ 1.5$ trillion divided by $\$ 2$ trillion, which equals 0.75 .

The marginal propensity to save (MPS) is the fraction of a change in disposable income that is saved. It is calculated as the change in saving $(\Delta \mathrm{S})$ divided by the change in disposable income ( $\Delta$ YD). The formula is

$$
M P S=\frac{\Delta S}{\Delta Y D}
$$

In the table in Fig. 28.1, when disposable income increases by $\$ 2$ trillion, saving increases by $\$ 0.5$ trillion. The MPS is $\$ 0.5$ trillion divided by $\$ 2$ trillion, which equals 0.25 .
Because an increase in disposable income is either spent on consumption or saved, the marginal propensity to consume plus the marginal propensity to save equals 1. You can see why by using the equation:

$$
\Delta C+\Delta S=\Delta Y D
$$

figure 28.2 The Marginal Propensities to Consume and Save


Divide both sides of the equation by the change in disposable income to obtain

$$
\frac{\Delta C}{\Delta Y D}+\frac{\Delta S}{\Delta Y D}=1
$$

$\triangle C / \triangle Y D$ is the marginal propensity to consume (MPC), and $\triangle S / \triangle Y D$ is the marginal propensity to save (MPS), so

$$
M P C+M P S=1
$$

## Slopes and Marginal Propensities

The slope of the consumption function is the marginal propensity to consume, and the slope of the saving function is the marginal propensity to save.
Figure 28.2(a) shows the MPC as the slope of the consumption function. An increase in disposable income of $\$ 2$ trillion is the base of the red triangle. The increase in consumption expenditure that results from this increase in disposable income is $\$ 1.5$ trillion and is the height of the triangle. The slope of the consumption function is given by the formula "slope equals rise over run" and is $\$ 1.5$ trillion divided by $\$ 2$ trillion, which equals 0.75-the MPC.
Figure 28.2(b) shows the MPS as the slope of the saving function. An increase in disposable income of $\$ 2$ trillion (the base of the red triangle) increases saving by $\$ 0.5$ trillion (the height of the triangle). The slope of the saving function is $\$ 0.5$ trillion divided by $\$ 2$ trillion, which equals 0.25-the MPS.

## Consumption as a Function of Real GDP

Consumption expenditure changes when disposable income changes and disposable income changes when real GDP changes. So consumption expenditure depends not only on disposable income but also on real GDP. We use this link between consumption expenditure and real GDP to determine equilibriumexpenditure. But before we do so, we need to look at one further component of aggregate expenditure: imports. Like consumption expenditure, imports are influenced by real GDP.

## Import Function

Of the many influences on U.S. imports in the short run, U.S. real GDP is the main influence. Otherthings remaining the same, an increase in U.S. real GDP increases the quantity of U.S. imports.
The relationship between imports and real GDP is determined by the marginal propensity to import, which is the fraction of an increase in real GDP that is spent on imports. It is calculated as the change in imports divided by the change in real GDP, other things remaining the same. For example, if an increase in real GDP of $\$ 1$ trillion increases imports by $\$ 0.25$ trillion, the marginal propensity to import is 0.25 .

## Real GDP with a Fixed Price Level

You are now going to see how, at a given price level, aggregate expenditure plans determine real GDP. Westart by looking at the relationship between aggregateplanned expenditure and real GDP. This relationshipcan be described by an aggregate expenditure schedule or an aggregate expenditure curve. The aggregate expenditure schedule lists aggregate planned expenditure generated at each level of real GDP. The aggregate expenditure curve is a graph of the aggregate expenditure schedule.

## ECONOMICS IN ACTION

## The U.S. Consumption Function

The figure shows the U.S. consumption function. Each point identified by a blue dot represents consumption expenditure and disposable income for a particular year. (The dots are for the years 1980 to 2014, and the dots for five of those years are identified in the figure.)

The U.S. consumption function is $C F_{0}$ in 1980 and $C F_{1}$ in 2014.

The slope of the consumption function in the figure is 0.9 , which means that a $\$ 1$ increase in disposable income increases consumption expenditure by 90 cents. This slope, which is an estimate of the marginal propensity to consume, is an assumption that is at the upper end of the range of values that economists have estimated for the marginal propensity to consume.

The consumption function shifts upward over time as other influences on consumption expenditure change. Of these other influences, the real interest rate and wealth fluctuate and so bring upward and downward shifts in the consumption function.

But increasing wealth and increasing expected future income bring a steady upward shift in the consump tion function. As the consumption function shifts upward, autonomous consumption increases.


The U.S. Consumption Function

## Aggregate Planned Expenditure

The table in Fig. 28.3 sets out an aggregate expenditure schedule. To calculate aggregate planned expenditure at a given real GDP, we add the expenditure components together. The first column of the table shows real GDP, and the second column shows the planned consumption at each level of real GDP. A $\$ 1$ trillion increase in real GDP increases consumption expenditure by $\$ 0.7$ trillion-the MPC is 0.7 .
The next two columns show investment and government expenditure on goods and services, both of which are independent of the level of real GDP. Investment depends on the real interest rate and the expected profit (see Chapter 24, p. 613). At a given point in time, these factors generate a given level of investment. Suppose this level of investment is $\$ 2.5$ trillion. Also, suppose that government expenditure is $\$ 3.5$ trillion.
The next two columns show exports and imports.
Exports are influenced by events in the rest of the world, prices of foreign-produced goods and services relative to the prices of similar U.S.-produced goods and services, and exchange rates. But they are not directly affected by U.S. real GDP. Exports are a constant $\$ 2.0$ trillion. Imports increase as U.S. real GDP increases. A \$1 trillion increase in U.S. real GDP generates a $\$ 0.2$ trillion increase in imports-the marginal propensity to import is 0.2 . The final column shows aggregate planned expenditure-the sum of planned consumption expenditure, investment, government expenditure on goods and services, and exports minus imports.
Figure 28.3 plots an aggregate expenditurecurve.
Real GDP is shown on the $x$-axis, and aggregate planned expenditure is shown on the $y$ axis. The aggregate expenditure curve is the red line $A E$. Points $A$ through F on that curve correspond to the rows of the table. The $A E$ curve is a graph of aggregate planned expenditure (the last column) plotted against real GDP (the first column).
Figure 28.3 also shows the components of aggregate expenditure. The constant components- investment (I), government expenditure on goods and services (G), and exports (X)-are shown by the horizontal lines in the figure. Consumption expenditure (C) is the vertical gap between the lines labeled

$$
I+G+X \text { and } I+G+X+C .
$$

To construct the $A E$ curve, subtract imports ( $M$ ) from the $I+G+X+$ Cline. Aggregate expenditure is expenditure on U.S.-produced goods and services. But the components of aggregate expenditure-C, $I$, and G-include expenditure on imported goods and services. For example, if you buy a new cellphone, your expenditure is part of consumption expenditure. But if the cellphone is a Nokia made in Finland, your expenditure on it must be subtracted from consumption expenditure to find out how much is spent on goods and services produced in the United States - on U.S. real GDP. Money paid to Nokia for cellphone imports from Finland does not add to aggregate expenditure in the United States.
Because imports are only a part of aggregate expenditure, when we subtract imports from the other components of aggregate expenditure, aggregate planned expenditure still increases as real GDP increases, as you can see in Fig. 28.3.

FIGURE 28.3 Aggregate Planned Expenditure: The $A E$ Curve


Aggregate planned expenditure is the sum of planned consumption expenditure, investment, government expenditure on goods and services, and exports minus imports. For example, in row $C$ of the table, when real GDP is $\$ 14$ trillion, planned consumption expenditure is $\$ 9.8$ trillion, planned investment is $\$ 2.5$ trillion, planned government expenditure is $\$ 3.5$ trillion, planned exports are $\$ 2.0$ trillion, and planned imports are $\$ 2.8$ trillion. So when real GDP is $\$ 14$ trillion, aggregate planned expenditure is $\$ 15$ trillion $(\$ 9.8+\$ 2.5+\$ 3.5+$ $\$ 2.0-\$ 2.8$ ).

The schedule shows that aggregate planned expenditure increases as real GDP increases. This relationship is graphed as the aggregate expenditure curve $A E$. The components of aggregate expenditure that increase with real GDP are consumption expenditure and imports. The other componentsinvestment, government expenditure, and exports - do not vary with real GDP.

| Real GDP (Y) |  | Planned expenditure |  |  |  |  | $\begin{gathered} \begin{array}{c} \text { Aggregate } \\ \text { planned } \\ \text { expenditure } \end{array} \\ (A E=C+I+G+X-M) \end{gathered}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Consumption expenditure <br> (C) | Investment (I) | Government expenditure <br> (G) | Exports (X) | Imports <br> (M) |  |
| (trillions of $\mathbf{2 0 0 9}$ dollars) |  |  |  |  |  |  |  |
| A | 0 | 0 | 2.5 | 3.5 | 2.0 | 0.0 | 8.0 |
| B | 5 | 3.5 | 2.5 | 3.5 | 2.0 | 1.0 | 10.5 |
| $C$ | 14 | 9.8 | 2.5 | 3.5 | 2.0 | 2.8 | 15.0 |
| D | 15 | 10.5 | 2.5 | 3.5 | 2.0 | 3.0 | 15.5 |
| E | 16 | 11.2 | 2.5 | 3.5 | 2.0 | 3.2 | 16.0 |
| F | 17 | 11.9 | 2.5 | 3.5 | 2.0 | 3.4 | 16.5 |

Consumption expenditure minus imports, which varies with real GDP, is called induced expenditure. The sum of investment, government expenditure, and exports, which does not vary with real GDP, is called autonomous expenditure. Consumption expenditure and imports can also have an autonomous component-a component that does not vary with real GDP. Another way of thinking about autonomous expenditure is that it would be the level of aggregate planned expenditure if real GDP were zero.
In Fig. 28.3, autonomous expenditure is $\$ 8$ trillion-aggregate planned expenditure when real GDPis zero (point $A$ ). For each $\$ 1$ trillion increase in real GDP, induced expenditure increases by $\$ 0.5$ trillion. The aggregate expenditure curve summarizes the relationship between aggregate planned expenditure and real GDP. But what determines the point on the aggregate expenditure curve at which the economy operates? What determines actual aggregate expenditure?

## Actual Expenditure, Planned Expenditure, and Real GDP

Actual aggregate expenditure is always equal to real GDP, as we saw in Chapter 21 (p.532). But aggregate planned expenditure is not always equal to actual aggregate expenditure and
therefore is not always equal to real GDP. How can actual expenditure and planned expenditure differ? The answer is that firms can end up with inventories that are greater or smaller than planned. People carry out their consumption expenditure plans, the government implements its planned expenditure on goods and services, and net exports are as planned. Firms carry out their plans to purchase new buildings, plant, and equipment. But one component of investment is the change in firms' inventories. If aggregate planned expenditure is less than real GDP, firms sell less than they planned to sell and end up with unplanned inventories. If aggregate planned expenditure exceeds real GDP, firms sell more than they planned to sell and end up with inventories being toolow.

## Equilibrium Expenditure

Equilibrium expenditure is the level of aggregate expenditure that occurs when aggregate planned expenditure equals real GDP. Equilibrium expenditure is a level of aggregate expenditure and real GDP at which spending plans are fulfilled. At a given price level, equilibrium expenditure determines real GDP. When aggregate planned expenditure and actual aggregate expenditure are unequal, a process of convergence toward equilibrium expenditure occurs. Throughoutthis process, real GDP adjusts. Let's examine equilibrium expenditure and the process that brings it about.
Figure 28.4(a) illustrates equilibrium expenditure. The table sets out aggregate planned expenditure at various levels of real GDP. These values are plotted as points $A$ through $F$ along the $A E$ curve. The $45^{\circ}$ line shows all the points at which aggregate planned expenditure equals real GDP. So where the $A E$ curve lies above the $45^{\circ}$ line, aggregate planned expenditure exceeds real GDP; where the $A E$ curve lies below the $45^{\circ}$ line, aggregate planned expenditure is less than real GDP; and where the $A E$ curve intersects the $45^{\circ}$ line, aggregate planned expenditure equals real GDP. Point $D$ illustrates equilibrium expenditure. At this point, real GDP is $\$ 16$ trillion.

## Convergence to equilibrium

What are the forces that move aggregate expenditure toward its equilibrium level? To answer this question, we must look at a situation in which aggregate expenditure is away from its equilibrium level.

From Below Equilibrium Suppose that in Fig. 28.4, real GDP is $\$ 14$ trillion. With real GDP at $\$ 14$ trillion, actual aggregate expenditure is also $\$ 14$ trillion. But aggregate planned expenditure is $\$ 15$ trillion, point B in Fig. 28.4(a). Aggregate planned expenditure exceeds actual expenditure. When people spend $\$ 15$ trillion and firms produce goods and services worth $\$ 14$ trillion, firms' inventories fall by $\$ 1$ trillion, point B in Fig. 28.4(b). Because the change in inventories is part of investment, actual investment is $\$ 1$ trillion less than plannedinvestment.
Real GDP doesn't remain at \$14 trillion for very long. Firms have inventory targets based on their sales. When inventories fall below target, firms increase production to restore inventories to the tar- get level.

FIGURE 28.4 Equilibrium Expenditure


(b) Unplanned inventory changes

|  | Real GDP <br> $(\boldsymbol{Y})$ | Aggregate planned <br> expenditure <br> $(\boldsymbol{A E})$ | Unplanned <br> inventory change <br> $(\boldsymbol{Y}-\boldsymbol{A E})$ |  |
| :---: | :---: | :---: | :---: | :---: |
|  | (trillions of 2009 dollars) |  |  |  |
| A | 13 | 14.5 | -1.5 |  |
| B | 14 | 15.0 | -1.0 |  |
| C | 15 | 15.5 | -0.5 |  |
| D | 16 | 16.0 | 0 |  |
| E | 17 | 16.5 | 0.5 |  |
| F | 18 | 17.0 | 1.0 |  |

> The table shows expenditure plans at different levels of real GDP. When real GDP is $\$ 16$ trillion, aggregate planned expenditure equals real GDP.
> Part (a) of the figure illustrates equilibrium expenditure, which occurs when aggregate planned expenditure equals real GDP at the intersection of the $45^{\circ}$ line and the $A E$ curve. Part (b) of the figure shows the forces that bring about equilibrium expenditure.
> When aggregate planned expenditure exceeds real GDP, inventories decrease - for example, at point $B$ in both parts of the figure. So, firms increase production, and real GDP increases.
> When aggregate planned expenditure is less than real GDP, inventories increase-for example, at point $F$ in both parts of the figure. So, firms cut production, and real GDP decreases.
> When aggregate planned expenditure equals real GDP, there are no unplanned inventory changes and real GDP remains constant at equilibrium expenditure.

To increase inventories, firms hire additional labor and increase production. Suppose that they increase production in the next period by $\$ 1$ trillion. Real GDP increases by $\$ 1.0$ trillion to $\$ 15.0$ trillion. But again, aggregate planned expenditure exceeds real GDP. When real GDP is $\$ 15.0$ trillion, aggregate planned expenditure is $\$ 15.5$ trillion, point C in Fig. 28.4(a). Again, inventories decrease, but this time by less than before. With real GDP of $\$ 15.0$ trillion and aggregate planned expenditure of $\$ 15.5$ trillion, inventories decrease by $\$ 0.5$ trillion, point C in Fig. 28.4(b). Again, firms hire additional labor and production increases; real GDP increases yet further.
The process that we've just described-planned expenditure exceeds real GDP, inventories decrease, and production increases to restore inventories endswhen real GDP has reached $\$ 16$ trillion. At this real GDP, there is equilibrium. Unplanned inventory changes are zero. Firms do not change their production.

From Above Equilibrium If in Fig. 28.4, real GDP is $\$ 18$ trillion, the process that we've just described works in reverse. With real GDP at $\$ 18$ trillion, actual aggregate expenditure is also $\$ 18$ trillion. Butaggregate planned expenditure is $\$ 17$ trillion, point Fin Fig. 28.4(a). Actual expenditure exceeds planned expenditure. When people spend $\$ 17$ trillion and firms produce goods and services worth $\$ 18$ trillion, firms' inventories rise by $\$ 1$ trillion, point Fin Fig. 28.4(b). Now, real GDP begins to de:..crease. As long as actual expenditure exceeds plannedexpenditure, inventories rise, and production decreases. Again, the process ends when real GDP has reached $\$ 16$ trillion, the equilibrium at which unplanned inventory changes are zero and firms do not change their production.
We've learned that when the price level is fixed, real GDP is determined by equilibrium expenditure. And we have seen how unplanned changes in inventories and the production response they generate bring a convergence toward equilibrium expenditure. We're now going to study changes in equilibrium expenditure and discover an economic amplifier called the multiplier.

## The Multiplier

Investment and exports can change for many reasons. A fall in the real interest rate might induce firms to increase their planned investment. A wave of innovation, such as occurred with the spread of multimedia computers in the 1990s, might increase expected future profits and lead firms to increase their planned investment. An economic boom in Western Europe and Japan might lead to a large increase in their expenditure on U.S.-produced goods and services- on U.S. exports. These are all examples of increases in autonomous expenditure.
When autonomous expenditure increases, aggregate expenditure increases and so does equilibrium expenditure and real GDP. But the increase in real GDP is larger than the change in autonomous expenditure. The multiplier is the amount by which a
change in autonomous expenditure is magnified or multiplied to determine the change in equilibrium expenditure and real GDP.
To get the basic idea of the multiplier, we'll work with an example economy in which there are no income taxes and no imports. So we'll first assume that these factors are absent. Then, when you understand the basic idea, we'll bring these factors back into play and see what difference they make to the multiplier.

## The Basic Idea of the Multiplier

Suppose that investment increases. The additional expenditure by businesses means that aggregate expenditure and real GDP increase. The increase in real GDP increases disposable income, and with no income taxes, real GDP and disposable income increase by the same amount. The increase in disposable income brings an increase in consumption expenditure. And the increased consumption expenditure adds even more to aggregate expenditure. Real GDP and disposable income increase further, and so does consumption expenditure. The initial increase in investment brings an even bigger increase in aggregate expenditure because it induces an increase in consumption expenditure. The magnitude of the increase in aggregate expenditure that results from an increase in autonomous expenditure is determined by the multiplier.

The table in Fig. 28.5 sets out an aggregate planned expenditure schedule. Initially, when real GDP is $\$ 15$ trillion, aggregate planned expenditure is $\$ 15.25$ trillion. For each $\$ 1$ trillion increase in real GDP, aggregate planned expenditure increases by $\$ 0.75$ trillion. This aggregate expenditure schedule is shown in the figure as the aggregate expenditure curve $A E_{O}$. Initially, equilibrium expenditure is $\$ 16$ trillion. You can see this equilibrium in row $B$ of the table and in the figure where the curve $A E O$ intersects the $45^{\circ}$ line at the point marked $B$. Now suppose that autonomous expenditure increases by $\$ 0.5$ trillion. What happens to equilibrium expenditure? You can see the answer in Fig. 28.5. When this increase in autonomous expenditure is added to the original aggregate planned expenditure, aggregate planned expenditure increases by $\$ 0.5$ trillion at each level of real GDP. The new aggregate expenditure curve is $A E 1$. The new equilibrium expenditure, highlighted in the table (row $\mathrm{D}^{\prime}$ ), occurs where $A E 1$ intersects the $45^{\circ}$ line and is $\$ 18$ trillion (point $\mathrm{D}^{\prime}$ ). At this real GDP, aggregate planned expenditure equals real GDP.

## Multiplier Effect

In Fig. 28.5, the increase in autonomous expenditure of $\$ 0.5$ trillion increases equilibrium expenditure by $\$ 2$ trillion. That is, the change in autonomous expenditure leads, like a rock singer's electronic equipment, to an amplified change in equilibrium expenditure.
This amplified change is the multiplier effect - equilibrium expenditure increases by more than the increase in autonomous expenditure. The multiplier is greater than 1.
Initially, when autonomous expenditure increases, aggregate planned expenditure exceeds real GDP. As a result, inventories decrease. Firms respond by increasing production so as to restore their inventories to the target level. As production increases, so does real GDP. With a higher level of real GDP, induced expenditure increases. Equilibrium expenditure increases by the sum of the initial increase in autonomous expenditure and the increase in induced expenditure. In this example, equilibrium expenditure increases by $\$ 2$ trillion following the increase in autonomous expenditure of $\$ 0.5$ trillion, so induced expenditure increases by $\$ 1.5$ trillion.
Although we have just analyzed the effects of an increase in autonomous expenditure, this analysis also applies to a decrease in autonomous expenditure. Ifinitially the aggregate expenditure curve is AE1, equilibrium expenditure and real GDP are $\$ 18$ trillion. A decrease in autonomous expenditure of $\$ 0.5$ trillionshifts the aggregate expenditure curve downward by $\$ 0.5$ trillion to $A E_{0}$. Equilibrium expenditure decreases from $\$ 18$ trillion to $\$ 16$ trillion. The decrease in equilibrium expenditure ( $\$ 2$ trillion) is larger than the decrease in autonomous expenditure that brought it about ( $\$ 0.5$ trillion).

## Why Is the Multiplier Effect Greater Than 1?

We've seen that equilibrium expenditure increases by more than the increase in autonomous expenditure.This makes the multiplier greater than 1 . How come? Why does equilibrium expenditure increase by more than the increase in autonomous expenditure?

FIGURE 28.5 The Multiplier


|  | Aggregate planned expenditure |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
| Real GDP <br> $(\boldsymbol{Y})$ | Original <br> $\left(\boldsymbol{A} E_{0}\right)$ | New <br> $\left(A E_{1}\right)$ |  |  |
| (trillions of 2009 dollars) |  |  |  |  |
| $\mathbf{1 5}$ | $A$ | 15.25 | $A^{\prime}$ | 15.75 |
| $\mathbf{1 6}$ | B | $\mathbf{1 6 . 0 0}$ | $B^{\prime}$ | 16.50 |
| 17 | $C$ | 16.75 | $C^{\prime}$ | 17.25 |
| $\mathbf{1 8}$ | $D$ | 17.50 | $\boldsymbol{D}^{\prime}$ | $\mathbf{1 8 . 0 0}$ |
| 19 | $E$ | 18.25 | $E^{\prime}$ | 18.75 |

A $\$ 0.5$ trillion increase in autonomous expenditure shifts the $A E$ curve upward by $\$ 0.5$ trillion from $A E_{0}$ to $A E_{1}$.
Equilibrium expenditure increases by $\$ 2$ trillion from $\$ 16$ trillion to $\$ 18$ trillion. The increase in equilibrium expenditure is 4 times the increase in autonomous expenditure, so the multiplier is 4 .
The multiplier is greater than 1 because induced expenditure increases-an increase in autonomous expenditure induces further increases in expenditure. The NASA space program costs about $\$ 18$ billion a year. This expenditure adds $\$ 18$ billion a year directly to real GDP. But that is not the end of the story.

Astronauts and engineers now have more income, and they spend part of the extra income on goods and services. Real GDP now rises by the initial $\$ 18$ billion plus the extra consumption expenditure induced by the $\$ 18$ billion increase in income. The producers of cars, TVs, vacation cruises, and other goods and services now have increased incomes, and they, in turn, spend part of the increase in their incomes on consumption goods and services. Additional income induces additional expenditure, which creates additional income. How big is the multiplier effect?

## The Size of the Multiplier

Suppose that the economy is in a recession. Profit prospects start to look better, and firms are planning a $\cdot$ large increase in investment. The world economy is also heading toward expansion. The question on everyone's lips is: How strong will the expansion be? This is a hard question to answer, but an important ingredient in the answer is the size of the multiplier.
The multiplier is the amount by which a change in autonomous expenditure is multiplied to deter- mine the change in equilibrium expenditure that itgenerates. To calculate the multiplier, we divide the change in equilibrium expenditure by the change in autonomous expenditure.
Let's calculate the multiplier for the example in Fig. 28.5. Initially, equilibrium expenditure is $\$ 16$ trillion. Then autonomous expenditure increases by $\$ 0.5$ trillion, and equilibrium expenditure increases by $\$ 2$ trillion, to $\$ 18$ trillion. Then

$$
\begin{gathered}
\text { Multiplier }=\frac{\text { Change in equilibrium expenditure }}{\text { Change in autonomous expenditure }} \\
\text { Multiplier }=\frac{\$ 2 \text { trillion }}{\$ 0.5 \text { trillion }}=4
\end{gathered}
$$

## The Multiplier and the Slope of the AE Curve

The magnitude of the multiplier depends on the slope of the $A E$ curve. In Fig. 28.6, the $A E$ curve in part (a) is steeper than the $A E$ curve in part (b), and the multiplier is larger in part (a) than in part (b). To see why, let's do a calculation.
Aggregate expenditure and real GDP change because induced expenditure and autonomous expenditure change. The change in real GDP (,6.Y) equalsthe change in induced expenditure (,6.N) plus thechange in autonomous expenditure (M). That is,

$$
\Delta Y=\Delta N+\Delta A
$$

But the change in induced expenditure is determined by the change in real GDP and the slope of the $A E$ curve. To see why, begin with the fact that the slope of the $A E$ curve equals the "rise," , $6 . N$, divided by the "run," , $6 . Y$ That is,

$$
\text { Slope of } A E \text { curve }=\Delta N \div \Delta Y
$$

So

$$
\Delta N=\text { Slope of } A E \text { curve } \times \Delta Y
$$

Now, use this equation to replace $\Delta N$ in the first equation above to give

$$
\Delta Y=\text { Slope of } A E \text { curve } \times \Delta Y+\Delta A
$$

Now, solve for $\Delta Y$ as

$$
(1-\text { Slope of } A E \text { curve }) \times \Delta Y=\Delta A
$$

Now rearrange the equation to give

$$
\Delta Y=\frac{\Delta A}{1-\text { Slope of } A E \text { curve }}
$$

Finally, divide both sides of this equation by $\Delta A$ to give

$$
\text { Multiplier }=\frac{\Delta Y}{\Delta A}=\frac{1}{1-\text { Slope of } A E \text { curve }}
$$

If we use the example in Fig. 28.5, the slope of the $A E$ curve is 0.75 , so

$$
\text { Multiplier }=\frac{1}{1-0.75}=\frac{1}{0.25}=4
$$

Where there are no income taxes and no imports, the slope of the $A E$ curve equals the marginal propensity to consume (MPC). So

$$
\text { Multiplier }=\frac{1}{1-M P C}
$$

But ( $1-M P C$ ) equals MPS. So another formula is

$$
\text { Multiplier }=\frac{1}{M P S}
$$

Again using the numbers in Fig. 28.5, we have

$$
\text { Multiplier }=\frac{1}{0.25}=4
$$

Because the marginal propensity to save (MPS) is a fraction-a number between 0 and 1 - the multiplier is greater than 1.

FIGURE 28.6 The Multiplier and the Slope of the AE Curve


## Imports and Income Taxes

Imports and income taxes influence the size of the multiplier and make it smaller than it otherwise would be.
To see why imports make the multiplier smaller, think about what happens following an increase in investment. The increase in investment increases real GDP, which in turn increases consumption expenditure. But part of the increase in expenditure is on imported goods and services. Only expenditure on U.S.-produced goods and services increases U.S. real GDP. The larger the marginal propensity to import, the smaller is the change in U.S. real GDP. The Mathematical Note on pp. 732-735 shows the effects of imports and income taxes on the multiplier.
Income taxes also make the multiplier smaller than it otherwise would be. Again, think about what hap- pens following an increase in investment. The increase in investment increases real GDP. Income tax payments increase so disposable income increases by less than the increase in real GDP and consumption expenditure increases by less than it would if taxes had not changed. The larger the income tax rate, the smaller is the change in real GDP.
The marginal propensity to import and the income tax rate together with the marginal propensity to consume determine the multiplier. And their combined influence determines the slope of the $A E$ curve. Over time, the value of the multiplier changes as tax rates change and as the marginal propensity to consume and the marginal propensity to import change. These ongoing changes make the multiplier hard to predict. But they do not change the fundamental fact that an initial change in autonomous expenditure leads to a magnified change in aggregate expenditure and real GDP.

## The Multiplier Process

The multiplier effect isn't a one-shot event. It is a process that plays out over a few months. Figure 28.7 illustrate the multiplier process. Autonomous expenditure increases by $\$ 0.5$ trillion and real GDP increases by $\$ 0.5$ trillion (the green bar in round 1). This increase in real GDP increases induced expenditure in round 2 . With the slope of the $A E$ curveequal to 0.75 , induced expenditure increases by 0.75 times the increase in real GDP, so the increase in real GDP of $\$ 0.5$ trillion induces a further increase in expenditure of $\$ 0.375$ trillion. This change in induced expenditure (the green bar in round 2 ) when added
to the previous increase in expenditure (the bar in round 2 ) increases real GDP by $\$ 0.875$ trillion. The round 2 increase in real GDP induces a round 3 increase in induced expenditure. The process repeats through successive rounds. Each increase in real GDP is 0.75 times the previous increase and eventually real GDP increases by $\$ 2$ trillion.
figure 28.7 The Multiplier Process


Autonomous expenditure increases by $\$ 0.5$ trillion. In round 1 , real GDP increases by the same amount. With the slope of the $A E$ curve equal to 0.75 , each additional dollar of real GDP induces an additional 0.75 of a dollar of induced expenditure. The round 1 increase in real GDP brings an increase in induced expenditure of $\$ 0.375$ trillion in round 2.

At the end of round 2, real GDP has increased by $\$ 0.875$ trillion. The extra $\$ 0.375$ trillion of real GDP in round 2 brings a further increase in induced expenditure of $\$ 0.281$ trillion in round 3.

At the end of round 3, real GDP has increased by $\$ 1.156$ trillion. This process continues with real GDP increasing by ever-smaller amounts. When the process comes to an end, real GDP has increased by a total of $\$ 2$ trillion.

## Business Cycle Turning Points

At business cycle turning points, the economy moves from expansion to recession or from recession to expansion. Economists understand these turning points as seismologists understand earthquakes. They know quite a lot about the forces and mechanisms that produce them, but they can't predict them. The forces that bring business cycle turning points are the swings in autonomous expenditure, such as investment and exports. The multiplier that you've just studied is the mechanism that gives momentum to the economy's new direction.

## ECONOMICS IN ACTION


The aggregate expenditure model and its multiplier were developed during the 1930s by John Maynard Keynes to understand the most traumatic event in economic history, the Great Depression.

In 1929, the U.S. and global economies were booming. U.S. real GDP and real GDP per person had never been higher. By 1933, real GDP had fallen to 73 percent of its 1929 level and more than a quarter of the labor force was unemployed.

The table below shows the GDP numbers and components of aggregate expenditure in 1929 and 1933.

Autonomous expenditure collapsed as investment fell from $\$ 17$ billion to $\$ 3$ billion and exports fell by $\$ 3$ billion. Government expenditure held steady.

|  | 1929 <br> (billions of 1929 dollars) |  |
| :--- | :---: | :---: |
| Induced consumption | 47 | 34 |
| Induced imports | -6 | $\underline{-4}$ |
| Induced expenditure | 30 | 31 |
| Autonomous consumption | 17 | 30 |
| Investment | 10 | 10 |
| Government expenditure | $\underline{6}$ | $\underline{3}$ |
| Exports | $\underline{63}$ | $\underline{46}$ |
| Autonomous expenditure | $\underline{\mathbf{1 0 4}}$ | $\underline{=}$ |
| GDP |  |  |

Source of data: Bureau of Economic Analysis.

The figure uses the $A E$ model to illustrate the Great Depression. In 1929, with autonomous expenditure of $\$ 63$ billion, the $A E$ curve was $A E_{29}$. Equilibrium expenditure and real GDP were $\$ 104$ billion.

By 1933, autonomous expenditure had fallen by $\$ 17$ billion to $\$ 46$ billion and the $A E$ curve had shifted downward to $A E_{33}$. Equilibrium expenditure and real GDP had fallen to $\$ 76$ billion.

The decrease in autonomous expenditure of $\$ 17$ billion brought a decrease in real GDP of $\$ 28$ billion. The multiplier was $\$ 28 / \$ 17=1.6$. The slope of the $A E$ curve is 0.39 - the fall in induced expenditure, $\$ 11$ billion, divided by the fall in real GDP, $\$ 28$ billion. The multiplier formula, $1 /(1-$ Slope of $A E$ curve), delivers a multiplier equal to 1.6 .


Aggregate Expenditure in the Great Depression

## The Multiplier and the Price Level

We have just considered adjustments in spending that occur in the very short run when the price level is fixed. In this time frame, the economy's cobblestones, which are changes in investment andexports, are not smoothed by shock absorbers like those on a limousine. Instead, they are amplified like a rock singer's voice. But these outcomes occur only when the
price level is fixed. We now investigate what happens after a long enough time lapse for the price level to change.

## Adjusting Quantities and Prices

When firms can't keep up with sales and their inventories fall below target, they increase production, but at some point, they raise their prices. Similarly, when firms find unwanted inventories piling up, they decrease production, but eventually they cut their prices. So far, we've studied the macroeconomic con- sequences of firms changing their production levels when their sales change, but we haven't looked at the effects of price changes. When individual firms change their prices, the economy's price level changes. To study the simultaneous determination of real GDP and the price level, we use the $A S-A D$ model, which is explained in Chapter 27. But to understand how aggregate demand adjusts, we need to work out the connection between the $A S-A D m o d e l ~ a n d ~ t h e ~$ aggregate expenditure model thatwe've used in this chapter. The key to understanding the relationship between these two models is the distinction between the aggregate expenditure and aggregate demand and the related distinctionbetween the aggregate expenditure curve and the aggregate demandcurve.

## Aggregate Expenditure and Aggregate Demand

The aggregate expenditure curve is the relationship between the aggregate planned expenditure and real GDP, all other influences on aggregate planned expenditure remaining the same. The aggregate demand curve is the relationship between the aggregate quantity of goods and services demanded andthe price level, all other influences on aggregate demand remaining the same. Let's explore the links between these two relationships.

## Deriving the Aggregate Demand Curve

When the price level changes, aggregate planned expenditure changes and the quantity of real GDP demanded changes. The aggregate demand curve slopes downward. Why? There are two main reasons:

- Wealth effect
- Substitution effects

Wealth Effect Other things remaining the same, the higher the price level, the smaller is the purchasing power of wealth. For example, suppose you have $\$ 100$ in the bank and the price level is 105 . If the price level rises to 125 , your $\$ 100$ buys fewergoods and services. You are less wealthy. With less wealth, you will probably want to try to spend a bit less and save a bit more. The higher the price level, other things remaining the same, the lower is aggregate planned expenditure.

Substitution Effects For a given expectedfuture price level, a rise in the price level today makes cur- rent goods and services more expensive relative to future goods and services and results in a delay inpurchases-an intertemporal substitution. A rise in the U.S. price level, other things remaining the same, makes U.S.-produced goods and services more
expensive relative to foreign-produced goodsand services. As a result, U.S. imports increase and U.S. exports decrease-an internationalsubstitution.
When the price level rises, each of these effects reduces aggregate planned expenditure at each level of real GDP. As a result, when the price level rises, the aggregate expenditure curve shifts downward. A fall in the price level has the opposite effect. When the price levelfalls, the aggregate expenditure curve shifts upward.
Figure 28.8(a) shows the shifts of the $A B$ curve. When the price level is 110 , the aggregate expenditure curve is $A E_{0}$, which intersects the $45^{\circ}$ line at point $B$. Equilibrium expenditure is $\$ 16$ trillion.
If the price level rises to 130 , the aggregate expenditure curve shifts downward to AE1, which intersects the $45^{\circ}$ line at point $A$. Equilibrium expenditure decreases to $\$ 15$ trillion. If the price level falls to 90 , the aggregate expenditure curve shifts upward to $A E_{2}$, which intersects the $45^{\circ}$ line at point C. Equilibrium expenditure increases to $\$ 17$ trillion. We've just seen that when the price level changes, other things remaining the same, the aggregate expenditure curve shifts and the equilibrium expenditure changes. But when the price level changes and other things remain the same, there is a movement along the aggregate demand curve.
Figure 28.8(b) shows the movements along the aggregate demand curve. At a price level of 110, the aggregate quantity of goods and services demanded is $\$ 16$ trillion-point $B$ on the $A D$ curve. If the price level rises to 130 , the aggregate quantity of goods and services demanded decreases to $\$ 15$ trillion and there is a movement up along the aggregate demand curve to point $A$. If the price level falls to 90, the aggregate quantity of goods and services demanded increases to $\$ 17$ trillion and there is a movement down along the aggregate demandcurve to point C.
Each point on the aggregate demand curve corresponds to a point of equilibrium expenditure. Theequilibrium expenditure points $A, B$, and $C$ in Fig.28.8(a) correspond to the points $A, B$, and $C$ on the aggregate demand curve in Fig. 28.8(b).
Changes inAggregate Expenditureand Aggregate Demand When any influence on aggregate planned expenditure other than the price level changes, both the aggregate expenditure curve and the aggregate demand curve shift. For example, an increase in investment or exports increases both aggregate planned expenditure and aggregate demand and shifts both the $A E$ curve and the $A D$ curve. Figure 28.9 illustrates the effect of such an increase.
Initially, the aggregate expenditure curve is $A E_{O}$ in part (a) and the aggregate demand curve is $A D_{O}$ in part (b). The price level is 110, real GDP is $\$ 16$ trillion, and the economy is at point $A$ in both parts of Fig. 28.9. Now suppose that investment increases by $\$ 1$ trillion. At a constant price level of 110, the aggregate expenditure curve shifts upward to $A E 1 \bullet$ This curve intersects the $45^{\circ}$ line at an equilibrium expenditure of $\$ 18$ trillion (point $B$ ). This equilibrium expenditure of $\$ 18$ trillion is the aggregate quantity of goods and services demanded at a price levelof 110, as shown by point B in part (b). Point $B$ lies on a new aggregate demand curve. The aggregate demand curve has shifted rightward to $A D 1$ to pass through point $B$.

FIGURE 28.8 Equilibrium Expenditure and Aggregate Demand


## (b) Aggregate demand

A change in the price level shifts the $A E$ curve and results in a movement along the $A D$ curve. When the price level is 110, the $A E$ curve is $A E_{0}$ and equilibrium expenditure is $\$ 16$ trillion at point $B$. A rise in the price level to 130 shifts the $A E$ curve downward to $A E_{1}$. Equilibrium expenditure decreases to $\$ 15$ trillion at point $A$. A fall in the price level from 110 to 90 shifts the $A E$ curve upward to $A E_{2}$. Equilibrium expenditure increases to $\$ 17$ trillion at point $C$. Points $A, B$, and $C$ on the $A D$ curve in part (b) correspond to the equilibrium expenditure points $A, B$, and $C$ in part (a).

FIGURE 28.9 A Change in
Aggregate Demand


(b) Aggregate demand

An increase in autonomous expenditure increases aggregate demand. The price level is 110 . When the aggregate expenditure curve is $A E_{0}$ in part (a), the aggregate demand curve is $A D_{0}$ in part (b). The economy is at point $A$ in both parts of the figure. An increase in autonomous expenditure shifts the $A E$ curve upward to $A E_{1}$. The new equilibrium expenditure is $\$ 18$ trillion at point $B$ in part (a). Because the quantity of real GDP demanded at a price level of 110 in creases from $\$ 16$ trillion to $\$ 18$ trillion, the $A D$ curve shifts rightward to $A D_{1}$.

But how do we know by how much the $A D$ curve shifts? The multiplier determines the answer. The larger the multiplier, the larger is the shift in the aggregate demand curve that results from a given change in autonomous expenditure. In this example, the multiplier is 2. A $\$ 1$ trillion increase in investment produces a $\$ 2$ trillion increase in the aggregate quantity of goods and services demanded at each price level. That is, a $\$ 1$ trillion increase in autonomous expenditure shifts the aggregate demand curve rightward by $\$ 2$ trillion. A decrease in autonomous expenditure shifts the aggregate expenditure curve downward and shifts the aggregate demand curve leftward. You can see these effects by reversing the change that we've just described. If the economy is initially at point $B$ on the aggregate expenditure curve $A E_{1}$ and on the aggregate demand curve $A D_{1}$, a decrease in autonomous expenditure shifts the aggregate expenditure curve downward to $A E_{0}$. The aggregate quantity of goods and services demanded decreases from \$18 trillion to \$16 trillion, and the aggregate demand curve shifts leftward to $A D_{0}$.
Let's summarize what we have just discovered:
If some factor other than a change in the price level increases autonomous expenditure, then the $A E$ curve shifts upward and the $A D$ curve shifts rightward. The size of the $A D$ curve shift equals the change in autonomous expenditure multiplied by the multiplier.

## Equilibrium Real GDP and the Price Level

In Chapter 27, we learned that aggregate demand and short-run aggregate supply determine equilibrium real GDP and the price level. We've now put aggregate demand under a more powerful microscope and have discovered that a change in investment (or in any component of autonomous expenditure) changes aggregate demand and shifts the aggregate demand curve. The magnitude of the shift depends on the multiplier. But whether a change in autonomous expenditure results ultimately in a change in real GDP, a change in the price level, or a combination of the two depends on aggregate supply. There are two time frames to consider: the short run and the long run. First we'll see what happens in the short run.

An Increase in Aggregate Demand in the Short Run Figure 28.10 describes the economy. Initially, in part (a), the aggregate expenditure curve is $A E_{0}$ and equilibrium expenditure is $\$ 16$ trillion-point $A$. In part(b), aggregate demand is $A D_{0}$ and the short-run aggregate supply curve is SAS. (Chapter 27, pp.689-691, explains the SAS curve.) Equilibrium is at point $A$ in part (b), where the aggregate demand and short-run aggregate supply curves intersect. The price level is 110 , and real GDP is $\$ 16$ trillion.

Now suppose that investment increases by $\$ 1$ trillion. With the price level fixed at 110 , the aggregate expenditure curve shifts upward to $A E_{1}$. Equilibrium expenditure increases to $\$ 18$ trillion-point B in part (a). In part (b), the aggregate demand curve shifts rightward by \$2 trillion, from $A D_{0}$ to $A D_{1}$. How farthe aggregate demand curve shifts is determined by the multiplier when the price level is fixed.
But with this new aggregate demand curve, the price level does not remain fixed. The price level rises, and as it does, the aggregate expenditure curve shifts downward. The short-run equilibrium occurs whenthe aggregate expenditure curve has shifted downward to $A E_{2}$
and the new aggregate demand curve, $A D_{1}$, intersects the short-run aggregate supply curve at point $C$ in both part (a) and part (b). Real GDP is $\$ 17.3$ trillion, and the price level is 123. When price level effects are taken into account, the increase in investment still has a multiplier effect on real GDP, but the multiplier is smaller than it would be if the price level were fixed. The steeper the slope of the short-run aggregate supply curve, the larger is the increase in the price level and the smaller is the multiplier effect on real GDP.

An Increase in Aggregate Demand in the Long Run Figure 28.11 illustrates the long-run effect of an increase in aggregate demand. In the long run, real GDP equals potential GDP and there is full employment. Potential GDP is $\$ 16$ trillion, and the long-run aggregate supply curve is LAS. Initially, the economy is at point $A$ in parts (a) and (b).
Investment increases by $\$ 1$ trillion. In Fig. 28.11, the aggregate expenditure curve shifts to $A E_{1}$ and the aggregate demand curve shifts to $A D_{1}$. With no change in the price level, the economy would move to point B and real GDP would increase to $\$ 18$ trillion. But in the short run, the price level rises to 123 andreal GDP increases to only $\$ 17.3$ trillion. With the higher price level, the $A E$ curve shifts from $A E_{1}$ to $A B_{2}$. The economy is now in a short-run equilibrium at point $C$ in both part (a) and part (b).

Real GDP now exceeds potential GDP. The labor force is more than fully employed, and in the long run, shortages of labor increase the money wage rate. The higher money wage rate increases firms' costs, which decreases short-run aggregate supply and shiftsthe SAS curve leftward to $S A S_{1}$. The price level rises above 123 and real GDP decreases. There is a movement along $A D_{1}$, and the $A B$ curve shifts downward from $A B_{2}$ toward $A B_{0}$. When the money wage rate and the price level have increased by the same percentage, real GDP is again equal to potential GDP and the economy is at point $A^{\prime}$. In the long run, the multiplier is zero.
You are now ready to build on what you've learned about aggregate expenditure fluctuations. We'll study the business cycle and the roles offiscal policy and monetary policy in smoothing the cycle while achieving price stability and sustained economic growth. In Chapter 29 we study the U.S. business cycle and inflation, and in Chapters 30 and 31 we study fiscal policy and monetary policy, respectively. But before you leave the current topic, look at Economics in the News on pp. 692-693 and see the aggregate expenditure model in action in the economy during 2014.

FIGURE 28.10 The Multiplier in the Short Run


## (b) Aggregate demand

An increase in investment shifts the $A E$ curve from $A E_{0}$ to $A E_{1}$ and the $A D$ curve from $A D_{0}$ to $A D_{1}$. The price level rises, and the higher price level shifts the $A E$ curve downward from $A E_{1}$ to $A E_{2}$. The economy moves to point $C$ in both parts. In the short run, when prices are flexible, the multiplier effect is smaller than when the price level is fixed.

FIGURE 28.11 The Multiplier in the Long Run

(b) Aggregate demand

Starting from point $A$, an increase in investment shifts the $A E$ curve to $A E_{1}$ and the $A D$ curve to $A D_{1}$. In the short run at point $C$ the economy is in an above fullemployment equilibrium. In the long run, the money wage rate rises and the SAS curve shifts to $S A S_{1}$. As the price level rises, the $A E$ curve shifts down to $A E_{0}$ and the economy moves to point $A^{\prime}$. In the long run, the multiplier is zero.

## ECONOMIC ANALYSIS

- The BEA news release reports that real GDP increased in the second quarter of 2014 and identifies exports and investment as two sources of expansion. Business inventories also inreased.
- Table 1 shows the real GDP and aggregate expenditure numbers for the first two quarters of 2014 along with the change in the second quarter.
- Figure 1 shows the changes in inventories and real GDP. The two variables fluctuate together but real GDP has larger swings than inventories.
- Figure 2 interprets the data for 2014 using the Keynesian model of equilibrium expenditure.
- In 2014 Q1, the $A E$ curve was $A E_{0}$ and real GDP was $\$ 15.83$ trillion, which we assume to be an expenditure equilibrium.
- The slope of the $A E$ curve is 0.5 (an assumption).
- In Fig. 2(a), an increase in autonomous expenditure shifted the $A E$ curve upward to $A E_{1}$ and aggregate planned expenditure temporarily exceeded real GDP.
- In Fig. 2(b), an unplanned decrease in inventories occurred as real GDP increased toward its second quarter equilibrium.
- When real GDP reached its second quarter equilibrium, unplanned inventory changes had returned to zero.

Table 1 The Components of Aggregate Expenditure

|  | 2014 Q1 | 2014 Q2 | Change |
| :--- | ---: | :---: | ---: |
| Item | (billions of 2009 dollars) |  |  |
| Consumption <br> expenditure | 10,844 | 10,910 | 66 |
| Investment | 2,588 | 2,695 | 107 |
| Government <br> expenditure | 2,869 | 2,879 | 10 |
| Exports | 2,027 | 2,076 | 49 |
| Imports | 2,474 | 2,540 | 66 |
| Real GDP* | $\mathbf{1 5 , 8 3 2}$ | $\mathbf{1 5 , 9 9 4}$ | $\mathbf{1 6 3}$ |
| Change in <br> inventories | 35 | 84 | 49 |

*Chained-dollar real variables are calculated for each expenditure component independently of chained-dollar real GDP and the components don't exactly sum to real GDP.


Figure 1 Inventories and the Change in Real GDP

(a) Convergence to equilibrium expenditure in 2014

(b) Unplanned inventory change in 2014

Figure 2 Equilibrium Expenditure in 2014

## MATHEMATICAL NOTE

## The Algebra of the Keynesian Model

This mathematical note derives formulas for equilibrium expenditure and the multipliers when the price level is fixed. The variables are

- Aggregate planned expenditure, $A E$
- Real GDP, $Y$
- Consumption expenditure, $C$
- Disposable income, $Y D$
- Investment, $I$
- Government expenditure, $G$
- Exports, $X$
- Imports, $M$
- Net taxes, $T$
- Autonomous consumption expenditure, a
- Autonomous taxes, $T_{a}$
- Marginal propensity to consume, $b$
- Marginal propensity to import, $m$
- Marginal tax rate, $t$
- Autonomous expenditure, $A$


## Aggregute Expenditure

Aggregate planned expenditure $(A E)$ is the sum of the planned amounts of consumption expenditure $(C)$, investment $(I)$, government expenditure $(G)$, and exports $(X)$ minus the planned amount of imports ( $M$ ).

$$
A E=C+I+G+X-M
$$

Consumption Function Consumption expenditure $(C)$ depends on disposable income ( $Y D$ ), and we write the consumption function as

$$
C=a+b Y D .
$$

Disposable income $(Y D)$ equals real GDP minus net taxes $(Y-T)$. So if we replace $Y D$ with $(Y-T)$, the consumption function becomes

$$
C=a+b(Y-T)
$$

Net taxes, $T$, equal autonomous taxes (that are independent of income), $T_{a}$, plus induced taxes (that vary with income), $t Y$.

So we can write net taxes as

$$
T=T_{a}+t Y .
$$

Use this last equation to replace $T$ in the consumption function. The consumption function becomes

$$
C=a-b T_{a}+b(1-t) Y .
$$

This equation describes consumption expenditure as a function of real GDP.

Import Function Imports depend on real GDP, and the import function is

$$
M=m Y
$$

Aggregate Expenditure Curve Use the consumption function and the import function to replace $C$ and $M$ in the $A E$ equation. That is,

$$
A E=a-b T_{a}+b(1-t) Y+I+G+X-m Y
$$

Collect the terms that involve $Y$ on the right side of the equation to obtain
$A E=\left(a-b T_{a}+I+G+X\right)+[b(1-t)-m] Y$.
Autonomous expenditure $(A)$ is $\left(a-b T_{a}+I+G+X\right)$, and the slope of the $A E$ curve is $[b(1-t)-m]$. So the equation for the $A E$ curve, which is shown in Fig. 1, is

$$
A E=A+[b(1-t)-m] Y
$$



Figure 1 The AE Curve

## Equilibrium Expenditure

Equilibrium expenditure occurs when aggregate planned expenditure ( $A E$ ) equals real $\operatorname{GDP}(Y)$. That is,

$$
A E=Y
$$

In Fig. 2, the scales of the $x$-axis (real GDP) and the $y$-axis (aggregate planned expenditure) are identical, so the $45^{\circ}$ line shows the points at which aggregate planned expenditure equals real GDP.

Figure 2 shows the point of equilibrium expenditure at the intersection of the $A E$ curve and the $45^{\circ}$ line.

To calculate equilibrium expenditure, solve the equations for the $A E$ curve and the $45^{\circ}$ line for the two unknown quantities $A E$ and $Y$. So starting with

$$
\begin{gathered}
A E=A+[b(1-t)-m] Y \\
A E=Y,
\end{gathered}
$$

replace $A E$ with $Y$ in the $A E$ equation to obtain

$$
Y=A+[b(1-t)-m] Y .
$$

The solution for $Y$ is

$$
Y=\frac{1}{1-[b(1-t)-m]} A .
$$



Figure 2 Equilibrium Expenditure

## The Multiplier

The multiplier equals the change in equilibrium expenditure and real GDP $(Y)$ that results from a change in autonomous expenditure $(A)$ divided by the change in autonomous expenditure.

A change in autonomous expenditure $(\triangle A)$ changes equilibrium expenditure and real GDP by

$$
\begin{gathered}
\Delta Y=\frac{1}{1-[b(1-t)-m]} \Delta A \\
\text { Multiplier }=\frac{1}{1-[b(1-t)-m]}
\end{gathered}
$$

The size of the multiplier depends on the slope of the $A E$ curve, $b(1-t)-m$. The larger the slope, the larger is the multiplier. So the multiplier is larger,

- The greater the marginal propensity to consume (b)
- The smaller the marginal tax rate $(t)$
- The smaller the marginal propensity to import ( $m$ )

An economy with no imports and no income taxes has $m=0$ and $t=0$. In this special case, the multiplier equals $1 /(1-b)$. If $b$ is 0.75 , then the multiplier is 4 , as shown in Fig. 3.

In an economy with imports and income taxes, if $b=0.75, t=0.2$, and $m=0.1$, the multiplier equals 1 divided by $[1-0.75(1-0.2)-0.1]$, which equals 2. Make up some more examples to show the effects of $b, t$, and $m$ on the multiplier.


Figure 3 The Multiplier

## Government Expenditure Multiplier

The government expenditure multiplier equals the change in equilibrium expenditure and real GDP $(Y)$ that results from a change in government expenditure $(G)$ divided by the change in government expenditure. Because autonomous expenditure is equal to

$$
A=a-b T_{a}+I+G+X
$$

the change in autonomous expenditure equals the change in government expenditure. That is,

$$
\Delta A=\Delta G
$$

You can see from the solution for equilibrium expenditure $Y$ that

$$
\Delta Y=\frac{1}{1-[b(1-t)-m]} \Delta G
$$

The government expenditure multiplier equals

$$
\frac{1}{1-[b(1-t)-m]}
$$

In an economy in which $t=0$ and $m=0$, the government expenditure multiplier is $1 /(1-b)$. With $b=0.75$, the government expenditure multiplier is 4 , as Fig. 4 shows. Make up some examples and use the above formula to show how $b, m$, and $t$ influence the government expenditure multiplier.


Figure 4 Government Expenditure Multiplier

## Autonomous Tax Multiplier

The autonomous tax multiplier equals the change in equilibrium expenditure and real GDP $(Y)$ that results from a change in autonomous taxes $\left(T_{a}\right)$ divided by the change in autonomous taxes. Because autonomous expenditure is equal to

$$
A=a-b T_{a}+I+G+X
$$

the change in autonomous expenditure equals minus $b$ multiplied by the change in autonomous taxes. That is,

$$
\Delta A=-b \Delta T_{a}
$$

You can see from the solution for equilibrium expenditure $Y$ that

$$
\Delta Y=\frac{-b}{1-[b(1-t)-m]} \Delta T_{a}
$$

The autonomous tax multiplier equals

$$
\frac{-b}{1-[b(1-t)-m]}
$$

In an economy in which $t=0$ and $m=0$, the autonomous tax multiplier is $-b /(1-b)$. In this special case, with $b=0.75$, the autonomous tax multiplier equals -3 , as Fig. 5 shows. Make up some examples and use the above formula to show how $b$, $m$, and $t$ influence the autonomous tax multiplier.


Figure 5 Autonomous Tax Multiplier

## Balanced Budget Multiplier

The balanced budget multiplier equals the change in equilibrium expenditure and real GDP $(Y)$ that results from equal changes in government expenditure and lump-sum taxes divided by the change in government expenditure. Because government expenditure and autonomous taxes change by the same amount, the budget balance does not change.

The change in equilibrium expenditure that results from the change in government expenditure is

$$
\Delta Y=\frac{1}{1-[b(1-t)-m]} \Delta G
$$

And the change in equilibrium expenditure that results from the change in autonomous taxes is

$$
\Delta Y=\frac{-b}{1-[b(1-t)-m]} \Delta T_{a}
$$

So the change in equilibrium expenditure resulting from the changes in government expenditure and autonomous taxes is

$$
\begin{aligned}
\Delta Y= & \frac{1}{1-[b(1-t)-m]} \Delta G+ \\
& \frac{-b}{1-[b(1-t)-m]} \Delta T_{a}
\end{aligned}
$$

Notice that

$$
\frac{1}{1-[b(1-t)-m]}
$$

is common to both terms on the right side. So we can rewrite the equation as

$$
\Delta Y=\frac{1}{1-[b(1-t)-m]}\left(\Delta G-b \Delta T_{a}\right)
$$

The $A E$ curve shifts upward by $\Delta G-b \Delta T_{a}$ as shown in Fig. 6.

But the change in government expenditure equals the change in autonomous taxes. That is,

$$
\Delta G=\Delta T_{a}
$$

So we can write the equation as

$$
\Delta Y=\frac{1-b}{1-[b(1-t)-m]} \Delta G
$$

The balanced budget multiplier equals

$$
\frac{1-b}{1-[b(1-t)-m]}
$$

In an economy in which $t=0$ and $m=0$, the balanced budget multiplier is $(1-b) /(1-b)$, which equals 1, as Fig. 6 shows. Make up some examples and use the above formula to show how $b, m$, and $t$ influence the balanced budget multiplier.


Figure 6 Balanced Budget Multiplier

## CHAPTER 29: The Business Cycle, Inflation, and Deflation

After studying this chapter, you will be able to:

- Explain how aggregate demand shocks and aggregate supply shocks create the business cycle
- Explain how demand-pull and cost-push forces bring cycles in inflation and output
- Explain the causes and consequences of deflation
- Explain the short-run and long-run tradeoff between inflation and unemployment

We fear deflation because it brings stagnant incomes and high unemployment. And we worry about inflation because it raises our cost of living. We want low inflation, low unemployment, and rapid incomegrowth. But can we have all these things at the same time? Or do we face a tradeoff among them? As this chapter explains, we face a tradeoff in the short run but not in the long run.
At the end of the chapter, in Economics in the News, we examine a stagnating European economy and the lessons it holds for the United States and other countries.

## The Business Cycle

The business cycle is easy to describe but hard toexplain and the next peak or trough is impossible to predict. We'll look at two approaches to understanding the business cycle:

- Mainstream business cycle theory
- Real business cycle theory


## Mainstream Business Cycle Theory

The mainstream business cycle theory is that potential GDP grows at a steady rate while aggregate demand grows at a fluctuating rate. Because the money wage rate is sticky, if aggregate demand grows faster than potential GDP, real GDP moves above potential GDP and an inflationary gapemerges. And if aggregate demand grows slower than potential GDP, real GDP moves below potential GDP and a recessionary gap emerges. Ifaggregate demand decreases, real GDP also decreases in a recession.
Figure 29.1 illustrates this business cycle theory. Initially, potential GDP is $\$ 13$ trillion. The long-runaggregate supply curve is $L A S_{0}$, the aggregate demand curve is $A D_{0}$, and the price level is 100. The economy is at full employment at point $A$.
An expansion occurs when potential GDP increases and the LAS curve shifts rightward to $\mathrm{LAS}_{1}$. During an expansion, aggregate demandalso increases, and usually by more than potential GDP, so the price level rises. Assume that in the current expansion, the price level is expected to rise to 110 and the money wage rate has been set based on that expectation. The short-run aggregate supply curve is $S A S_{1}$.
If aggregate demand increases to $A D_{1}$, real GDP increases from $\$ 13$ trillion to $\$ 16$ trillion, the new level of potential GDP, and the price level rises, as expected, to 110. The economy remains at full employment but now at point $B$. If aggregate demand increases more slowly to $A D_{2}$; real GDP grows by less than potential GDP and the) economy moves to point C, with real GDP at $\$ 15.5$ trillion and the price level at 107. Real GDP growth is slower and inflation
is lower than expected. If aggregate demand increases more quickly to $A D_{3}$, real GDP grows by more than potential GDPand the economy moves to point $D$, with real GDP at $\$ 16.5$ trillion and the price level at 113, real GDP growth is faster and inflation is higher than expected. Growth, inflation, and the business cycle arise from the relentless increases in potential GDP, faster (on average) increases in aggregate demand, and fluctuations in the pace of aggregate demand growth.
This mainstream theory comes in a number of special forms that differ regarding the source offluctuations in aggregate demand growth and the source of money wage stickiness.

FIGURE 29.1 The Mainstream Business Cycle Theory


In a business cycle expansion, potential GDP increase) and the LAS curve shifts rightward from $L A S_{0}$ to $L A S_{1}$.Agreater than expected increase in aggregate demand brings inflation. If the aggregate demand curve shifts to $A D_{1}$ : economy remains at full employment. If the aggregate demand curve shifts to $A D_{2}$, a recessionary gap arises. If the aggregate demand curve shifts to $A D_{3}$ an inflationary gap arises.

Keynesian Cycle Theory In Keynesian cycle theory, fluctuations in investment driven by fluctuations in business confidence - summarized by the phrase "animal spirits"—are the main source of fluctuations in aggregate demand.

Monetarist Cycle Theory In monetarist cycle theory, fluctuations in both investment and consumption expenditure, driven by fluctuations in the growth rate of the quantity of money, are the main source of fluctuations in aggregate demand.

Both the Keynesian and monetarist cycle theories simply assume that the money wage rate is rigid and don't explain that rigidity.
Two newer theories seek to explain money wage rate rigidity and to be more careful about working out its consequences.

New Classical Cycle Theory In new classical cycle theory, the rational expectation of the price level, which is determined by potential GDP and expected aggregate demand, determines the money wage rate and the position of the SAS curve. In this theory, only unexpected fluctuations in aggregate demand bring fluctuations in real GDP around potential GDP.

New Keynesian Cycle Theory The new Keynesian cycle theory emphasizes the fact that today's money wage rates were negotiated at many past dates, which means that past rational expectations of the curref!.tprice level influences the money wage rate and the position of the SAS curve. In this theory, both unexpected and currently expected fluctuations in aggregate demand bring fluctuations in real GDP around potential GDP.
The mainstream cycle theories don't rule out the possibility that aggregate supply shocks might occur. An oil price rise, a widespread drought, a major hurricane, or another natural disaster, could, for ex- ample, bring a recession. But supply shocks are not the normal source of fluctuations in the mainstream theories. In contrast, real business cycle theory puts supply shocks at centerstage.

## Real Business Cycle Theory

The newest theory of the business cycle, known as real business cycle theory (or RBC theory), regards random fluctuations in productivity as the mainsource of economic fluctuations. These productivity fluctuations are assumed to result mainly from fluctuations in the pace of technological change, but they might also have other sources, such as international disturbances, climate fluctuations, or natural disasters. The origins of RBC theory can be traced to the rational expectations revolution set off by Robert E. Lucas, Jr., but the first demonstrations of the power of this theory were given by Edward Prescott and Finn Kydland and by John Long and Charles Plosser. Today, RBC theory is part of a broad research agenda called dynamic general equilibrium analysis, and hundreds of young macroeconomists do research on this topic.
We'll explore RBC theory by looking first at its impulse and then at the mechanism that converts that impulse into a cycle in real GDP.

The RBC Impulse The impulse in RBC theory is the growth rate of productivity that results from techno-logical change. RBC theorists believe this impulse tobe generated mainly by the process of research and development that leads to the creation and use of new technologies (see Economics inAction).
The pace of technological change and productivity growth is not constant. Sometimes productivity growth speeds up, sometimes it slows, and occasion- ally it evenfalls-labor and capital become less productive, on average. A period of rapid productivity growth brings a business cycle expansion, and a slow-down or fall in productivity triggers a recession.

It is easy to understand why technological change brings productivity growth. But how does it decrease productivity? All technological change eventually increases productivity. But if initially, technological change makes a sufficient amount of existing capital-especially human capital-obsolete, productivity can temporarily fall. At such a time, more jobs are destroyed than created and more businesses fail than start up.

The RBC Mechanism Two effects follow from a change in productivity that sparks an expansion or acontraction: Investment demand changes and the demand for labor changes. We'll study these effects and their consequences during a recession. In an expansion, they work in the direction opposite to what is described here.
Technological change makes some existing capital obsolete and temporarily decreases productivity.
Firms expect the future profits to fall and see their labor productivity falling. With lower profit expectations, they cut back their purchases of new capital, and with lower labor productivity, they plan to lay off some workers. So the initial effect of a temporary fall in productivity is a decrease in investment demand and a decrease in the demand for labor. Figure 29.2 illustrates these two initial effects of a decrease in productivity. Part (a) shows the effects of a decrease in investment demand in the loanable funds market. The demand for loanable funds curve is DLF and the supply of loanable funds curve is SLF (both of which are explained in Chapter 24, pp. 613-615). Initially, the demand for loanable fundscurve is $D L F O$ and the equilibrium quantity of funds is $\$ 2$ trillion at a real interest rate of 6 percent a year.
A decrease in productivity decreases investment demand, and the demand for loanable funds curve shifts leftward from $D L F_{o}$ to $D L F_{1}$. The real interest rate falls to 4 percent a year, and the equilibrium quantity of loanable funds decreases to $\$ 1.7$ trillion. Figure 29.2(b) shows the demand for labor and supply of labor (which are explained in Chapter 23, pp. 584-585). Initially, the demand for labor curve is $L D_{0}$, the supply of labor curve is $L S_{0}$, and equilibrium employment is 200 billion hours a year at a real wage rate of $\$ 35$ an hour. The decrease in productivity decreases the demand for labor, and the demand for labor curve shifts leftward from $L D_{0}$ to $L D_{1}$.
Before we can determine the new level of employment and real wage rate, we need to look at a ripple effect - the key effect in RBC theory.

The Key Decision: When to Work? According to RBC theory, people decide when to work by doing a cost-benefit calculation. They compare the return from working in the current period with the expected return from working in a later period. You make such a comparison every day in school. Suppose your goal in this course is to get an A. To achieve this goal, you work hard most of the time. But during the few days before the midterm and final exams, you work especially hard. Why? Because you believe that the return from studying close to the exam is greater than the return from studying when the exam is a long time away. So during the term, you take time off for the movies and other leisure pursuits, but at exam time, you study every evening and weekend. RBC theory says that workers behave like you.

They work fewer hours, sometimes zero hours, when the real wage rate is temporarily low, and they work more hours when the real wage rate is temporarily high. But to properly compare the current wage rate with the expected future wage rate, workers must use the real interest rate. If the real interest rate is 6 per- cent a year, a real wage of $\$ 1$ an hour earned this week will become $\$ 1.06$ a year from now. If the real wage rate is expected to be $\$ 1.05$ an hour next year, today's real wage of $\$ 1$ looks good. By working longer hours now and shorter hours a year from now, a person can get a 1 percent higher real wage. But suppose the real interest rate is 4 percent a year. In this case, $\$ 1$ earned now is worth $\$ 1.04$ next year.
Working fewer hours now and more next year is the way to get a 1 percent higher real wage.

## ECONOMICS IN ACTION

## The Real Business Cycle Impuise

To isolate the RBC impulse, economists measure the change in the combined productivity of capital and labor-called total factor productivity. The figure shows the RBC impulse for the United States from 1963 through 2013.

You can see that the productivity growth rate fluctuations are not directly correlated with real GDP fluctuations. Their influence on real GDP growth is spread out over time.

You can also see that the fluctuations in real GDP growth have wider swings than those of productivity growth.

Real business cycle theory explains these facts.


So the when-to-work decision depends on the real interest rate. The lower the real interest rate, other things remaining the same, the smaller is the supply of labor today. Many economists believe this intertemporal substitution effect to be of negligible size. RBC theorists believe that the effect is large, and it is the key feature of the RBC mechanism.

FIGURE 29.2 Loanable Funds and Labor Markets in a Real Business Cycle

(a) Loanable funds and interest rate

In part (a), the supply of loanable funds SLF and the initial demand for loanable funds $D L F_{0}$ determine the real interest rate at 6 percent a year. In part (b), the initial demand for labor $L D_{0}$ and the supply of labor $L S_{0}$ determine the real wage rate at $\$ 35$ an hour and employment at 200 billion hours. A technological change temporarily decreases productivity, and both the demand for loanable funds and the

(b) Labor and wage rate
demand for labor decrease. The two demand curves shift leftward to $D L F_{1}$ and $L D_{1}$. In part (a), the real interest rate falls to 4 percent a year. In part (b), the fall in the real interest rate decreases the supply of labor (the when-to-work decision) and the supply of labor curve shifts leftward to $L S_{1}$. Employment decreases to 195 billion hours, and the real wage rate falls to $\$ 34.50$ an hour. A recession is under way.

You saw in Fig. 29.2(a) that the decrease in the demand for loanable funds lowers the real interest rate. This fall in the real interest rate lowers the return tocurrent work and decreases the supply of labor.
In Fig. 29.2(b), the labor supply curve shifts leftward to $L S_{1}$. The effect of the decrease in productivity on the demand for labor is larger than theeffect of the fall in the real interest rate on the supply of labor. That is, the demand curve shifts farther leftward than does the supply curve. As a result, the real wage rate falls to $\$ 34.50$ an hour and employment decreases to 195 billion hours. A recession has begun and is intensifying.

What happened to Money? The name realbusiness cycle theory is no accident. It reflects the central prediction of the theory. Real things, not nominal or monetary things, cause the business cycle. If the quantity of money changes, aggregate demand changes. But if there is no real change - with no change in the use of resources and no change in potential GDP - the change in the quantity of money changes only the price level. In RBC theory, this outcome occurs because the aggregate supply curve is the LAS curve, which pins real GDP down at potential GDP, so when demand aggregate changes, only the price level changes.

Cycles and Growth The shock that drives the business cycle of RBC theory is the same as the force that generates economic growth: technological change.

On average, as technology advances, productivity grows; but as you saw in Economics in Action on pg. 744, it grows at an uneven pace. Economic growth arises from the upward trend in productivity growth and, according to RBC theory, the mostly positive but occasionally negative higher frequency shocks to productivity bring the business cycle.

Criticisms and Defenses of RBC Theory The three main criticisms of RBC theory are that

1. The money wage rate is sticky, and to assume otherwise is at odds with a clear fact.
2. Intertemporal substitution is too weak a force to account for large fluctuations in labor supply and employment with small real wage rate changes.

Productivity shocks are as likely to be caused by changes in aggregate demand as by technological change.
If aggregate demand fluctuations cause the fluctuations in productivity, then the traditional aggregate demand theories are needed to explain them.
Fluctuations in productivity do not cause the business cycle but are caused by it! Building on this theme, the critics point out that the so-called productivity fluctuations that growth accounting measures are correlated with changes in the growth rate of money and other indicators of changes in aggregate demand.
The defenders of RBC theory claim that the theory explains the macroeconomic facts about the business cycle and is consistent with the facts abouteconomic growth. In effect, a single theory explains both growth and the business cycle. The growth accounting exercise that explains slowly changing trends also explains the more frequent business cycleswings. Its defenders also claim that RBC theory is consistent with a wide range of microeconomic evidence about labor supply decisions, labor demand and investment demand decisions, and information on the distribution of income between labor and capital.

## Inflation Cycles

Inthe long run, inflation is a monetary phenomenon. It occurs if the quantity of money grows faster than potential GDP. But in the short run, many factors can start inflation, and real GDP and the price level interact. To study these interactions, we distinguish between two sources of inflation:

- Demand-pull inflation
- Cost-push inflation


## Demand-Pull Inflation

An inflation that starts because aggregate demand increases is called demand-pull inflation. Demand- pull inflation can be kicked off by any of the factors that change aggregate demand. Examples are a cut in the interest rate, an increase in the quantity of money, an increase in government expenditure, a tax cut, an increase in exports, or an increase in investment stimulated by an increase in expected future profits.

Initial Effect of an Increase in Aggregate Demand Suppose that last year the price level was 110 and realGDP was $\$ 16$ trillion. Potential GDP was also $\$ 16$ trillion. Figure 29.3(a) illustrates this situation. The aggregate demand curve is $A D_{0}$, the short-run aggregate supply curve is SASO, and the long-run aggregate supply curve is LAS.
Now suppose that the Fed cuts the interest rate. The quantity of money increases and the aggregate demand curve shifts from $A D_{0}$ to $A D_{1}$. With no change in potential GDP and no change in the moneywage rate, the long-run aggregate supply curve and the short-run aggregate supply curve remain at $L A S$ and $S A S_{0}$, respectively.
The price level and real GDP are determined at the point where the aggregate demand curve $A D_{1}$ intersects the short-run aggregate supply curve. The price level rises to 113 and real GDP increases above potential GDP to $\$ 16.5$ trillion. Unemployment falls below its natural rate. The economy is at an above full-employment equilibrium and there is an inflationary gap. The next step in the unfolding story is a rise in the money wage rate.

FIGURE 29.3 A Demand-Pull Rise in the Price Level

(a) Initial effect

In part (a), the aggregate demand curve is $A D_{0}$, the short-run aggregate supply curve is $S A S_{0}$, and the long-run aggregate supply curve is LAS. The price level is 110 , and real GDP is $\$ 16$ trillion, which equals potential GDP. Aggregate demand increases to $A D_{1}$. The price level rises to 113 , and real GDP increases to $\$ 16.5$ trillion.

(b) The money wage rate adjusts

In part (b), starting from the above full-employment equilibrium, the money wage rate begins to rise and the shoritrun aggregate supply curve shifts leffward toward $S A S_{1}$. The price level rises further, and real GDP returns to potential GDP.

Money Wage Rate Response Real GDP cannot remain above potential GDP forever. With unemployment below its natural rate, there is a shortage oflabor. In this situation, the money wage rate begins to rise. As it does so, short-run aggregate supplydecreases and the SAS curve starts to shift leftward.
The price level rises further, and real GDP begins to decrease.
With no further change in aggregate demand- that is, the aggregate demand curve remains at $A D_{1}$-this process ends when the short-run aggregate supply curve has shifted
to $S A S_{1 \text { in }}$ Fig. 29.3(b). At this time, the price level has increased to 121 and real GDP has returned to potential GDP of $\$ 16$ trillion, the level at which it started.

A Demand-Pull Inflation Process The events that we've just described bring a one-time rise in the price level, not an inflation. For inflation to proceed, aggregate demand must persistently increase.
The only way in which aggregate demand can persistently increase is if the quantity of money persistently increases. Suppose the government has a budget deficit that it finances by selling bonds.
Also suppose that the Fed buys some of these bonds. When the Fed buys bonds, it creates more money. In this situation, aggregate demand increases year after year. The aggregate demand curve keeps shifting rightward. This persistent increase in aggregate demand puts continual upward pressure on the price level. The economy now experiences demand-pull inflation.
Figure 29.4 illustrates the process of demand-pull inflation. The starting point is the same as that shown in Fig. 29.3. The aggregate demand curve is $A D_{0}$, the short-run aggregate supply curve is $S A S_{0}$, and the long-run aggregate supply curve is LAS. Real GDPis $\$ 16$ trillion, and the price level is 110 . Aggregatedemand increases, shifting the aggregate demand curve to $A D_{2}$. Real GDP increases to $\$ 16.5$ trillion, and the price level rises to 113 . The economy is at an above full-employment equilibrium. There is a shortage of labor, and the money wage rate rises. The short-run aggregate supply curve shifts to $S A S_{2}$ Theprice level rises to 121, and real GDP returns to potential GDP.
But the Fed increases the quantity of money again, and aggregate demand continues to increase. The aggregate demand curve shifts rightward to $A D_{2}$. The price level rises further to 125 , and real GDP again exceeds potential GDP at $\$ 16.5$ trillion. Yet again, the money wage rate rises and decreases short-run aggregate supply. The SAS curve shifts to SAS2, and the price level rises further, to 133. As the quantity of: money continues to grow, aggregate demandin- creases and the price level rises in an ongoing demand-pull inflation process.
The process you have just studied generates inflation-a persistently rising price level.

Demand-Pull Inflation in Kalamazoo Youmay better understand the inflation process that we've just described by considering what is going on in an individual part of the economy, such as a Kalamazoo soda-bottling plant. Initially, when aggregate demandincreases, the demand for soda increases and the price, of soda rises. Faced with a higher price, the soda plant"works overtime and increases production. Conditions are good for workers in Kalamazoo, and the soda factory finds it hard to hang on to its best people. To doso, it offers a higher money wage rate. As the wage rate rises, so do the soda factory's costs.
What happens next depends on aggregate demand. If aggregate demand remains constant, the firm's costs increase but the price of soda does not increase as quickly as its costs. In this case, the firm cuts production. Eventually, the money wage rate and costs increase by the same percentage as the rise in the price of soda. In real terms, the soda factory is in the same situation as it was initially. It produces the same amount of soda and employs the same amount of labor as before the increase in demand.

But if aggregate demand continues to increase, so does the demand for soda and the price of soda rises at the same rate as wages. The soda factory continues to operate at above full employment and there is a persistent shortage of labor. Prices and wages chase each other upward in a demand-pull inflation spiral.

FIGURE 29.4 A Demand-Pull Inflation Spiral


Each time the quantity of money increases, aggregate demand increases and the aggregate demand curve shifts rightward from $A D_{0}$ to $A D_{1}$ to $A D_{2}$, and so on. Each time real GDP increases above potential GDP, the money wage rate rises and the short-run aggregate supply curve shifts leftward from $S A S_{0}$ to $S A S_{1}$ to $S A S_{2}$, and so on. The price level rises from 110 to $113,121,125,133$, and so on. There is a demand-pull inflation spiral. Real GDP fluctuates between $\$ 16$ trillion and $\$ 16.5$ trillion.

Demand-Pull Inflation in the United States A demand-pull inflation like the one you've just studied occurred in the United States during the late 1960s. In 1960, inflation was a moderate 2 percent a year, but its rate increased slowly to 3 percent by 1966. Then, in 1967, a large increase in government expenditure on the Vietnam War and an increase in spending on social programs, together with an increase in the growth rate of the quantity of money, increased aggregate demand more quickly. Consequently, the rightward shift of the aggregate demand curve accelerated and the price level increased more quickly. Real GDP moved above potential GDP, and the unemployment rate fell below its natural rate. With unemployment below its natural rate, the money wage rate started to rise more quickly and the short-run aggregate supply curve shifted leftward.

The Fed responded with a further increase in the money growth rate, and a demand-pull inflation spiral unfolded. By 1970, the inflation rate had reached 5 percent ayear. For the next few years, aggregate demand grew even more quickly and the inflation rate kept rising. By 1974, the inflation rate had reached 11 percent a year. Next, let's see how shocks to aggregate supply can create cost-push inflation.

## Cost-Pull Inflation

An inflation that is kicked off by an increase in costsis called cost-push inflation. The two main sources of cost increases are

1. An increase in the money wage rate
2. An increase in the money prices of raw materials

At a given price level, the higher the cost of production, the smaller is the amount that firms are willing to produce. So if the money wage rate rises or if the prices of raw materials (for example, oil) rise, firms decrease their supply of goods and services. Aggregate supply decreases, and the short-run aggregate supply curve shifts leftward. Let's trace the effects of such a decrease in short-run aggregate sup- ply on the price level and real GDP.

Initial Effectofa Decrease in Aggregate Supply Suppose that last year the price level was 110 and realGDP was $\$ 16$ trillion. Potential real GDP was also $\$ 16$ trillion. Figure 29.5(a) illustrates this situation. The aggregate demand curve was $A D_{0}$, the short-run aggregate supply curve was $S A S O$, and the long- run aggregate supply curve was $L A S$. In the current year, the world's oil producers form a price-fixing organization that strengthens their market power and increases the relative price of oil. They raise the price of oil, and this action decreases short-run aggregate supply. The short-run aggregate supply curve shifts leftward to $S A S_{1}$. The price level rises to 117 , and real GDP decreases to $\$ 15.5$ trillion. The economy is at a below full-employment equilibrium and there is a recessionarygap. This event is a one-time rise in the price level. Itis not inflation. In fact, a supply shock on its own can- not cause inflation. Something more must happen to enable a one-time supply shock, which causes a one time rise in the price level, to be converted into a process of ongoing inflation. The quantity of moneymust persistently increase. Sometimes it does increase, as you will now see.

Aggregate Demand Response When real GDP decreases, unemployment rises above its natural rate. In such a situation, there is often an outcry of concern and a call for action to restore full employment. Suppose that the Fed cuts the interest rate and increases the quantity of money. Aggregate demand increases. In Fig. 29.5(b), the aggregate demand curve shifts rightward to $A D_{1}$ and full employment is restored. But the price level rises further to 121.

FIGURE 29.5 A Cost-Push Rise in the Price Level

(a) Initial cost push

Initially, the aggregate demand curve is $A D_{0}$, the short-run aggregate supply curve is $S A S_{0}$, and the long-run aggregate supply curve is $L A S$. A decrease in aggregate supply (for example, resulting from a rise in the world price of oil) shifts the short-run aggregate supply curve to $S A S_{1}$. The economy moves to the point where the short-run aggregate supply curve $S A S_{1}$ intersects the aggregate demand curve

(b) The Fed responds
$A D_{0}$. The price level rises to 117, and real GDP decreases to $\$ 15.5$ trillion.

In part (b), if the Fed responds by increasing aggregate demand to restore full employment, the aggregate demand curve shifts rightward to $A D_{1}$. The economy returns to full employment, but the price level rises further to 121 .

A Cost-Push Inflation Process The oil producers now see the prices of everything they buy increasing, so oil producers increase the price of oil again to restore itsnew high relative price. Figure 29.6 continues the story. The short-run aggregate supply curve now shifts to SAS2. The price level rises and real GDP decreases.
The price level rises further, to 129, and real GDP decreases to $\$ 15.5$ trillion.
Unemployment increases above its natural rate. If the Fed respondsyet again with an increase in the quantity of money, aggregate demand increases and the aggregatedemand curve shifts to $A D_{2}$. The price level rises even higher-to 133-and full employment is again restored. A cost-push inflation spiral results. The combination of a rising price level and decreasing real GDP is called stagflation.
You can see that the Fed has a dilemma. Ifit does not respond when producers raise the oil price, the economy remains below full employment. If the Fed increases the quantity of money to restore full employment, it invites another oil price hike that will call forth yet a further increase in the quantity of money.
If the Fed responds to each oil price hike by increasing the quantity of money, inflation will rage along at a rate decided by oil producers. But if the Fed keeps the lid on money growth, the economy remains below fullemployment.

Cost-Push Inflation in Kalamazoo What is going on in the Kalamazoo soda-bottling plant when the economy is experiencing cost-push inflation?

When the oil price increases, so do the costs of bottling soda. These higher costs decrease the supplyof soda, increasing its price and decreasing the quantity produced. The soda plant lays off some workers.
This situation persists until either the Fed increases aggregate demand or the price of oil falls. If the Fed increases aggregate demand, the demand for soda increases and so does its price. The higher price of soda brings higher profits, and the bottling plant increases its production. The soda factory rehires the laid-off workers.

FIGURE 29.6 A Cost-Push Inflation Spiral


Each time a cost increase occurs, the short-run aggregate supply curve shifts leftward from $S A S_{0}$ to $S A S_{1}$ to $S A S_{2}$, and so on. Each time real GDP decreases below potential GDP, the Fed increases the quantity of money and the aggregate demand curve shifts rightward from $A D_{0}$ to $A D_{1}$ to $A D_{2}$, and so on. The price level rises from 110 to 117 , 121, 129, 133, and so on. There is a cost-push inflation spiral. Real GDP fluctuates between $\$ 16$ trillion and $\$ 15.5$ trillion.

Cost-Push Inflation in the United States A cost-push inflation like the one you've just studied occurred in the United States during the 1970s. It began in 1974 when the Organization of the Petroleum ExportingCountries (OPEC) raised the price of oil fourfold. The higher oil price decreased aggregate supply, which caused the price level to rise more quickly and real GDP to shrink. The Fed then faced a dilemma: Would it increase the quantity of money and accommodate the cost-push forces, or would it keep aggregate demand growth in check by limiting money growth? In 1975, 1976, and 1977, the Fed repeatedly allowed the quantity of money to grow quickly and inflation proceeded at a rapid rate. In 1979 and 1980, OPEC was again able to push oil prices higher. On that occasion, the Fed
decided not to respond to the oil price hike with an increase in the quantity of money. The result was a recession but also, eventually, a fall in inflation.

## Expected Inflation

If inflation is expected, the fluctuations in real GDPthat accompany demand-pull and costpush inflation that you've just studied don't occur. Instead, inflation proceeds as it does in the long run, with real GDP equal to potential GDP and unemployment at its natural rate. Figure 29.7 explains why.
Suppose that last year the aggregate demand curve was $A D_{0}$ the aggregate supply curve was $S A S_{0}$, and the long-run aggregate supply curve was LAS. The price level was 110, and real GDP was $\$ 16$ trillion, which is also potential GDP.
To keep things as simple as possible, suppose that potential GDP does not change, so the LAS curve doesn't shift. Also suppose that aggregate demand is expected to increase to $A D_{1}$. In anticipation of this increase in aggregate demand, the money wage rate rises and the short-run aggregate supply curve shifts leftward. If the money wage rate rises by the same percentage as the price level is expected to rise, the short-run aggregate supply curve for next year is $S A S_{1}$.
If aggregate demand turns out to be the same as expected, the aggregate demand curve is $A D_{1}$. The short-run aggregate supply curve, $S A S_{1}$, and AD1 determine the actual price level at 121. Between last year and this year, the price level increased from 110to 121 and the economy experienced an inflation rateequal to that expected. If this inflation is ongoing, aggregate demand increases (as expected) in the following year and the aggregate demand curve shifts to $A D_{2}$. The money wage rate rises to reflect the expected inflation, and the short-run aggregate supply curve shifts to $S A S_{2}$. The price level rises, as expected, to 133.

What caused this inflation? The immediate answer is that because people expected inflation, the moneywage rate increased and the price level increased. Butthe expectation was correct. Aggregate demand was expected to increase, and it did increase. It is the actual and expected increase in aggregate demand that caused the inflation.
An expected inflation at full employment is exactly the process that the quantity theory of money predicts. To review the quantity theory of money, see Chapter 25, pp. 646-647. This broader account of the inflation process and its short-run effects shows why the quantity theory of money doesn't explain the fluctuations ininflation.
The economy follows the course described in Fig. 29.7, but as predicted by the quantity theory, only if aggregate demand growth is forecasted correctly.

## Forecasting Inflation

To anticipate inflation, people must forecast it. Some economists who work for macroeconomic forecasting agencies, banks, insurance companies, labor unions, and large corporations specialize in inflation forecasting. The best forecast available is one that is based on all the relevant information and is called a rational expectation. A rational expectation is not necessarily a correct forecast. It is simply the best forecast with the information available. It will often turn out to be wrong, but no other forecast that could have been made with the information available could do better.

FIGURE 29.7 Expected Inflation


Potential real GDP is $\$ 16$ trillion. Last year, aggregate demand was $A D_{0}$ and the short-run aggregate supply curve was $S A S_{0}$. The actual price level was the same as the expected price level: 110 . This year, aggregate demand is expected to increase to $A D_{1}$ and the price level is expected to rise from 110 to 121 . As a result, the money wage rate rises and the short-run aggregate supply curve shifts to $S A S_{1}$. If aggregate demand actually increases as expected, the actual aggregate demand curve $A D_{1}$ is the same as the expected aggregate demand curve. Real GDP is $\$ 16$ trillion, and the actual price level rises to 121. The inflation is expected. Next year, the process continues with aggregate demand increasing as expected to $A D_{2}$ and the money wage rate rising to shift the short-run aggregate supply curve to $S A S_{2}$. Again, real GDP remains at $\$ 16$ trillion, and the price level rises, as expected, to 133 .

## Inflation and the Business Cycle

When the inflation forecast is correct, the economy operates at full employment. If aggregate demand grows faster than expected, real GDP rises above potential GDP, the inflation rate exceeds its expected rate, and the economy behaves like it does in a demandpull inflation. If aggregate demand grows more slowly than expected, real GDP falls below potential GDP and the inflation rate slows.

## Deflation

Art economy experiences deflation when it has a persistently falling price level. Equivalently, during a period of deflation, the inflation rate is negative.
In most economies and for most of the time, the inflation rate is positive - the price level is rising-and deflation is rare. But deflation does happen, and most recently it was present in Japan (see Economics in Action on p. 754).
We're going to answer three questions about deflation:

- What causes deflation?
- What are the consequences of deflation?
- How can deflation be ended?


## What Causes Deflation?

The starting point for understanding the cause of deflation is to distinguish between a one-time fall in the price level and a persistently falling price level. A one-time fall in the price level is not deflation. Deflation is a persistent and ongoing falling price level.

A One-Time Fall in the Price Level The price level can fall either because aggregate demand decreases or because short-run aggregate supply increases. So anyof the influences on aggregate demand and short-run aggregate supply that you studied in Chapter 27 canbring a one-time fall in the price level.
Some examples on the demand side are a fall in global demand for a country's exports, or a fall in profit expectations that lowers business investment. Some examples on the supply side are an increase in capital or advance in technology that increases potential GDP, or (unlikely but possible) a fall in the money wage rate.
But none of these sources of a decrease in aggregate demand or increase in aggregate supply can bring a persistently falling price level.

A Persistently Falling Price Level The price level falls persistently if aggregate demand increases at a persistently slower rate than aggregate supply. The trend rate of increase in aggregate supply is determined by the forces that make potential GDP grow. These forces are the growth rates of the labor force and capital stock and the growth rate of productivity that results from technological change. Notice that all thesevariables are real, not monetary, and they have trendsthat change slowly. In contrast, the forces that drive aggregate demand include the quantity of money. And this quantity can grow as quickly or as slowly as the central bank chooses.
In most situations, the central bank doesn't have a target for the money stock or its growth rate and instead sets the interest rate. But the money stock is under central bank control, and its growth rate has a powerful effect on the growth rate of aggregate demand. To see the effect of growth in the money stock in the long term, we need to return to the quantitytheory of money.

The Quantity Theory and Deflation The quantity theory of money explains the trends in inflation by focusing on the trend influences on aggregate supply and aggregate demand. The foundation of the quantity theory is the equation of exchange (see Chapter 25, p.646), which in its growth rate version and solved for the inflation rate states

$$
\underset{\text { rate }}{\text { Inflation }}=\underset{\text { growth rate }}{\text { Money }}+\begin{gathered}
\text { Rate of } \\
\text { velocity } \\
\text { change }
\end{gathered}-\underset{\text { Real GDP }}{\text { growth rate }}
$$

This equation, true by definition, derives from the fact that the amount of money spent on real GDP, $M V$, equals the money value of GDP, $P Y$. ( $M$ is the money stock, Vis its velocity of circulation, $P$ is the price level, and $Y$ is real GDP.)
The quantity theory adds to the equation of exchange two propositions. First, the trend rate of change in the velocity of circulation does not depend on the money growth rate and is determined by decisions about the quantity of money to hold and to spend. Second, the trend growth rate of real GDP equals the growth rate of potential GDP and, again, is independent of the money growth rate.
With these two assumptions, the equation of exchange becomes the quantity theory of money and predicts that a change in the money growth rate brings an equal change in the inflation rate.
For example, suppose velocity increases by 2 percent per year and potential GDP grows by 3 percent per year. Then the quantity theory predicts that the trend inflation rate equals the money growth rate minus 1 percent. If the central bank makes the quantity of money grow by 1 percent, the inflation rate will be zero. If money grows at a rate faster than 1 percent, the economy will experience inflation. And if money grows at a slower rate than 1 percent, the economy will experience deflation.

Japan Example In the example of Japan during the1990s and 2000s (see Economics in Action on previous page), the money (M2) growth rate during the 15 years 1998-2013 was 2.5 percent per year. The velocity growth rate was negative and it decreased at a rate of 3 percent per year. Potential GDP grew at an average rate of 0.8 percent per year. Combining these numbers, the quantity theory predicts an inflation rate equal to -1.3 percent per year:

$$
+2.5+(-3)-0.8=-1.3
$$

In fact, the average inflation rate was -1.2 percent per year. So the quantity theory prediction is not exactly correct, but it is close. Its prediction of the deflation rate of 1.3 percent per year is off by only 0.1 . You now know what causes deflation. Let's turn to its consequences.

## ECONOMICS IN ACTION

Fifteen Years of Deflation in Japan
Japan experienced deflation for the 15 years from 1998 to 2013 .

## Japan's Deflation Rate

Figure 1 shows the inflation rate in Japan from 1990 to 2013. The inflation rate fluctuated between -1 percent and -2 percent per year and accumulated to a 17 percent fall in the price level.

## Cause of Japan's Deflation

Deflation, like its opposite, inflation, is primarily a monetary phenomenon. Japan's money stock grew too slowly during the deflation years.

Figure 2 shows the facts about inflation and money growth in Japan from 1995 to 2013. The relevant money growth rate that brings inflation or deflation is that of money itself plus the trend rate of change in the velocity of circulaton minus the growth rate of potential GDP. That is the money growth rate shown in Fig. 2 and except for one year, 1997, it is negative, which means that Japan's money stock did not grow fast enough to accommodate the growth of potential GDP and a trend rise in velocity.

## Consequences of Japan's Deflation

At first, Japan's deflation was unexpected and loan and wage contracts had been entered into that anticipated an ongoing low but positive inflation rate. So when the price level started to fall, the real value of debt increased and the real wage rate increased.

With higher real debt and wages, businesses cut back on both investment and hiring labor and cut production. Real GDP fell and the recessionary gap increased.

Because investment decreased, the capital stock increased more slowly and the growth rate of potential GDP slowed. From being one of the world's most dynamic rich economies, Japan became the world's most sluggish.

Figure 3 tells the story. The 1960s saw Japan doubling its real GDP in seven years. The growth rate slowed in the 1970s and 1980s but remained one of the world's fastest. Then, during the deflation years, the growth rate dropped to 1.5 percent (in the 1990s) and 0.5 percent (in the 2000s).

Japan's inflation rate turned positive in 2014, and real GDP growth picked up, but money growth rate remained too low. Without a sustained increase in money growth, deflation cannot end.


Figure 1 Japan's Long Deflation


Figure 2 Money Growth Rate Too Low


Figure 3 Japan's Decade Average Real GDP Growth Rates
Sources of data: Financial Stotistics and World Economic Outlook, International Monetary Fund, Washington, DC.

## What are the Consequences of Deflation?

Chapter 22 (p. 562) discusses why deflation and inflation are problems. But with what you now know about aggregate supply and aggregate demand and the determinants of potential GDP and its growthrate, you can gain deeper insight into the costs of deflation (and the related costs of inflation.)
The effects of deflation (like those of inflation) depend on whether it is anticipated or unanticipated. But because inflation is normal and deflation is rare, when deflation occurs, it is usually unanticipated. Unanticipated deflation redistributes income and wealth, lowers real GDP and employment, and diverts resources from production.
Workers with long-term wage contracts find their real wages rising. But on the other side of the labormarket, employers respond to a higher and rising realwage by hiring fewer workers. So the level of employment and output falls.
With lower output and profits, firms re-evaluate their investment plans and cut back on projects that they now see as unprofitable. This fall in investment slows the pace of capital accumulation and slows the growth rate of potential GDP.
Another consequence of deflation is a low nominal interest rate, which, in turn, brings an increase in the quantity of money that people plan to hold and a de- crease in the velocity of circulation. A lower velocity adds to the deflationary forces and, if unattended to, lowers the inflation rate yet further. So, what is the cure for deflation?

## How Can Deflation be Ended?

Deflation can be ended by removing its cause: The quantity of money is growing too slowly. If the central bank ensures that the quantity of money grows at the target inflation rate plus the growth rate of potential GDP minus the growth rate of the velocity of circulation, then, on average, the inflation rate will turn out to be close to target. In the example of Japan, if the Bank of Japan, the central bank, wanted to get a 2 percent inflation rate, and other things remaining the same, it would have needed to make the quantity of money grow at an annual average rate of 5.8 percent. (Money growth 5.8 plus velocity growth of -3 minus potential GDP growths of 0.8 equals target inflation of 2 percent.) If raising the inflation rate brought faster potential GDP growth, a yet higher money growth rate would be needed to sustain the higher inflation rate.

Money Growth, Not Quantity Notice that it is an increase in the growth rate of the money stock, not a one-time increase in the quantity of money that is required to end deflation. Central banks sometimes increase the quantity of money and fail to increase its growth rate. An increase in the level with no change in the growth rate brings a temporary inflation as the price level adjusts but not ongoing inflation, so it does not end deflation.

## The Phillips Curve

The Phillips curve is a relationship between inflation and unemployment. It is so named because it was first suggested by New Zealand economist A.W (Bill) Phillips. We distinguish between two time frames for the Phillips curve (similar to the two aggregate supply time frames). We study

- The short-run Phillips curve
- The long-run Phillips curve


## The Short-Run Phillips Curve

The short-run Phillips curve is the relationship between inflation and unemployment, holding constant

1. The expected inflation rate
2. The natural unemployment rate You've seen what determines the expected inflation rate earlier in this chapter (see p. 752) and the influences on the natural unemployment rate were explained in Chapter 22 (pp. 559-560).
Figure 29.8 on page 758 shows a short-run Phillips curve, SRPC. Suppose that the expected inflation rate is 10 percent a year and the natural unemployment rate is 6 percent, point $A$ in the figure. A short-runPhillips curve passes through this point. Ifinflation rises above its expected rate, unemployment falls below its natural rate in a movement up along the short-run Phillips curve from point $A$ to point $B$. Similarly, if inflation falls below its expected rate, unemployment rises above its natural rate in a movement down along the short-run Phillips curve from point $A$ to point $C$.

## The Long-Run Phillips Curve

The long-run Phillips curve is the relationship between inflation and unemployment when the actual inflation rate equals the expected inflation rate. The long- run Phillips curve is vertical at the natural unemployment rate because, in the long run, any expected inflation rate is possible. In Fig. 29.9(a), the longrun Phillips curve is the vertical line LRPC.

Change in Expected Inflation A change in the expected inflation rate shifts the short-run Phillips curve, but it does not shift the long-run Phillips curve. In Fig. 29.9(a), if the expected inflation rate is 10 percent a year, the short-run Phillips curve is $S R P C_{0}$. If the expected inflation rate falls to 6 percent a year, the short-run Phillips curve shifts downward to $S R P C_{1}$. The vertical distance by which the short-run Phillips curve shifts from point $A$ to point $D$ is equal to the change in the expected inflation rate. If the actual inflation rate also falls from 10 percent to 6 percent, there is a movement down the long-run Phillips curve from $A$ to $D$. An increase in the expected inflation rate has the opposite effect to that shown in Fig. 29.9(a). The other source of a shift in the Phillips curve is a change in the natural unemployment rate.

Change in Natural Unemployment Rate A change in the natural unemployment rate shifts both the short-run and long-run Phillips curves. Figure 29.9(b) illustrates such shifts. If the natural unemployment rate increases from 6 percent to 9 percent, the long-run Phillips curve shifts from $\angle R P C_{0}$ to $L R P C_{1}$, and if expected inflation is constant at 10 percent a year, the short-run Phillips curve shifts from $S R P C_{0}$ to $S R P C_{1}$. Because the expected inflation rate is constant, $S R P C_{1}$ intersects the long-run curve $L R P C_{1}$ (point $E$ ) at the same inflation rate at which $S R P C_{0}$ intersects the long-runcurve $\angle R P C_{0}$ (point $A$ ).

FIGURE 29.9 Short-Run and Long-Run

## Phillips Curves


(a) A change in expected inflation

(b) A change in natural unemployment

In part (a), the long-run Phillips curve is LRPC. A fall in expected inflation shifts the short-run Phillips curve downward from $S R P C_{0}$ to $S R P C_{1}$. The long-run Phillips curve does not shift. In part (b), a change in the natural unemployment rate shifts both the short-run and long-run Phillips curves.

## ECONOMICS IN ACTION

## The U.S. Phillips Curve

The figure below is a scatter diagram of the U.S. inflation rate (measured by the GDP deflator) and the unemployment rate since 2001. LRPC is at a natural unemployment rate of 5.5 percent and $S R P C$ at an expected inflation rate of 2 percent. The dots for each year (five of which are identified) show that the SRPC jumps around as inflation expectations change.


The U.S. Phillips Curve in the 2000s

## ECONOMIC ANALYSIS

- The Eurozone is the group of 18 European countries that use the euro as their money and for which the European Central Bank (ECB) makes monetary policy decisions.
- The Eurozone economy is stagnating and has a high unemployment rate.
- Figure 1 shows the Eurozone unemployment rate compared with that of the United States.
- The Eurozone unemployment rate has been persistently higher than that of the United States and the average difference is structural, not cyclical.
- A high structural unemployment rate in the Eurozone results from high minimum wages, generous unemployment benefits and welfare payments, and extensive regulation of the labor market.
- ECB monetary policy can do nothing to lower the structural unemployment rate. But it can act to lower the cyclical unemployment rate.
- Eurozone also has a low inflation rate that is below the ECB target rate of 2 percent per year.
- Figure 2 shows the Eurozone inflation rate compared with that of the United States. Both economies had inflation rates below 2 percent per year in 2013, but in the Eurozone inflation had been below 2 percent for 6 years.
- The high unemployment and stagnating real GDP result from real structural problems that make the Eurozone natural unemployment rate high and from high cyclical unemployment and below-target inflation that result from insufficient aggregate demand.
- The aggregate demand problem arises from the fact that the ECB has not expanded the money stock quickly enough.
- Figure 3 shows the growth rate of money plus the growth rate of velocity minus the growth rate of potential GDP.
- The growth rate of money plus the growth rate of velocity minus the growth rate of potential GDP equals the inflation rate that can be sustained at full employment.
- To lower cyclical unemployment, the growth rate of money plus the growth rate of velocity minus the growth rate of potential GDP must exceed the target and expected inflation rate.
- If, as in 2009 and 2010, the growth rate of money plus the growth rate of velocity minus the growth rate of potential GDP decreases, cyclical unemployment will increase and inflation will decrease.
- To end stagnation, the ECB must buy assets and increase the growth rate of money. A big one-off asset purchase will not do the job required.


Figure 1 The Stagnating Eurozone Economy


Figure 2 Inflation Rates Miss Targets


Figure 3 Money Growth Rate Too Low

## PART TEN: MACROECONOMIC POLICY

## CHAPTER 30: Fiscal Policy

After studying this chapter, you will be able to:

- Describe the federal budget process and the recent history of outlays, receipts, deficits, and debt
- Explain the supply-side effects of fiscal policy
- Explain how fiscal policy choices redistribute benefits and costs across generations
- Explain how fiscal stimulus is used to fight a recession

Most governments have a budget deficit and debt, but Japan's tops them all and might be a warning to the United States and others about dangers that lie ahead. How do government deficits and debt influence the economy? Do they create jobs, or do they destroy them? Do they slow economic growth? Do they impose a burden on future generations?
This chapter studies these fiscal policy questions in the U.S. economy today. In Economics in the News at the end of the chapter, we look at the extreme fiscal policy challenges facing Japan and the lesson theyhold for the United States.

## The Federal Budget

The federal budget is an annual statement of the out- lays and receipts of the government of the United States together with the laws and regulations that approve and support them. The federal budget has two purposes:

1. To finance federal government programs and activities, and
2. To achieve macroeconomic objectives

The first purpose of the federal budget was its only purpose before the Great Depression of the 1930 s . The second purpose arose as a reaction to the Great Depression and the rise of the ideas of economist John Maynard Keynes. The use of the federal budget to achieve macroeconomic objectives such as full employment, sustained economic growth, and price level stability is called fiscal policy. It is this aspect of the budget that is the focus of this chapter.

The Institutions and Laws Fiscal policy is made by the president and Congress on an annual timeline that is shown in Fig. 30.1 for the 2015 budget.

The Roles of the President and Congress The president proposes a budget to Congress each February. Congress debates the proposed budget and passes the budget acts in September. The president either signsthose acts into law or vetoes the entire budget bill. The president does not have the veto power to eliminate specific items in a budget bill and approve others - known as a line-item veto. Many state governors have long had line-item veto authority. Congress attempted to grant these powers to the president of the United States in 1996, but in a 1998 Supreme Court ruling, the line-item veto for the president was declared unconstitutional. Although the president proposes and ultimately approves
the budget, the task of making the tough decisions on spending and taxes rests with Congress.
Congress begins its work on the budget with the president's proposal. The House of Representativesand the Senate develop their own budget ideas in their respective House and Senate Budget Committees. Formal conferences between the two houses eventually resolve differences of view, and a series ofspending acts and an overall budget act are usually passed by both houses before the start of the fiscal year. A fiscal year is a year that runs from October 1 to September 30 in the next calendar year. Fiscal 2015 is the fiscal year that begins on October 1, 2014.
During a fiscal year, Congress often passes supplementary budget laws, and the budget outcome is influenced by the evolving state of the economy. Forexample, if a recession begins, tax revenues fall and welfare payments increase.

FIGURE 30.1 The Federal Budget Timeline in Fiscal 2015


Congress debates, amends, and enacts the budget.

September
Oct 1, 2014
Fiscal year 2015 begins.

Supplementary budget laws may be passed.
State of economy influences outlays, receipts, and the budget deficit.

Sept 30, 2015 Fiscal year 2015 ends.
Accounts of fiscal year 2015 are prepared. Outlays, receipts, and the budget deficit are reported.

The federal budget process begins with the president's request in February. Congress debates and amends the request and enacts a budget before the start of the fiscal year on October 1. The president signs the budget acts into law or vetoes the entire budget bill. Throughout the fiscal year, Congress might pass supplementary budget laws. The budget outcome is calculated after the end of the fiscal year.

The Employment Act of 1946 Fiscal policy operates within the framework of the landmark Employment Act of 1946 in which Congress declared that ... it is the continuing policy and responsibility of the Federal Government to use all practicable means ... to coordinate
and utilize all its plans, functions, and resources ... to promote maximum employment, production, and purchasing power.
This act recognized a role for government actions to keep unemployment low, the economy expanding, andinflation in check. The Full Employment and Balanced Growth Act of 1978, more commonly known as the Humphrey-Hawkins Act, went farther than the Employment Act of 1946 and set a specific target of 4 percent for the unemployment rate. But this target hasnever been treated as an unwavering policy goal. Under the 1946 act, the president must describe the currenteconomic situation and the policies he believes are needed in the annual Economic Report of the President, which the Council of Economic Advisers writes.

The Council of Economic Advisers The president's Council of Economic Advisers was established in the Employment Act of 1946. The Council consists of achairperson and two other members, all of whom are economists on a one- or two-year leave from their regular university or public service jobs. In 2014, the chair of President Obama's Council of EconomicAdvisers was Jason Furman formerly at the Brookings Institution. The Council of Economic Advisers monitors the economy and keeps the President and the publicinformed about the current state of the economy and the best available forecasts of where it is heading.
This economic intelligence activity is one source of data that informs the budget-making process. Let's look at the most recent federal budget.

## Highlights of the 2015 Budget

Table 30.1 shows the main items in the federal budget proposed by President Obama for 2015. The numbers are projected amounts for the fiscal year beginning on October I, 2014Fiscal 2015. Notice the three main parts of the table: Receipts are the government's tax revenues, outlays are the government's payments, and the deficit is the amount by which the government's outlays exceed its receipts.

Receipts were projected to be $\$ 3,514$ billion in Fiscal 2015. These receipts come from four sources:

1. Personal income taxes
2. Social Security taxes
3. Corporate income taxes
4. Indirect taxes and other receipts

The largest source of receipts is personal income taxes, which in 2015 are expected to be $\$ 1,505$ billion. These taxes are paid by individuals on their incomes. The second largest source is SocialSecurity taxes. These taxes are paid by workers and their employers to finance the government's Social Security programs. Third in size are corporate income taxes. These taxes are paid by companies on theirprofits. Finally, the smallest source of federal receiptsis what are called indirect taxes. These taxes are on the sale of gasoline, alcoholic beverages, and a few other items.

TABLE 30.1 Federal Budget in Fiscal 2015

| Item | Projections <br> (billions of dollars) |  |
| :--- | ---: | ---: |
|  | 3,514 |  |
| Receipts |  | 1,505 |
| Personal income taxes |  | 1,176 |
| Social Security taxes | 537 |  |
| Corporate income taxes |  | 296 |
| Indirect taxes and other receipts |  |  |
| Outlays | $\mathbf{4 , 1 5 8}$ |  |
| Transfer payments |  | 2,649 |
| Expenditure on goods and services |  | 1,030 |
| Debt interest | $\mathbf{6 4 4}$ |  |
| Deficit |  |  |

Source of data: Budget of the United States Government, Fiscal Year 2015, Table 14.1.

Outlays are classified into three categories:

1. Transfer payments
2. Expenditure on goods and services
3. Debt interest

The largest item of outlays, transfer payments, is the payment to individuals, businesses, other levels of government, and the rest of the world. In 2015, this item is expected to be $\$ 2,649$ billion. It includes Social Security benefits, Medicare and Medicaid, unemployment checks, welfare payments, farm subsidies, grants to state and local governments, and payments to international agencies. It also includes capital transfers to bail out failing financial institutions. Transfer payments, especially those for Medicare and Medicaid, are sources of persistent growth in government expenditures and are a major source of concern and political debate.
Expenditure ongoods and services is the expenditure on final goods and services, and in 2015 it isexpected to total $\$ 1,030$ billion. This expenditure, which includes that on national defense, homelandsecurity, research on cures for AIDS, computers forthe Internal Revenue Service, government cars andtrucks, and federal highways, has decreased in recent years. This component of the federal budget is the government expenditure ongoods and services that appears in the circular flow of expenditure and income and in the National Income and Product Accounts (see Chapter 21, pp. 531-532).
Debt interest is the interest on the government debt. In 2015, this item is expected to be $\$ 479$ billion - almost 12 percent of total expenditure. This interest payment is large
because the government has a debt of almost $\$ 13$ trillion, which has arisen from many years of budget deficits during the 1970s, 1980s, 1990s, and 2000s.

Surplus or Deficit The government's budget balance is equal to receipts minus outlays.
Budget balance = Receipts - Outlays.

If receipts exceed outlays, the government has a budget surplus. If outlays exceed receipts, the government has a budget deficit. If receipts equal outlays, the government has a balanced budget. For Fiscal 2015, with projected outlays of $\$ 4,158$ billion and receipts of $\$ 3,514$ billion, the government projected a budget deficit of $\$ 644$ billion.
Big numbers like these are hard to visualize and hard to compare over time. To get a better sense ofthe magnitude of receipts, outlays, and the deficit, we often express them as percentages of GDP.
Expressing them in this way lets us see how large government is relative to the size of the economy, and italso helps us to study changes in the scale of government over time. How typical is the federal budget of Fiscal 2015?
Let's look at the recent history of the budget.

FIGURE 30.2 The Budget Surplus and Deficit


## The Budget in Historical Perspective

Figure 30.2 shows the government's receipts, outlays, and budget surplus or deficit since 1990. You can see that except for the four years around 2000, the budget has been in persistent deficit.
You can also see that after 2008, the deficit was extraordinarily large, peaking in 2010 at more than 10 percent of GDP and remaining close to 10 percent for three years.

An earlier large deficit in 1992 gradually shrank through the 1990s expansion and in 1998 the first budget surplus since 1969 emerged. But by 2002, thebudget was again in deficit and during the 2008-2009recessions, the deficit reached a new all-time high.
Why did the budget deficit grow during the early 1990s, vanish in the late 1990s, and reemerge in the 2000s? Did outlays increase, or did receipts shrink, and which components of outlays and receiptschanged most to swell and then shrink the deficit? Let's look at receipts and outlays in a bit more detail.

Receipts Figure 30.3(a) on page 771 shows the components of government receipts as percentages of GDP from 1990 to 2015 . Total receipts fluctuate because personal income taxes and corporate income taxes fluctuate. Other receipts (Social Security taxes and indirect taxes) are a near-constant percentage of GDP.
Personal and corporate income tax receipts trended upward during the 1990s, downward during the 2000s, and then upward again after 2010.

Outlays Figure 30.3(b) on page 771 shows the components of government outlays as percentages of GDP from 1990 to 2015 . Two features of government outlays stand out. First, expenditure on goods and services decreased from 1990 through 2000 and then increased. The increase after 2000 was mainly on security-related goods and services in the wake of the attacks that occurred on September 11, 2001, and defense expenditure. Second, transfer payments increased over the entire period and exploded after 2008 when the government tried to stimulate economic activity.
You've seen that the U.S. government budget deficit is large. But how does it compare to the deficits of other countries? The answer: It is one of the largest as Economics in Action (p. 736) shows. Of the major economies, only Japan has a larger deficit as apercentage of GDP. Deficits bring debts, as you will now see.

## Budget Balance and Debt

When the government has a budget deficit it borrows, and when it has a budget surplus it makes loanrepayments. Government debt is the total amount that the government has borrowed. It is the sum of past budget deficits minus the sum of past budget surpluses. A government budget deficit increases government debt. A persistent budget deficit feeds itself: It leads to increased borrowing, which leads to largerinterest payments, which in turn lead to a larger deficit. That is the story of an increasing budget deficit during the 1970s and 1980s and again today.
Figure 30.4 shows government debt since 1940, measured as a percentage of GDP-the debt-to-GDP ratio. The government debt-to-GDP ratio was at an all-time high at the end of World War II when it exceeded 110 percent. Budget surpluses and rapid economic growth, especially during the 1960s, lowered the debt-to-GDP ratio through 1974. Small budget deficits increased the ratio slightly through the 1970s, and large budget deficits increased it dramatically during the 1980s and the 1990-1991 recession.
The growth rate of the debt-to-GDP ratio slowed as the economy expanded during the mid-1990s and fell when the budget went into surplus in the late 1990s and early 2000s.

After the global financial crisis of 2008, when the budget deficit reached a record high for peacetime, and real GDP stopped growing, the debt-to-GDPratio climbed again, and steeply.

Debt and Capital Businesses and individuals incur debts to buy capital-assets that yield a return. In fact, the main point of debt is to enable people to buy assets that will earn a return that exceeds theinterest paid on the debt. The government is similarto individuals and businesses in this regard. Muchstate government expenditure is on public assets suchas highways, public schools, and universities, that yield a social return greater than the interest rate.
Federal government assets, most of which are national defense capital, were valued at $\$ 1.8$ trillion in 2013. But federal debt, at $\$ 12.6$ trillion, is seven times the value of the Federal government's capital stock. So some government debt has been incurred to finance public consumption expenditure and transfer payments, which do not have a social return. Futuregenerations bear the cost of this debt.
figure 30.4 The Federal Government Debt


Government debt (the accumulation of past budget deficits minus past budget surpluses) was at its highest at the end of World War II. Debt as a percentage of GDP fell through 1974 but then started to increase. After a further brief decline during the 1990s, it exploded during the 2010s.

## ECONOMICS IN ACTION

## The U.S. Covernmerat Budger im Globoll Perspective

How does the U.S. government budget deficit compare with those of other major economies?

## Comparing Like with Like

To compare the budget deficits of governments across economies, we must take into account the fact that some countries, and the United States is one of them, have large state and local governments, while others, and the United Kingdom is one, have a large central government and small local governments. These differences make the international comparison more valid at the level of total government.

## Deficits Almost Everywhere

The figure shows the budget balances of all levels of government in eight economies in 2014. Fiscal stimulus to fight the global recession of 2008 resulted in deficits almost everywhere. Of the countries shown here, only Germany had a budget surplus in 2014.

Japan had the largest deficit and the United States had the second largest. The United Kingdom and some other European countries also had large deficits.

Italy and other advanced economies as a group, which includes the newly industrialized economies of Asia (Hong Kong, South Korea, Singapore, and Taiwan) had the smallest deficits.


The total government sector of the United States includes state and local governments as well as the federal government. In Fiscal 2015, when federal government outlays were \$4, 158 billion, state and local outlays were a further $\$ 2,700$ billion. Most of these expenditures were on public schools, colleges, and universities ( $\$ 550$ billion); local police and fire services; and roads.
It is the combination of federal, state, and local government receipts, outlays, and budget deficits that influences the economy. But state and local budgets are not designed to stabilize the aggregate economy.

So sometimes, when the federal government cuts taxes or outlays, state and local governments do the reverse and, to a degree, cancel out the effects of the federal actions. For example, since 2000, federal taxes have decreased as a percentage of GDP, but state and local taxes and total government taxes have increased. Now that you know what the federal budget is and what the main components of receipts and out-lays are, it is time to study the effects of fiscal policy.

We begin by learning about the effects of taxes on employment, aggregate supply, and potential GDP. Then we see how fiscal policy brings redistribution across generations. Finally, we look at fiscal stimulus and see how it might be used to speed recoveryfrom recession and stabilize the business cycle.

## Supply-Side Effects of Fiscal Policy

How do taxes on personal and corporate income affect real GDP and employment? The answer to these questions is controversial. Some economists, known as supply-siders, believe these effects to be large and an accumulating body of evidence suggests that they are correct. To see why these effects might be large, we'll begin with a refresher on how full employment and potential GDP are determined in the absence of taxes. Then we'll introduce an income tax and see how it changes the economic outcome.

## Full Employment and Potential GDP

Youlearned in Chapter 23 (pp. 584-586) how the full-employment quantity of labor and potential GDP are determined. At full employment, the real wage rate adjusts to make the quantity of labor demanded equal the quantity of labor supplied. Potential GDP is the real GDP that the full-employment quantity of labor produces.
Figure 30.5 illustrates a full-employment situation.
In part (a), the demand for labor curve is $L D$ and the supply of labor curve is $L S$. At a real wage rate of $\$ 30$ an hour, 250 billion hours of labor a year are employed and the economy is at full employment. In Fig. 30.5(b), the production function is PF. When 250 billion hours of labor are employed, real GDP - which is also potential GDP - is \$16 trillion. Let's now see how an income tax changes potential GDP.

FIGURE 30.5 The Effects of the Income Tax on Aggregate Supply

(a) Income tax and the labor market

(b) Income tax and potential GDP

In part (a), with no income tax, the real wage rate is $\$ 30$ an hour and employment is 250 billion hours. In part (b), potential GDP is $\$ 16$ trillion. An income tax shifts the supply of labor curve leftward to $L S+$ tax. The before-tax wage rate rises to $\$ 35$ an hour, the after-tax wage rate falls to $\$ 20$ an hour, and the quantity of labor employed decreases to 200 billion hours. With less labor, potential GDP decreases.

## The Effects of the Income Tax

The tax on labor income influences potential GDPand aggregate supply by changing the full-employment quantity of labor. The income tax weakens the incentive to work and drives a wedge between the take-home wage of workers and the cost of labor to businesses. The result is a smaller quantity of labor and a smaller potential GDP.
Figure 30.5 shows this outcome. In the labor market, the income tax has no effect on the demand for labor, which remains at $L D$. The reason is that the quantity of labor that businesses plan to hire depends only on how productive labor is and what it costs-its real wage rate.

But the supply of labor does change. With no income tax, the real wage rate is $\$ 30$ an hour and 250 billion hours of labor a year are employed. An income tax weakens the incentive to work and decreases the supply of labor. The reason is that for each dollar of before-tax earnings, workers must pay the government an amount determined by the income tax code. So workers look at the after-tax wage rate when they decide how much labor to supply. An income tax shifts the supply curve leftward to $L S+$ tax. The vertical distance between the $L S$ curve and the $L S+$ tax curve measures the amount of income tax. With the smaller supply of labor, the before-tax wage rate rises to $\$ 35$ an hour but the after-tax wage rate falls to $\$ 20$ an hour. The gap created between the before-tax and after-tax wage rates iscalled the tax wedge.
The new equilibrium quantity of labor employed is 200 billion hours a year-less than in the no-taxcase. Because the full-employment quantity of labor decreases, so does potential GDP. And a decrease in potential GDP decreases aggregate supply.
In this example, the tax rate is high- $\$ 15$ tax on a $\$ 35$ wage rate is a tax rate of about 43 percent. A lower tax rate would have a smaller effect on employment and potential GDP. An increase in the tax rate to above 43 percent would decrease the supply of labor by more than the decrease shown in Fig. 30.5. Equilibrium employment and potential GDP would also decrease still further. A tax cut would increase the supply oflabor, increase equilibrium employment, and increase potential GDP.

## Taxes on Expenditure and the Tax Wedge

The tax wedge that we've just considered is only a part of the wedge that affects labor-supply decisions. Taxes on consumption expenditure add to the wedge. The reason is that a tax on consumption raises the prices paid for consumption goods and services and is equivalent to a cut in the real wage rate.
The incentive to supply labor depends on the goods and services that an hour of labor can buy. The higher the taxes on goods and services and the lowerthe after-tax wage rate, the less is the incentive to supply labor. If the income tax rate is 25 percent andthe tax rate on consumption expenditure is 10 percent, a dollar earned buys only 65 cents worth of goods and services. The tax wedge is 35 percent.

## ECONOMICS IN ACTION

## Some Real-World Tax Wedges

Edward C. Prescott of Arizona State University, who shared the 2004 Nobel Prize for Economic Science, has estimated the tax wedges for a number of countries, among them the United States, the United Kingdom, and France.

The U.S. tax wedge is a combination of 13 percent tax on consumption and 32 percent tax on incomes. The income tax component of the U.S. tax wedge includes Social Security taxes and is the marginal tax rate-the tax rate paid on the marginal dollar earned.

Prescott estimates that in France, tax rates on consumption are 33 percent and on incomes are 49 percent.

The estimates for the United Kingdom fall between those for the United States and France. The figure shows these components of the tax wedges in the three countries.

## Does the Tax Wedge Maiter?

According to Prescott's estimates, the tax wedge has a powerful effect on employment and potential GDP. Potential GDP in France is 30 percent below that of the United States (per person), and the entire difference can be attributed to the difference in the tax wedge in the two countries.

Potential GDP in the United Kingdom is 28 percent below that of the United States (per person), and about a third of the difference arises from the different tax wedges. (The rest is due to different productivities.)


Three Tax Wedges

## Taxes and the Incentive to Save and Invest

A tax on interest income weakens the incentive to save and drives a wedge between the aftertax interest rateearned by savers and the interest rate paid by firms.
These effects are analogous to those of a tax on labor income. But they are more serious for two reasons.
First, a tax on labor income lowers the quantity of labor employed and lowers potential GDP, while a tax on capital income lowers the quantity of saving and investment and slows the growth rate of real GDP. Second, the true tax rate on interest income is much higher than that on labor income because of the way in which inflation and taxes on interest income interact. Let's examine this interaction.

Effect of Tax Rate on Real Interest Rate The interest rate that influences investment and saving plans is the real after-tax interestrate. The real after-taxinterest rate subtracts the income tax rate paid on interest income from the real interest rate. But the taxes depend on the nominal interest rate, not the real interest rate. So the higher the inflation rate, the higher is the true tax rate on interest income. Here is an example. Suppose the real interest rate is 4 percent a year and the tax rate is 40 percent.
If there is no inflation, the nominal interest rate equals the real interest rate. The tax on 4 percent interest is 1.6 percent ( 40 percent of 4 percent), so the real after-tax interest rate is 4 percent minus 1.6 percent, which equals 2.4 percent.
If the inflation rate is 6 percent a year, the nominal interest rate is 10 percent. The tax on 10 percent interest is 4 percent ( 40 percent of 10 percent), so the real after-tax interest rate is 4 percent minus 4 per-cent, which equals zero. The true tax rate in this case is not 40 percent but 100 percent!

Effect of Income tax on Saving and Investment In Fig. 30.6, initially there are no taxes. Also, the government has a balanced budget. The demand for loanable funds curve, which is also the investment demand curve, is DLF. The supply of loanable fundscurve, which is also the saving supply curve, is SLF. The equilibrium interest rate is 3 percent a year, and the quantity of funds borrowed and lent is $\$ 2$ trilliona year.
A tax on interest income has no effect on the demand for loanable funds. The quantity of investmentand borrowing that firms plan to undertake depends only on how productive capital is and what it costs - in real interest rate. But a tax on interest income weakens the incentive to save and lend and decreases the supply of loanable funds. For each dollar of before-tax interest,savers must pay the government an amount determinedby the tax code. So savers look at the after-tax real interest rate when they decide how much to save.

## figure 30.6 The Effects of a Tax on Capital Income



The demand for loanable funds and investment demand curve is $D L F$, and the supply of loanable funds and saving supply curve is SLF. With no income tax, the real interest rate is 3 percent a year and investment is $\$ 2$ trillion. An income tax shifis the supply curve leftward to SLF + tax. The interest rate rises to 4 percent a year, the after-tax interest rate falls to 1 percent a year, and investment decreases to $\$ 1.8$ trillion. With less investment, the real GDP growth rate decreases.
When a tax is imposed, saving decreases and the supply of loanable funds curve shifts leftward to $S L F+$ tax. The amount of tax payable is measured by the vertical distance between the SLF curve andthe SLF+tax curve. With this smaller supply of loanable funds, the interest rate rises to 4 percent a year but the after-tax interest rate falls to 1 percent a year. A tax wedge is driven between the interest rate and the after-tax interest rate, and the equilibrium quantity of loanable funds decreases. Saving and investmentalso decrease.

Tax Revenues and the Laffer Curve An interesting consequence of the effect of taxes on employment and saving is that a higher tax rate does not always bring greater tax revenue. A higher tax rate brings in more revenue per dollar earned. But because a higher tax rate decreases the number of dollars earned, two forces operate in opposite directions on the tax revenue collected.
The relationship between the tax rate and the amount of tax revenue collected is called the Laffer curve. The curve is so named because Arthur B. Laffer, a member of President

Reagan's Economic Policy Advisory Board, drew such a curve on a table napkin and launched the idea that tax cuts could increase tax revenue.
Figure 30.7 shows a Laffer curve. The tax rate is on the $x$-axis, and total tax revenue is on the $y$-axis. For tax rates below $T^{*}$, an increase in the tax rateincreases tax revenue; at $T^{*}$, tax revenue is maxmized; and a tax rate increase above $T^{*}$ decreasestax revenue.
Most people think that the United States is on the upward-sloping part of the Laffer curve; so is the United Kingdom. But France might be close to the maximum point or perhaps even beyondit.

## The Supply-Side Debate

Before 1980, few economists paid attention to the supply-side effects of taxes on employment and potential GDP. Then, when Ronald Reagan took office as president, a group of supply-siders began to argue the virtues of cutting taxes. Arthur Laffer was one of them. Laffer and his supporters were not held in high esteem among mainstream economists, but they were influential for a period. They correctly argued that tax cuts would increase employment and increase output. But they incorrectly argued that tax cuts would increase tax revenues and decrease the budget deficit. For this prediction to be correct, the United States would have had to be on the "wrong" side of the Laffer curve. Given that U.S. tax rates are among the lowest in the industrial world, it is unlikely that this condition was met. And when the Reagan administration did cut taxes, the budget deficit increased a fact that reinforces this view.
Supply-side economics became tarnished because of its association with Laffer and came to be called "voodoo economics." But mainstream economists, including Martin Feldstein, a Harvard professor whowas Reagan's chief economic adviser, recognized the power of tax cuts as incentives but took the standard view that tax cuts without spending cuts would swell the budget deficit and bring further serious problems. This view is now widely accepted by economists of allpolitical persuasions.

## Generational Effects of Fiscal Policy

Is a budget deficit a burden on future generations? If it is, how will the burden be borne? And is the budget deficit the only burden on future generations? What about the deficit in the Social Security fund?
Does it matter who owns the bonds that the government sells to finance its deficit? What about the bonds owned by foreigners? Won't repaying those bonds impose a bigger burden than repaying bonds owned by Americans?
To answer questions like these, we use a tool called generational accounting-an accounting systemthat measures the lifetime tax burden and benefits of each generation. This accounting system was developed by Alan Auerbach of the University of Pennsylvania and Laurence Kodikoff of Boston University. Generational accounts for the United States have been prepared by Jagadeesh Gokhale of the Cato Institute and Kent Smetters of the Wharton School at the University of Pennsylvania.
figure 30.7 A Laffer Curve


A Laffer curve shows the relationship between the tax rate and tax revenues. For tax rates below $T^{*}$, an increase in the tax rate increases tax revenue. At the tax rate $T^{\star}$, tax revenve is maximized. For tax rates above $T^{*}$, an increase in the tax rate decreases tax revenue.

## Generating Accounting and Present Value

Income taxes and Social Security taxes are paid by people who have jobs. Social Security benefits are paid to people after they retire. So to compare taxes and benefits, we must compare the value of taxes paid by people during their working years with the benefits received in their retirement years. To compare the value of an amount of money at one date with that at a later date, we use the concept of present value.A present value is an amount of money that, if invested today, will grow to equal a given future amount when the interest that it earns is taken into account. We can compare dollars today with dollars in 2065 or any other future year by using present values.
For example, if the interest rate is 5 percent a year, $\$ 1,000$ invested today will grow, with interest, to $\$ 11,467$ after 50 years. So the present value (in 2015) of $\$ 11,467$ in 2065 is \$1,000.
By using present values, we can assess the magnitude of the government's debts to older Americans in the form of pensions and medical benefits.
But the assumed interest rate and growth rate of taxes and benefits critically influence the answers we get. For example, at an interest rate of 3 percent a year, the present value (in 2015) of $\$ 11,467$ in 2065 is $\$ 2,616$. The lower the interest rate, the greater isthe present value of a given future amount.
Because there is uncertainty about the proper interest rate to use to calculate present values, plausible alternative numbers are used to estimate a rangeof present values.

Using generational accounting and present values, economists have studied the situation facing the federal government arising from its Social Security obligations, and they have found a time bomb!

## The Social Security Time Bomb

When Social Security was introduced in the New Deal of the 1930s, today's demographic situation was not foreseen. The age distribution of the U.S. population today is dominated by the surge in the birth rate after World War II that created what is called the "baby boom generation." There are 77 million "baby boomers."
The first of the baby boomers started collecting Social Security pensions in 2008 and in 2011 they became eligible for Medicare benefits. By 2030, allthe baby boomers will have reached retirement ageand the population supported by Social Security andMedicare benefits will have doubled.
Under the existing laws, the federal government has an obligation to this increasing number of citizens to pay pensions and Medicare benefits on an already declared scale. These obligations are a debtowed by the government and are just as real as the bonds that the government issues to finance its cur-rent budget deficit.
To assess the full extent of the government's obligations, economists use the concept of fiscal imbalance. Fiscal imbalance is the present value of the government's commitments to pay benefits minus the present value of its tax revenues. Fiscal imbalance is an attempt to measure the scale of the government's true liabilities.
In an update, Gokhale estimates that the Social Security and Medicare fiscal imbalance was $\$ 68$ trillion in 2014. To put the $\$ 68$ trillion in perspective, note that U.S. GDP in 2014 was $\$ 17$ trillion. So the fiscal imbalance was 4 times the value of one year's production. Furthermore, the fiscal imbalance grows every year by an amount that in 2014 was approaching $\$ 2$ trillion.
These are enormous numbers and point to a catastrophic future. How can the federal government meet its Social Security obligations? Gokhale and Smetters consider four alternatives. They are

- Raise income taxes
- Raise Social Security taxes
- Cut Social Security benefits
- Cut federal government discretionary spending

Gokhale and Smetters estimate that if we had started in 2003 and made only one of these changes, income taxes would need to be raised by 69 percent, or Social Security taxes raised by 95 percent, or Social Security benefits cut by 56 percent. Even if the government stopped all its discretionary spending, including that on national defense, it would not be able to pay its bills. By combining the four measures, the pain from each could be lessened, but the pain would still be severe.
A further way of meeting these obligations is to pay by printing money. As you learned in Chapter 25 (see pp. 646-647), the consequence of this solution would be a seriously high inflation rate.

## Generational Imbalance

A fiscal imbalance must eventually be corrected andwhen it is, people either pay higher taxes or receivelower benefits. The concept of generational imbalancetells us who will pay.
Generational imbalance is the division of the fiscal imbalance between the current and future generations, assuming that the current generation will enjoy the existing levels of taxes and benefits.
Figure 30.8 on page 781 shows an estimate of how the fiscal imbalance is distributed across the current generation (those born before 1988) and the future generation (those born in or after 1988). The generational imbalance also shows that the major source of the imbalance is Medicare. Social Security pension benefits create a fiscal imbalance, but these benefits will be more than fully paid for by the current generation. But the current generation will not pay for all its Medicare costs, and the balance will fall on future generations. If we sum all the items, the current generation will pay 83 percent and future generations will pay 17 percent of the fiscalimbalance.
Because the estimated fiscal imbalance is solarge, it is not possible to predict how it will be resolved. But we can predict that the outcome will involve both lower benefits and higher taxes, or paying bills with new money and creating inflation.
The Fed would have to cooperate if inflation were to be used to deal with the imbalance, and this cooperation might be hard to obtain.

## International Debt

So far in our discussion of government deficits and debts, we've ignored the role played by the rest of the world. We'll conclude this discussion by considering the role and magnitude of international debt.
You've seen that borrowing from the rest of the world is one source of loanable funds. And you've also seen that this source of funds became larger during the late 1990s and 2000s. How large is the contribution of the rest of the world? How much business investment have we paid for by borrowing from the rest of the world? And how much U.S. government debt is held abroad?
Table 30.2 answers these questions. In June 2014, the United States had a net debt to the rest of the world of $\$ 11.7$ trillion. Of that debt, $\$ 5.8$ trillionwas U.S. government borrowing-about 48 percent of total U.S. government debt. U.S. corporations had used $\$ 8.6$ trillion of foreign funds (in bonds and equities).
The international debt of the United States is important because, when that debt is repaid, the United States will transfer real resources to the rest of the world. Instead of running a large net exportsdeficit, the United States will need a surplus ofexports over imports. To make a surplus possible, U.S. saving must increase and consumption must decrease. Some tough choices lie ahead.

| TABLE 30.2 What the United | ates Owed d in |
| :---: | :---: |
|  | \$ trillions |
| (a) U.S. liabilities |  |
| Deposits in U.S. banks | 1.2 |
| U.S. government securities | 5.8 |
| U.S. corporate bonds and equities | 8.6 |
| Other items (net) | -3.9 |
| Total | $\underline{\underline{11.7}}$ |
| (b) U.S. government securities |  |
| Held by rest of world | 5.8 |
| Held in the United States | 6.2 |
| Total | $\underline{12.0}$ |
| Source of data: Federol Reserve Board. |  |

## Fiscal Stimulus

The 2008-2009 recession brought Keynesian macroeconomic ideas (see Chapter 27, pp. 702-703) backinto fashion and put a spotlight on fiscal stimulus - the use of fiscal policy to increase production andemployment. But whether fiscal policy is truly stimulating, and if so, how stimulating, are questions that generate much discussion and disagreement. You'renow going to explore these questions.
Fiscal stimulus can be either automatic or discretionary. A fiscal policy action that is triggered by the state of the economy with no action by governmentis called automatic fiscal policy. The increase in total unemployment benefits triggered by the massive rise in the unemployment rate through 2009 is an example of automatic fiscal policy. A fiscal policy action initiated by an act of Congress is called discretionary fiscal policy. Itrequires a change in a spending program or in a tax law. A fiscal stimulus act passed by Congress in 2009 (see Economics in Action on p. 784) is an example of discretionary fiscal policy. Whether automatic or discretionary, an increase in government outlays or a decrease in government receipts can stimulate production and jobs. An increase in expenditure on goods and services directly increases aggregate expenditure. And an increase intransfer payments (such as unemployment benefits) or a decrease in tax revenues increases disposable income, which enables people to increase consumption expenditure. Lower taxes also strengthen the incentives to work and invest.
We'll begin by looking at automatic fiscal policy and the interaction between the business cycle and the budget balance.

## Automatic Fiscal Policy and Cyclical and Structural Budget Balances

Two items in the government budget change automatically in response to the state of the economy. They are tax revenues and needs-tested spending.

Automatic Changes in Tax Revenues The tax laws that Congress enacts don't legislate the number of tax dollars the government will raise. Rather they define the tax rates that people must pay. Tax dollars paid depend on tax rates and incomes. But incomes varywith real GDP, so tax revenues depend on real GDP. When real GDP increases in a business cycle expansion, wages and profits rise, so tax revenues from these incomes rise. When real GDP decreases in a recession, wages and profits fall, so tax revenues fall.

Needs-Tested Spending The government creates pro- grams that pay benefits to qualified people and businesses. The spending on these programs results in transfer payments that depend on the economic state of individual citizens and businesses. When theeconomy expands, unemployment falls, the number of people experiencing economic hardship decreases, so needs-tested spending decreases. When the economy is in a recession, unemployment is high and the number of people experiencing economic hardship increases, so needs-tested spending on unemployment benefits and food stamps increases.

Automatic Stimulus Because government receipts fall and outlays increase in a recession, the budget pro- vides automatic stimulus that helps to shrink the recessionary gap. Similarly, because receipts rise and outlays decrease in a boom, the budget provides automatic restraint to shrink an inflationary gap.

Cyclical and Structural Budget Balances To identify the government budget deficit that arises from the business cycle, we distinguish between the structural surplus or deficit, which is the budget balance that would occur if the economy were at full employment, and the cyclical surplus or deficit, which is the actual surplus or deficit minus the structural surplus or deficit. Figure 30.9 illustrates these concepts. Outlays decrease as real GDP increases, so the outlays curve slopes downward; and receipts increase as real GDP increases, so the receipts curve slopes upward.
In Fig. 30.9(a), potential GDP is $\$ 17$ trillion and if real GDP equals potential GDP, the government has a balanced budget. There is no structural surplus or deficit. But there might be a cyclical surplus or deficit. If real GDP is less than potential GDP at $\$ 16$ trillion, outlays exceed receipts and there is a cyclical deficit. If real GDP is greater than potential GDP at $\$ 18$ trillion, outlays are less than receipts and there is a cyclical surplus.
In Fig. 30.9(b), if potential GDP equals $\$ 17$ trillion (line $B$ ), the structural balance is zero. But if potential GDP is $\$ 16$ trillion (line $A$ ), the government budget has a structural deficit. And if potential GDP is $\$ 18$ trillion (line C), the government budget has a structural surplus.
U.S. Structural Budget Balance in 2014 The U.S. federal budget in 2014 was in deficit at $\$ 0.64$ trillionand the recessionary gap (the gap between real GDPand potential GDP) was $\$ 0.7$

FIGURE 30.9 Cyclical and Structural Surpluses and Deficits

(a) Cyclical deficit and cyclical surplus

(b) Structural deficit and structural surplus

In part (a), potential GDP is $\$ 17$ trillion. When real GDP is less than potential GDP, the budget is in deficit and it is a cyclical deficit. When real GDP exceeds potential GDP, the budget is in surplus and it is a cyclical surplus. The government has a balanced budget when real GDP equals potential GDP.

In part (b), if potential GDP is $\$ 16$ trillion, the deficit is a structural deficit and if potential GDP is $\$ 18$ trillion, the surplus is a structural surplus. If potential GDP is $\$ 17$ trillion, the budget is in structural balance.
trillion. With a large recessionary gap, you would expect some of the deficit to be cyclical. But how much of the 2014 deficit wascyclical and how much was structural?
The Congressional Budget Office (CBO) answers this question by analyzing the detailed items in the budget. According to the CBO, the cyclical deficit in 2014 was $\$ 0.18$ trillion and the structural deficit was $\$ 0.46$ trillion. Figure 30.10 shows the cyclical and structural deficit between 1990 and 2015.
You can see that the structural deficit was small in 2007, increased in 2008, and exploded in 2009. The 2009 fiscal stimulus package (see Economics in Action) created most of this structural deficit.
When full employment returns, which the CBO says will be in 2018, the cyclical deficit will vanish. But the structural deficit must be addressed by further actsof Congress. No one knows the discretionary measures that will be taken to reduce the structural deficit andthis awkward fact creates enormous uncertainty.

FIGURE 30.10 U.S. Cyclical and Structural Budget Balance


As real GDP shrank in the 2008-2009 recession, receipts fell, outlays increased, and the budget deficit increased. But the cyclical deficit is small compared to the actual deficit.

## ECONOMICS IN ACTION

## The 2009 fiscal Simuluts froctrage

Congress passed the American Recovery and Reinvestment Act of 2009 (the 2009 Fiscal Stimulus Act) in February 2009, and President Obama signed it into law at an economic forum he hosted in Denver. This act was the third and most ambitious in a series of stimulus packages and its purpose was to increase investment and consumer expenditure and lead to the creation of jobs.

The total package added $\$ 862$ billion to the federal government's budget deficit: $\$ 288$ billion from tax cuts and the rest from increased spending. The spending increases included payments to state and local governments ( $\$ 144$ billion), spending on infrastructure and science projects (\$111 billion), and programs in healthcare ( $\$ 59$ billion), education and training ( $\$ 53$ billion), and energy ( $\$ 43$ billion).


The Components of the 2009 Fiscal Stimulus Act

## Discretionary Fiscal Stimulus

Most discussion of discretionary fiscal stimulus focuses on its effects on aggregate demand. But you've seen(on pp. 775-777) that taxes influence aggregate sup-ply and that the balance of taxes and spending - the government budget deficit - can crowd out investment and slow the pace of economic growth. So discretionary fiscal stimulus has both supply-side and demand-side effects that end up determining its overall effectiveness.

We're going to begin our examination of discretionary fiscal stimulus by looking at its effects on aggregate demand.

Fiscal Stimulus and Aggregate Demand Changes in government expenditure and changes in taxes changeaggregate demand by their influence on spending plans, and they also have multiplier effects. Let's look at the two main fiscal policy multipliers: the government expenditure and tax multipliers.
The government expenditure multiplier is the quantitative effect of a change in government expenditure on real GDP. Because government expenditure is a component of aggregate expenditure, an increase in government spending increases aggregate expenditure and real GDP. But does a $\$ 1$ billion increase in government expenditure increase real GDP by $\$ 1$ billion, or more than $\$ 1$ billion, or less than $\$ 1$ billion? When an increase in government expenditure increases real GDP, incomes rise and the higher incomes bring an increase in consumption expenditure. If this were the only consequence of increased government expenditure, the government expenditure multiplier would be greater than 1.
But an increase in government expenditure increases government borrowing (or decreases government lending if there is a budget surplus) and raises the real interest rate. With a higher cost of borrowing, investment decreases, which partly offsets the increase in government spending. If this were the only consequence of increased government expenditure, the multiplier would be less than 1.
The actual multiplier depends on which of the above effects is stronger and the consensus is that the crowding-out effect is strong enough to make the government expenditure multiplier less than 1.

The tax multiplier is the quantitative effect of a change in taxes on real GDP. The demand-side effectsof a tax cut are likely to be smaller than an equivalentincrease in government expenditure. The reason is that a tax cut influences aggregate demand by increasing disposable income, only part of which gets spent. So the initial injection of expenditure from a $\$ 1$ billion tax cut is less than $\$ 1$ billion.
A tax cut has similar crowding-out consequences to a spending increase. It increases government borrowing (or decreases government lending), raises the real interest rate, and cuts investment.
The tax multiplier effect on aggregate demand depends on these two opposing effects and is probably quite small.

Graphical Illustration of Fiscal Stimulus Figure 30.11 shows how fiscal stimulus is supposed to work if it isperfectly executed and has its desired effects.
Potential GDP is $\$ 17$ trillion and real GDP is below potential at $\$ 16$ trillion so the economy has a recessionary gap of $\$ 1$ trillion. To restore full employment, the government passes a fiscal stimulus package. An increase in government expenditure and a tax cut increase aggregate expenditure by $\Delta E$. If this were the only change in spending plans, the $A D$ curve would shift rightward to become the curve labeled $A D_{0}+\Delta E$ in Fig. 30.11. But if fiscal stimulus sets off a multiplier process that increases consumption expenditure, and does
not crowd out much investment expenditure, aggregate demand increases further and the $A D$ curve shifts to $A D_{1}$.

FIGURE 30.11 Expansionary Fiscal Policy


Potential GDP is $\$ 17$ trillion, real GDP is $\$ 16$ trillion, and there is a $\$ 1$ trillion recessionary gap. An increase in government expenditure and a tax cut increase aggregate expenditure by $\Delta E$. The multiplier increases consumption expenditure. The $A D$ curve shifts rightward to $A D_{1}$, the price level rises to 115 , real GDP increases to $\$ 17$ trillion, and the recessionary gap is eliminated.
With no change in the price level, the economy would move from point $A$ to point $B$ on $A D_{1}$. But the increase in aggregate demand brings a rise in the pricelevel along the upward-sloping SAS curve and the economy moves to point $C$. At point C , the economy returns to full employment and the recessionary gap is eliminated.

Fiscal Stimulus and Aggregate Supply You've seen earlier in this chapter that taxes influence aggregate supply. A tax on labor income (on wages) drives a wedge between the cost of labor and the take-homepay of workers and lowers employment and output (p. 776). A tax on capital income (on interest) drives a wedge between the cost of borrowing and the return to lending and lowers saving and investment (p. 777).
With less saving and investment, the real GDP growth rate slows.
These negative effects of taxes on real GDP and its growth rate and on employment mean that a tax cut increases real GDP and its growth rate and increases employment.

These supply-side effects of a tax cut occur along with the demand-side effects and are probably much larger than the demand-side effects and make the overall tax multiplier much larger than the government expenditure multiplier-see Economics in Action. An increase in government expenditure financed by borrowing increases the demand for loanable funds and raises the real interest rate, which in turn lowersinvestment and private saving. This cut in investment is the main reason why the government expenditure multiplier is so small and why a deficit-financed increase in government spending ends up making only a small contribution to job creation. And because government expenditure crowds out investment, it lowers future real GDP.
So a fiscal stimulus package that is heavy on tax cuts and light on government spending works. But an increase in government expenditure alone is not an effective way to stimulate production and create jobs.
The description of the effects of discretionary fiscal stimulus and its graphical illustration in Fig. 30.11 make it look easy: Calculate the recessionary gap andthe multipliers, change government expenditure and taxes, and eliminate the gap. In reality, things are notthat easy. Getting the magnitude and the timing right is difficult, and we'll now examine this challenge.

Magnitude of Stimulus Economists havediverging views about the size of the government spending and tax multipliers because there is insufficient empirical evidence on which to pin their size with accuracy.
This fact makes it impossible for Congress to determine the amount of stimulus needed to close a given output gap. Further, the actual output gap is not known and can only be estimated with error. Forthese two reasons, discretionary fiscal policy is risky.

Time Lags Discretionary fiscal stimulus actions are also seriously hampered by three time lags:

- Recognition lag
- Law-making lag
- Impact lag

Recognition Lag is the time it takes to figure out that fiscal policy actions are needed. This process involves assessing the current state of the economy and forecasting its future state.

Law-Making Lagis the time it takes Congress to pass the laws needed to change taxes or spending. This process takes time because each member of Congress has a different idea about what is the best tax or spending program to change, so long debates and committee meetings are needed to reconcile conflicting views. The economy might benefit from fiscal stimulation today, but by the time Congress acts, a different fiscal medicine might be needed.
Impact Lag is the time it takes from passing a tax or spending change to its effects on real GDP being felt. This lag depends partly on the speed with which government agencies can act and partly on the timing of changes in spending plans by households and businesses. These changes are spread out over a number of quarters and possibly a number of years. Economic forecasting is steadily improving, but it remains inexact and subject to error. The range of uncertainty about the magnitudes of the spendingand tax multipliers make
discretionary fiscal stimulus an imprecise tool for boosting production and jobs and the crowding out consequences raise serious questions about its effects on long-term economic growth.

## ECONOMIC ANALYSIS

- Japan's fiscal policy challenge described in the news article is finding a way to stop and then reverse an ever rising government debt ratio-government debt as a percentage of GDP.
- Figure 1 shows the scale and upward direction of the problem. Gross debt was 250 percent of GDP in 2014, up from about 70 percent in 1990.
- The news article says that only faster growth or inflation can fix the debt problem. The debt ratio $=$ Debt $\div$ GDP. And GDP $=P Y$, where $P$ is the price level and $Y$ is real GDP, so the debt ratio $=$ Debt $\div$ PY. Faster growth increases $Y$ faster and inflation increases $P$ faster, either of which lowers the debt ratio.
- There is a third way to fix the debt problem: Cut government spending. And that is the only effective way. The reason is that high government spending crowds out investment and without an increase in investment, real GDP cannot grow faster.
- Figure 2 shows how government spending has crowded out investment. In 1990, investment at 34 percent of GDP was larger than government spending lexpenditure on goods and services and transfer payments) at 30 percent. By 2014, investment had shrunk to 20 percent and government spending had increased to 40 percent of GDP.
- The problem with cutting government spending is that in the short run, it decreases aggregate demand and widens the output gap.
- Figure 3 illustrates this short-run consequence of a government spending cut.
- In 2013, Japan's potential GDP was 537 trillion yen, shown by the long-run aggregate supply curve, LAS. The short-run aggregate supply curve was $S A S$.
- With the 2013 aggregate demand curve $A D_{0}$, real GDP was 525 trillion yen and the price level was 91 ( 91 percent of its 2009 level).
- Japan was experiencing stagnation and deflation.
- A fiscal policy aimed at cutting government expenditure and lowering the budget deficit by also increasing taxes would decrease aggregate demand.
- If aggregate demand decreased to $A D_{1}$, real GDP would decrease to 500 trillion yen and the recessionary gap would widen.
- Monetary policy might be used alongside fiscal policy to avoid this outcome and prevent real GDP from falling.
- Japan's government budget and debt is an extreme version of the United States'.


Figure 1 Japan's Government Debt


Figure 2 Crowding Out


Figure 3 Short-Run Effects of Budget Cuts

- Like Japan, the United States has an aging population that will bring persistently increasing government expenditure on healthcare and Social Security.
- The United States' challenge is to contain these expenditures and avoid crowding out investment and stopping growth.


## Chapter 31: Monetary Policy

After studying this chapter, you will able to:

- Describe the objectives of U.S. monetary policy and the framework for setting and achieving them
- Explain how the Federal Reserve makes its interest rate decision and achieves its interest rate target
- Explain the transmission channels through which the Federal Reserve influences real GDP, jobs, and inflation
- Explain the Fed's extraordinary policy actions

At eight regularly scheduled meetings a year and in an emergency between regular meetings, the Federal Reserve decides whether to change its interest rate target. How does the Fed make its interest rate decision? Can the Fed speed up economic growth by lowering the interest rate and can it keep inflation in check by raising the interest rate? What special measures can the Fed take in a financial crisis like the one that engulfed the U.S. and global economiesin 2008?
This chapter answers these questions and Economics in the News at the end of the chapter looks at the Fed's attempt to restore full employment.

## Monetary Policy Objectives and Framework

A nation's monetary policy objectives and the frame- work for setting and achieving those objectives stem from the relationship between the central bank and the government. We'll describe the objectives of U.S. monetary pol- icy and the framework and assignment of responsibility for achieving those objectives.

## Monetary Policy Objectives

The objectives of U.S. monetary policy are set out in the mandate of the Board of Governors of theFederal Reserve System, which is defined by the Federal Reserve Act of 1913 and its subsequent amendments, the most recent of which was passed in 2000.

Federal Reserve Act The Fed's mandate was most recently clarified in amendments to the Federal Reserve Act passed by Congress in 2000. The 2000law states that mandate in the following words:
The Board of Governors of the Federal Reserve System and the Federal Open Market Committee shall maintain long-run growth of the monetary and credit aggregates commensurate with the economy's long-run potential to increase production, so as to promote effectively the goals of maximum employment, stable prices, and moderate long-term interest rates.

Goals and Means This description of the Fed's monetary policy objectives has two distinct parts: a statement of the goals, or ultimate objectives, and aprescription of the means by which the Fed should pursue its goals.

Goals of Monetary Policy The goals are "maximum employment, stable prices, and moderate long-term interest rates." In the long run, these goals are in harmony and reinforce each other. But in the short run, these goals might come into conflict. Let's examine these goals a bit more closely.
Achieving the goal of "maximum employment" means attaining the maximum sustainable growthrate of potential GDP and keeping real GDP close to potential GDP. It also means keeping the unemployment rate close to the natural unemploymentrate.
Achieving the goal of "stable prices" means keeping the inflation rate low (and perhaps close to zero).
Achieving the goal of "moderate long-term interest rates" means keeping long-term nominal interest rates close to (or even equal to) long-term real interestrates.
Price stability is the key goal. It is the source of maximum employment and moderate longterm interest rates. Price stability provides the best avail- able environment for households and firms to make the saving and investment decisions that bring economic growth. So price stability encourages the maxi- mum sustainable growth rate of potential GDP. Price stability delivers moderate long-term interest rates because the nominal interest rate reflects the inflation rate. The nominal interest rate equals the real interest rate plus the inflation rate. With stable prices, the nominal interest rate is close to the real interest rate, and most of the time, this rate is likely to be moderate.
In the short run, the Fed faces a tradeoff between inflation and interest rates and between inflation and ; real GDP, employment, and unemployment. Taking ' an action that is designed to lower the inflation rate and achieve stable prices might mean raising interest rates, which lowers employment and real GDP and increases the unemployment rate in the short run.

Means for Achieving the Goals The 2000 law instructs the Fed to pursue its goals by "maintain[ing]long-run growth of the monetary and credit aggregates commensurate with the economy's long-run potential to increase production." You perhaps recognize this statement as being consistent with the quantity theory of money that you studied in Chapter 25 (see pp. 646-647). The "economy's long-run potential to increase production" is the growth rate of potential GDP. The "monetary and credit aggregates" are the quantities of money and loans. By keeping the growth rate of the quantity of money in line with the growth rate of potential GDP, the Fed is expected to be able to maintain full employment and keep the price level stable.
To pursue the goals of monetary policy, the Fed must make the general concepts of price stability and maximum employment precise and operational.

## Operational "Stable Prices" Goal

The Fed pays attention to two measures of inflation: the Consumer Price Index (CPI) and the personal consumption expenditure (PCE) deflator. But the core PCB deflator, which excludes food and fuel prices, is the Fed's operational guide and the Fed defines the rate of increase in the core PCE deflator as the core inflation rate.

The Fed focuses on the core inflation rate because it is less volatile than the total CPI inflation rate and the Fed believes that it provides a better indication of whether price stability is being achieved.
The Fed has not defined price stability, but it almost certainly doesn't regard it as meaning a core inflation rate equal to zero.
A former Fed Chairman, Alan Greenspan, sug-gested that "price stability is best thought of as an environment in which inflation is so low and stable over time that it does not materially enter into the decisions of households and firms." He also believes that a "specific numerical inflation target would represent an unhelpful and false precision." Ben Bernanke, Alan Greenspan's successor, was more precise and suggested that a core inflation rate of between 1 and 2 percent a year is the equivalent of price stability. This inflation range came to be known as the Fed's "comfort zone."
Figure 31.1 shows the core inflation rate since 2000 along with the Fed's comfort zone. You can see that most of the time, the Fed has kept the core inflation rate inside its comfort zone. But between 2004 and 2008, a period during which a major financial crisis occurred, inflation was above its comfort zone.

## Operational Maximum Employment Goal

The Fed regards stable prices (a core inflation rate of 1 to 2 percent a year) as the primary goal of monetary policy and as a means to achieving the other two goals. But the Fed also pays attention to the business cycle and tries to steer a steady course between inflation and recession. To gauge the state of output and employment relative to full employment, the Fed looks at a large number of indicators that include the labor force participation rate, the unemployment rate, measures of capacity utilization, activity in the housing market, the stock market, and regional information gathered by the regional Federal Reserve Banks. All these data that describe the current state of the economy are summarized in the Fed's Beige Book.

While the Fed considers a vast range of data, one number stands out as a summary of the overall state of aggregate demand relative to potential GDP. That number is the output gap-the percentage deviation of real GDP from potential GDP.
When the output gap is positive, it is an inflationary gap that brings an increase in the inflation rate. And when the output gap is negative, it is a recessionary gap that results in lost output and in employment being below its full-employment equilibrium level. So the Fed tries to minimize the output gap.

## Responsibility for Monetary Policy

Who is responsible for monetary policy in the UnitedStates? What are the roles of the Fed, Congress, andthe president?

The Role of the Fed The Federal Reserve Act makes the Board of Governors of the Federal ReserveSystem and the Federal Open Market Committee (FOMC) responsible for the conduct of monetary policy. We described the composition of the FOMC in Chapter 25 (see p. 635). The FOMC makes a monetary policy decision at eight scheduled meetings each year and
communicates its decision with a brief explanation. Three weeks after an FOMC meeting, the full minutes are published.

The Role of Congress Congress plays no role in making monetary policy decisions but the Federal Reserve Act requires the Board of Governors to report on monetary policy to Congress. The Fed makes two reports each year, one in February and another in July. These reports and the Fed chairman's testimony before Congress along with the minutes of the FOMC communicate the Fed's thinking on monetary policy to lawmakers and the public.

The Role of the President The formal role of the president of the United States is limited to appointing the members and the chairman of the Board of Governors. But some presidents - Richard Nixon was one - have tried to influence Fed decisions.

FIGURE 31.1 Operational Price Stability Goal: Core Inflation


The core inflation rate-based on the core PCE deflatorwas inside the Fed's comfort zone between 2000 and 2004 and after 2008 but above the comfort zone upper limit between 2004 and 2008.

You now know the objectives of monetary policy and can describe the framework and assignment of responsibility for achieving those objectives. Your next task is to see how the Federal Reserve conducts its monetary policy.

## The Conduct of Monetary Policy

How does the Fed conduct its monetary policy? This question has two parts:

- What is the monetary policy instrument?
- How does the Fed make its policy decisions?


## The Monetary Policy Instrument

A monetary policy instrument is a variable that the Fed can directly control or at least very closely target. The Fed has two possible instruments: the monetary base or the interest rate at which banks borrow and lend monetary base overnight.
The Fed's choice of monetary policy instrument is the interest rate at which the banks make overnight loans to each other. The market in which the banks borrow and lend overnight is called the federal funds market and the interest rate in that market is called the federal funds rate.
Figure 31.2 shows the federal funds rate from 2000 to 2014 . You can see that the federal funds rateranges between a high of 6.8 percent a year and a low of 0.2 percent a year. In 2000 and 2006, when the federal funds rate was high, the Fed's actions wereaimed at lowering the inflation rate.

Between 2002 and 2004 and again in and since 2008, the federal funds rate was set at historically low levels. During these years, inflation was well anchored at close to or below 2 percent a year, and the Fed was less concerned about inflation than it was about recession and high unemployment. So the Fed set a low interest rate to fight recession.
Although the Fed can change the federal funds rate by any (reasonable) amount that it chooses, it normally changes the federal funds rate by only a quarter of a percentage point. Having decided the appropriate level for the federal funds rate, how does the Fed move the rate to its target level? The answer is by using open-market operations (see Chapter 25, pp. 636-638) to adjust the quantity of monetary base.
To see how an open market operation changes the federal funds rate, we need to examine the federal funds market and the market for bank reserves. In the federal funds market, the higher the federal funds rate, the greater is the quantity of overnight loans supplied and the smaller is the quantity of overnight loans demanded. The equilibrium federal funds rate balances the quantities demanded and supplied.
figure 31.2 The Federal Funds Rate


The Fed sets a target for the federal funds rate and then takes actions to keep the rate close to its target. When the Fed wants to slow inflation, it takes actions that raise the federal funds rate. When inflation is low and the Fed wants to avoid recession, it takes actions that lower the federal funds rate.

Source of data: Board of Governors of the Federal Reserve System.
An equivalent way of looking at the forces that determine the federal funds rate is to consider the demand for and supply of bank reserves. Banks hold reserves to meet the required reserve ratio and so that they can make payments. But reserves are costly to hold because they can be loaned in the federal funds market and earn the federal funds rate. So the higher the federal funds rate, the smaller is the quantity of reserves demanded.
Figure 31.3 illustrates the demand for bank reserves. The $x$-axis measures the quantity of reserves that banks hold on deposit at the Fed, and the $y$-axismeasures the federal funds rate. The demand for reserves is the curve labeled $R D$.
The Fed's open market operations determine the supply of reserves, which is shown by the supply curve $R S$. Equilibrium in the market for bankreserves determines the federal funds rate where the quantity of reserves demanded by the banks equals the quantity of reserves supplied by the Fed. By using open market operations, the Fed adjusts the supply of reserves to keep the federal funds rate on target.
Next, we see how the Fed makes its policy decisions.

FIGURE 31.3 The Market for Reserves


The demand curve for reserves is $R D$. The quantity of reserves demanded decreases as the federal funds rate rises because the federal funds rate is the opportunity cost of holding reserves. The supply curve of reserves is $R S$. The Fed uses open market operations to make the quantity of reserves supplied equal the quantity of reserves demanded ( $\$ 50$ billion in this case) at the federal funds rate target (5 percent a year in this case).

The Fed's Decision-Making Strategy
The Fed's decision making begins with the Beige Book exercise described in Economics in Action on the next page. The Fed then turns to forecasting three key variables: the inflation rate, the unemployment rate, and the output gap.

Inflation Rate The Fed's forecasts of the inflation rate are a crucial ingredient in its interest rate decision. Ifinflation is above or is expected to move above the top of the comfort zone, the Fed considers raising the federal funds rate target; and if inflation is below or is expected to move below the bottom of the comfort zone, it considers lowering the interest rate.

Unemployment Rate The Fed monitors and forecasts the unemployment rate and its relation to the natural unemployment rate (see Chapter 22, pp. 559-561). If the unemployment rate is below the natural rate, a labor shortage might put upward pressure on wage rates, which might feed through to increase the inflation rate. So a higher interest rate might be called for. If the unemployment rate is above the natural rate, a lower inflation rate is expected, which indicates the need for a lower interest rate.

Output Gap The Fed monitors and forecasts real GDP and potential GDP and the gap between them, the output gap (see Chapter 27, pp. 698-699). If the output gap is positive, an inflationary gap, the inflation rate will most likely accelerate, so a higher interest rate might be required. If the output gap is negative, a recessionary gap, inflation might ease, which indicates room to lower the interest rate. We next look at the transmission of monetary policy and see how it achieves its goals.

## ECONOMICS IN ACTION

## FOMC Decision Making

The Fed's decision making begins with an intensive assessment of the current state of the economy, which is conducted by the Federal Reserve districts and summarized in the Beige Book. Today, the Beige Book is a web posting at http://www.federalreserve.gov/ monetarypolicy/beigebook/default.htm (see the screenshot).
The FOMC then turns its attention to the likely near-future evolution of the economy and the interest rate change that will keep inflation in check and the economy expanding at close to full employment. In making this assessment, the FOMC pays close attention to the inflation rate, the unemployment rate, and the output gap.

Balancing the signals that it gets from monitoring the three main features of macroeconomic performance, the FOMC meets in its imposing room (see the photo) and makes a decision on whether to change its federal funds rate target and if so, what the new target should be.

Having decided on the appropriate target for the federal funds rate, the FOMC instructs the New York Fed to conduct open market operations aimed at hitting the federal funds rate target.
If the goal is to raise the federal funds rate, the New York Fed sells securities in the open market. If the goal is to lower the federal funds rate, the New York Fed buys securities in the open market.

## Beige Book

Summary of Commentary on Current Economic Conditions by Federal Reserve District

Commonly known as the Beige Book this report is published eight times per year. Each Federal Reserve Bank gathers anecdotal information on current economic conditions in its District through reports from Bank and 8ranch directors and interviews with key business contacts. economists market experts. and other sources. The Beige Book summarizes this information by District and sector An overall summary of the twelve district reports is prepared by a designated Federal Reserve Bank on a rotating basis.

| January | 15 | HTM I P PDE |
| :---: | :---: | :---: |
| March | 05 | HTML I PDF |
| April | 16 | HTME I PDF |
| June | 04 | HTM I PDF |
| July | 16 | HTU \| PDF |
| September | 03 | HTM : PDF |
| October | 15 |  |
| December | 03 |  |



## Monetary PolicyTransmission

You've seen that the Fed's goal is to keep the price level stable (keep the inflation rate around 2 percent a year) and to achieve maximum employment (keep the output gap close to zero). And you've seen how the Fed can use its power to set the federal funds rate at its desired level. We're now going to trace the events that follow a change in the federal funds rate and see how those events lead to the ultimate policy goal. We'll begin with a quick overview of the trans- mission process and then look at each step a bit more closely.

## Quick Overview

When the Fed lowers the federal funds rate, the Fed securities in an open market operation and other short-term interest rates and the exchange rate also fall. The quantity of money and the supply of loan-able funds increase. The long-term real interest ratefalls. The lower real interest rate increases consumption expenditure and investment. And the lower exchange rate makes U.S. exports cheaper and imports more costly, so net exports increase. Easierbank loans reinforce the effect of lower interest rates on aggregate expenditure. Aggregate demand increases, which increases real GDP and the price level relative to what they would have been had theFed not lowered the federal funds rate. Real GDP growth and inflation speed up. When the Fed raises the federal funds rate, the Fed sells securities in an open market operation and as the sequence of events that we've just reviewed plays out, the effects are in the opposite directions.
Figure 31.4 provides a schematic summary of these ripple effects for both a cut and a rise in the federal funds rate.
These ripple effects stretch out over a period of between one and two years. The interest rate and exchange rate effects are immediate. The effects on money and bank loans follow in a few weeks and run for a few months. Real long-term interest rates change quickly and often in anticipation of the short- term interest rate changes. Spending plans change and real GDP growth changes after about one year. The inflation rate changes between one year and two years after the change in the federal funds rate. But these time lags are not entirely predictable and can be longer or shorter.
We're going to look at each stage in the transmission process, starting with the interest rate effects.

## Interest Rate Changes

The first effect of a monetary policy decision by the FOMC is a change in the federal funds rate. To achieve that, the Fed conducts an open marketoperation. Other interest rates then change. These interest rate effects occur quickly and relatively predictably.
Figure 31.5 shows the fluctuations in three interest rates: the federal funds rate, the 3month Treasury bill rate, and the long-term bond rate.

Federal Funds Rate As soon as the FOMC announces a new setting for the federal funds rate, the New York Fed undertakes the necessary open market operations to hit the target. There is no doubt about where the interest rate changes shown in Fig. 31.5 are generated. They are driven by the Fed's monetary policy.

FIGURE 31.4 The Ripple Effects of a Change in the Federal Funds Rate


Short-Term Bill Rate The short-term bill rate isthe interest rate paid by the U.S. government on 3- month Treasury bills. It is similar to the interest rate paid by U.S. businesses on shortterm loans. Noticehow closely the 3-month Treasury bill rate follows the federal funds rate. The two rates are almost identical.
A powerful substitution effect keeps these two interest rates close. Commercial banks have a choice about how to hold their short-term liquid assets, and an overnight loan to another bank is a close substitute for short-term securities such as Treasury bills. If the interest rate on Treasury bills is higher than the federal funds rate, the quantity of overnight loans supplied decreases and the demand for Treasury bills increases. The price ofTreasury bills rises and the interest rate falls.
Similarly, if the interest rate on Treasury bills is lower than the federal funds rate, the quantity of overnight loans supplied increases and the demand for Treasury bills decreases. The price of Treasury bills falls, and the interest rate rises. When the interest rate on Treasury bills is close to the federal funds rate, there is no incentive for a bank to switch between making
an overnight loan and buying Treasury bills. Both the Treasury bill market and the federal funds market are in equilibrium.

FIGURE 31.5 Three Interest Rates


The short-term interest rates-the federal funds rate and the 3 -month Treasury bill rate-move closely together. The longterm bond rate is higher than the short-term rates, and it fluctuates less than the short-term rates.

The Long-Term Bond Rate The long-term bond rate is the interest rate paid on bonds issued by large corporations. It is this interest rate that businesses pay on the loans that finance their purchase of new capital and that influences their investment decisions.
Two features of the long-term bond rate stand out: It is higher than the short-term rates, and it fluctuates less than the short-term rates.
The long-term interest rate is higher than the two short-term rates because long-term loans are riskier than short-term loans. To provide the incentive that brings forth a supply of long-term loans, lenders must be compensated for the additional risk. Without compensation for the additional risk, only short-term loans would be supplied. The long-term interest rate fluctuates less than the short-term rates because it is influenced by expectations about future short-term interest rates as well as current shortterm interest rates. The alternative to borrowing or lending long term is to borrow or lend using a sequence of short-term securities. If the long-term interest rate exceeds the expected average of future short-term interest rates, people will lend long term and borrow short term. The long-term interest rate will fall. And if the long-term interest rate is below the expected average of future short-term interest rates, people will borrow long term and lend short term. The long-term interest rate will rise.

These market forces keep the long-term interest rate close to the expected average of future short-term interest rates (plus a premium for the extra risk associated with long-term loans). The expected average future short-term interest rate fluctuates less than the current shortterm interest rate

## Exchange Rate Fluctuations

The exchange rate responds to changes in the interest rate in the United States relative to the interest rates $m$ other countries-the U.S. interest rate differential. Weexplain this influence in Chapter 26 (see p. 663).
When the Fed raises the federal funds rate, the interest rate differential rises and, other things remaining the same, the U.S. dollar appreciates, andwhen the Fed lowers the federal funds rate, the U.S. interest rate differential falls and, other things remaining the same, the U.S. dollar depreciates.
Many factors other than the U.S. interest rate differential influence the exchange rate, so when the Fed changes the federal funds rate, the exchange rate does not usually change in exactly the way it would with other things remaining the same. So while monetary policy influences the exchange rate, many other factors also make the exchange rate change.

## Moneyand Bank Loans

The quantity of money and bank loans change when the Fed changes the federal funds rate target. A rise in the federal funds rate decreases the quantity of money and bank loans, and a fall in the federal funds rate increases the quantity of money and bank loans. These changes occur for two reasons: The quantity of deposits and loans created by the banking system changes and the quantity of money demanded changes.
You've seen that to change the federal funds rate, the Fed must change the quantity of bank reserves. A change in the quantity of bank reserves changes the monetary base, which in turn changes the quantity of deposits and loans that the banking system can create. A rise in the federal funds rate decreases reserves and decreases the quantity of deposits and bank loans created; and a fall in the federal funds rate increases reserves and increases the quantity of deposits and bank loans created.
The quantity of money created by the banking sys- tem must be held by households and firms. The change in the interest rate changes the quantity of money demanded. A fall in the interest rate increases the quantity of money demanded, and a rise in the interest ratedecreases the quantity of money demanded.
A change in the quantity of money and the supply of bank loans directly affects consumption and investment plans. With more money and easier access to loans, consumers and firms spend more. With less money and loans harder to get, consumers and firms spend less.

## The Long-Term Real Interest Rate

Demand and supply in the market for loanable funds determine the long-term real interest rate, which equals the long-term nominal interest rate minus the expected inflation rate. The long-term real interest rate influences expenditure decisions.

In the long run, demand and supply in the loan- able funds market depend only on real forces-on saving and investment decisions. But in the short run, when the price level is not fully flexible; the sup- ply of loanable funds is influenced by the supply of bank loans. Changes in the federal funds rate change the supply of bank loans, which changes the supply of loanable funds and changes the interest rate in theloanable funds market. A fall in the federal funds rate that increases the supply of bank loans increases the supply of loanable funds and lowers the equilibrium real interest rate. A rise in the federal funds rate that decreases the supply of bank loans decreases the supply of loanable funds and raises the equilibrium real interest rate.
These changes in the real interest rate, along with the other factors we've just described, change expenditure plans.

## Expenditure Plans

The ripple effects that follow a change in the federal funds rate change three components of aggregate expenditure:

- Consumption expenditure
- Investment
- Net exports

Consumption Expenditure Other things remaining the same, the lower the real interest rate, the greater is the amount of consumption expenditure and the smaller is the amount of saving.

Investment Other things remaining the same, the lower the real interest rate, the greater is the amount of investment.

Net Exports Other things remaining the same, the lower the interest rate, the lower is the exchange rate and the greater are exports and the smaller are imports.
So eventually, a cut in the federal funds rate increases aggregate expenditure and a rise in the federal funds rate curtails aggregate expenditure. These changes in aggregate expenditure plans change aggregate demand, real GDP, and the price level.

## The Change in Aggregate Demand, Real GDP, and the Price Level

The final link in the transmission chain is a change in aggregate demand and a resulting change in real GDP and the price level. By changing real GDP and the price level relative to what they would have been without a change in the federal funds rate, the Fed influences its ultimate goals: the inflation rate and the output gap.

## The Fed Fights Recession

If inflation is low and real GDP is below potential GDP, the Fed takes actions that are designed to restore full employment. Figure 31.6 shows the effects of the Fed's actions, starting in the market for bank reserves and ending in the market for real GDP.

Market for Bank Reserves In Fig. 3I.6(a), which shows the market for bank reserves, the FOMC lowers the target federal funds rate from 5 percent to 4 percent a year. To achieve the new target, the NewYork Fed buys securities and increases the supply of reserves of the banking system from $R S_{0}$ to $R S_{1}$.
Money Market With increased reserves, the banks create deposits by making loans and the supply of money increases. The short-term interest rate falls and the quantity of money demanded increases. In Fig. 31.6(b), the supply of money increases from $M S_{0}$ to $M S_{1}$, the interest rate falls from 5 percent to 4 percent a year, and the quantity of money increases from $\$ 3$ trillion to $\$ 3.1$ trillion. The interest rate in the money market and the federal funds rate are kept close to each other by the powerful substitution effect described on $p$. 802.

Loanable Funds Market Banks create money by making loans. In the long run, an increase in the sup- ply of bank loans is matched by a rise in the price level and the quantity of real loans is unchanged. But in the short run, with a sticky price level, an increase in the supply of bank loans increases the supply of (real) loanablefunds.
figure 31.6 The Fed Fights Recession

(a) The market for bank reserves

In part (a), the FOMC lowers the federal funds rate target from 5 percent to 4 percent. The New York Fed buys securities in an open market operation and increases the supply of reserves from $R S_{0}$ to $R S_{1}$ to hit the new federal funds rate target.

(b) Money market

In part (b), the supply of money increases from $M S_{0}$ to $M S_{1}$, the short-term interest rate falls, and the quantity of money demanded increases. The short-erm interest rate and the federal funds rate change by similar amounts.

In Fig. 31.6(c), the supply of loanable funds curve shifts rightward from $S L F_{0}$ to $S L F_{1}$. With the demand for loanable funds at $D L F$, the real interest rate falls from 6 percent to 5.5 percent a year. (We're assuming a zero inflation rate so that the real interest rate equals the nominal interest rate.) The long-term interest rate changes by a smaller amount than the change in the short-term interest rate for the reason explained on p. 802.

The Market for Real GDP Figure 31.6(d) shows aggregate demand and aggregate supply-the demand for and supply of real GDP. Potential GDP is $\$ 16$ trillion, where LAS is located. The short-run aggregate supply curve is SAS, and initially, the aggregate demand curve is $A D_{0}$. Real GDP is $\$ 15.8$ trillion, which is less than potential GDP, so there is a recessionary gap. The Fed is reacting to this recessionary gap.
The increase in the supply of loans and the decrease in the real interest rate increase aggregate planned expenditure. (Not shown in the figure, a fall in the interest rate lowers the exchange rate, which increases net exports and aggregate planned expenditure.) The increase in aggregate expenditure, $\Delta E$, increases aggregate demand and shifts the aggregate demand curve rightward to $A D_{0}+\Delta E$. A multiplier process begins. The increase in expenditure increases income, which induces an increase in consumption expenditure. Aggregate demand increases further, and the aggregate demand curve eventually shifts rightward to $A D_{1}$.
The new equilibrium is at full employment. Real GDP is equal to potential GDP. The price level rises to 110 and then becomes stable at that level. So after a one-time adjustment, there is price stability.
In this example, we have given the Fed a perfect hit at achieving full employment and keeping the price level stable. Itis unlikely that the Fed would be able to achieve the precision of this example. If the Fed stimulated. demand by too little and too late, the economy would experience a recession. And if the Fed hit the gas pedal too hard, it would push the economy from recession to inflation.

(c) The market for loanable funds

In part (c), the increase in the quantily of money increases the supply of bank loans. The supply of loanable funds increases and shiffs the supply curve from $S L F_{0}$ to $S L F_{1}$. The real interest rate falls and investment increases.

(d) Real GDP and the price level

In part (d), the increase in investment increases aggregate planned expenditure. The aggregate demand curve shifts to $A D_{0}+\Delta E$ and eventually it shifts rightward to $A D_{1}$. Real GDP increases to potential GDP, and the price level rises.

## The Fed Fights Inflation

If the inflation rate is too high and real GDP is above potential GDP, the Fed takes actions that are designed to lower the inflation rate and restore price stability. Figure 31.7 shows the effects of the Fed's actions starting in the market for reserves and ending in the market for real GDP.

Market for Bank Reserves In Fig. 31.7(a), which shows the market for bank reserves, the FOMC raises the target federal funds rate from 5 percent to 6 per-cent a year. To achieve the new target, the New YorkFed sells securities and decreases the supply of reserves of the banking system from $R S_{0}$ to $R S_{1}$.

Money Market With decreased reserves, the banks shrink deposits by decreasing loans and the supply ofmoney decreases. The short-term interest rate rises and the quantity of money demanded decreases. In Fig. 31.7(b), the supply of money decreases from $M S 0$ to $M S_{1}$, the interest rate rises from 5 percent to 6 percent a year, and the quantity of money decreases from $\$ 3$ trillion to $\$ 2.9$ trillion.

Loanable Funds Market With a decrease in reserves, banks must decrease the supply of loans. The supply of (real) loanable funds decreases, and the supply ofloanable funds curve shifts leftward in Fig. 31.7(c) from $S L F_{0}$ to $S L F_{1}$. With the demand for loanable funds at DLF, the real interest rate rises from 6 per-cent to 6.5 percent a year. (Again, we're assuming a zero inflation rate so that the real interest rate equals the nominal interest rate.)

The Market for Real GDP Figure 31.7(d) shows aggregate demand and aggregate supply in the marketfor real GDP. Potential GDP is $\$ 16$ trillion where LAS is located. The short-run aggregate supply curve is $S A S$ and initially the aggregate demand is $A D_{0}$.
Now, real GDP is $\$ 16.2$ trillion, which is greater than potential GDP, so there is an inflationary gap. The Fed is reacting to this inflationary gap.

The increase in the short-term interest rate, the decrease in the supply of bank loans, and the increase in the real interest rate decrease aggregate planned expenditure. (Not shown in the figures, a rise in the interest rate raises the exchange rate, which decreases net exports and aggregate planned expenditure.)
The decrease in aggregate expenditure, $M$, decreases aggregate demand and shifts the aggregate demand curve to $A D_{0}-\Delta E$. A multiplier process begins. The decrease in expenditure decreases income, which induces a decrease in consumption expenditure. Aggregate demand decreases further, and the aggregate demand curve eventually shifts leftward to $\mathrm{AD}_{1}$.
The economy returns to full employment. Real GDP is equal to potential GDP. The price level falls to 110 and then becomes stable at that level. So after a one-time adjustment, there is price stability. Again, in this example, we have given the Fed a perfect hit at achieving full employment and keeping the price level stable. If the Fed decreased aggregate demand by too little and too late, the economy would have remained with an inflationary gap and the inflation rate would have moved above the rate that is consistent with price
stability. And if the Fed hit the brakes too hard, it would push the economy from inflation torecession.

FIGURE 31.7 The Fed Fights Inflation

(a) The market for bank reserves

In part (a), the FOMC raises the federal funds rate from 5 percent to 6 percent. The New York Fed sells securities in an open market operation to decrease the supply of reserves from $R S_{0}$ to $R S_{1}$ and hit the new federal funds rate target.

(b) Money market

In part (b), the supply of money decreases from $M S_{0}$ to $M S_{1}$, the short-term interest rate rises, and the quantity of money demanded decreases. The short-term interest rate and the federal funds rate change by similar amounts.

## Loose Links and Long and Variable Lags

The ripple effects of monetary policy that we've justanalyzed with the precision of an economic model are, in reality, very hard to predict and anticipate.
To achieve price stability and full employment, the Fed needs a combination of good judgment and good luck. Too large an interest rate cut in an underemployed economy can bring inflation, as it did during the 1970s. And too large an interest rate rise in an inflationary economy can create unemployment, as it did in 1981 and 1991. Loose links between the federal funds rate and the ultimate policy goals make unwanted outcomes inevitable and long and variable time lags add to the Fed's challenges.

Loose Link from federal Funds rate to Spending The real long-term interest rate that influences spending plans is linked only loosely to the federal funds rate. Also, the response of the reallong-term interest rate to a change in the nominal interest rate depends on how inflation expectations change. And the response of expenditure plans to changes in the real interest rate depend on many factors that make the response hard to predict.

Time Lags in the Adjustment Process The Fed is especially handicapped by the fact that the monetary policytransmission process is long and drawn out. Also, theeconomy does not always respond in exactly the sameway to a policy change. Further, many factors otherthan policy are constantly changing and bringing newsituations to which policy must respond.

(c) The market for loanable funds

In part (c), the decrease in the quantity of money decreases the supply of bank loans. The supply of loanable funds decreases and the supply curve shifts from $S L F_{0}$ to $S L F_{1}$. The real interest rate rises and investment decreases.

(d) Real GDP and the price level

In part (d), the decrease in investment decreases aggregate planned expenditure. Aggregate demand decreases and the $A D$ curve shifts leffward from $A D_{0}$ fo $A D_{1}$. Real GDP decreases to potential GDP, and the price level falls.

## ECONOMICS IN ACTION

## A View of the Long and Variable Lag

You've studied the theory of monetary policy. Does it really work in the way we've described? It does, and the figure opposite provides some evidence to support this claim.

The blue line in the figure is the federal funds rate that the Fed targets minus the long-term bond rate. (When the long-term bond rate exceeds the federal funds rate, this gap is negative.)

We can view the gap between the federal funds rate and the long-term bond rate as a measure of how hard the Fed is trying to steer a change in the economy's course.

When the Fed is more concerned about recession than inflation and is trying to stimulate real GDP growth, it cuts the federal funds rate target and the gap between the long-term bond rate and the federal funds rate widens.

When the Fed is more concerned about inflation than recession and is trying to restrain real GDP growth, it raises the federal funds rate target and the gap between the long-term bond rate and the federal funds rate narrows.

The red line in the figure is the real GDP growth rate two years later. You can see that when the FOMC raises the federal funds rate, the real GDP growth rate slows two years later. And when the Fed lowers


Interest Rates and Real GDP Growth
Sources of data: Interest rates, see Fig. 31.5; real GDP growth rate, Bureau of Economic Analysis.
the federal funds rate, the real GDP growth rate speeds up two years later.

Not shown in the figure, the inflation rate increases and decreases corresponding to the fluctuations in the real GDP growth rate. But the effects on the inflation rate take even longer and are not as strong as the effects on the real GDP growth rate.

## Extraordinary Monetary Stimulus

During the financial crisis and recession of 2008-2009, the Fed lowered the federal funds rate target to the floor. The rate can't go below zero, so what can the Fed do to stimulate the economy when it can't lower the interest rate any further?
The Fed has answered this question with some extraordinary policy actions. To understand those actions, we need to dig a bit into the anatomy of the financial crisis to which the Fed is responding. That'swhat we'll now do. We'll look at the key elements in the financial crisis and then look at the Fed's response.

## The Key Elements of the Crisis

We can describe the crisis by identifying the events that changed the values of the assets and liabilities of banks and other financial institutions.
Figure 31.8 shows the stylized balance sheet of a bank: deposits plus equity equals reserves plus loans and securities (see Chapter 25, p. 632). Deposits and own capital - equity - are the bank's sources of funds (other borrowing by banks is ignored here).
Deposits are the funds loaned to the bank by house- holds and firms. Equity is the capital provided by the bank's stockholders and includes the bank's undistributed profits (and losses). The bank's reserves are currency and its deposit at the Fed. The bank's loans and securities are the loans made by the bank and government bonds, private bonds, assetbacked bonds, and other securities that the bank holds.

Three main events can put a bank under stress:

1. Widespread fall in asset prices
2. A significant currency drain
3. A run on the bank

Figure 31.8 on page 810 summarizes the problems that each event presents to a bank. A widespread fall in asset means that the bank suffers a capital loss. It must write down the value of its assets and the value of the bank's equity decreases by the same amount as the fall in the value of its securities. If the fall in asset prices is large enough, the bank's equity might fall to zero, in which case the bank is insolvent. It fails.
A significant currency drain means that depositors withdraw funds and the bank loses reserves. This event puts the bank in a liquidity crisis. It is short of cash reserves. A run on the bank occurs when depositors lose confidence in the bank and massive withdrawals of deposits occur. The bank loses reserves and must call in loans and sell off securities at unfavorable prices. Its equity shrinks.
The red arrows in Fig. 31.8 summarize the effects of these events and the problems they brought in the 2007-2008 financial crisis. A wide-spread fall in asset prices was triggered by the bursting of a house-price bubble that saw house prices switch from rapidly rising to falling. With falling house prices, sub-prime mortgage defaults occurred and the prices of mortgage-backed securities and derivatives whose values are based on these securities began to fall.
People with money market mutual fund deposits began to withdraw them, which created a fear of a massive withdrawal of these funds analogous to arun on a bank. In the United Kingdom, one bank,Northern Rock, experienced a bank run.
With low reserves and even lower equity, banks turned their attention to securing their balance sheetsand called in loans. The loanable funds market and money market dried up.
Because the loanable funds market is global, the same problems quickly spread to other economies, and foreign exchange markets became highly volatile.
Hard-to-get loans, market volatility, and increased uncertainty transmitted the financial and monetary crisis to real expenditure decisions.

## The Policy Actions

Policy actions in response to the financial crisis dribbled out over a period of more than a year. But by November 2008, eight groups of policies designed to contain the crisis and minimize its impact on the real economy were in place. Figure 31.9 summarizes them, describes their effects on a bank's balance sheet, and identifies the problem that each action sought to address.
An open market operation is the classic policy (see Chapter 25, pp. 636-638) for providing liquidity and enabling the Fed to hit its interest rate target. With substantial interest rate cuts, open market operations were used on a massive scale to keep the banks well supplied with reserves. This action lowered bank holdings of securities and increased their reserves. By extending deposit insurance (see Chapter 25, p. 632), the FDIC gave depositors greater security and less incentive to withdraw their bank deposits. This action increased both deposits andreserves.

Three actions by the Fed provided additional liquidity in exchange for troubled assets. Term auction credit, primary dealer and broker credit, and the asset-backed commercial paper money market mutual fund liquidity facility enabled institutions toswap troubled assets for reserves or safer assets. All of these actions decreased bank holdings of securities and increasedreserves.
The Troubled Asset Relief Program (TARP) was an action by the U.S. Treasury, so technically it isn't a monetary policy action, but it had a direct impact on banks and other financial institutions. The programwas funded by $\$ 700$ billion of national debt.
The original intent (we'll call it TARP 1) was for the U.S. Treasury to buy troubled assets from banks and other holders and replace them with U.S. government securities. Implementing this program proved more difficult than initially anticipated and the benefits of the action came to be questioned.
So instead of buying troubled assets, the Treasury decided to buy equity stakes in troubled institutions (we'll call it TARP 2). This action directly increased the institutions' reserves and equity.
The final action was neither monetary policy nor fiscal policy but a change in accounting standards. It relaxed the requirement for institutions to value their assets at current market value-called "mark-to-market" -and permitted them, in rare conditions, to use a model to assess "fair market value."
Taken as a whole, a huge amount of relief was thrown at the financial crisis but the economy continued to perform poorly through 2009 and 2010.

## Persistently Slow Recovery

Despite extraordinary monetary (and fiscal) stimulus, at the end of 2010, the U.S. economy remained stuck with slow real GDP growth and an unemployment rate close to 10 percent. Why?
No one knows for sure, but the Fed's critics say that the Fed itself contributed to the problem morethan to the solution. That problem is extreme uncertainty about the future that is keeping business investment low. Critics emphasize the need for greater clarity about monetary policy strategy. We'll conclude this review of monetary policy by looking at two suggested policy strategies.
FIGURE 31.9 Policy Actions in a Financial and Banking Crisis
Action
Open market operation
Extension of deposit insurance
Term auction credit
Primary dealer and other broker credit + Equity $=$ Reserves + Loans and securities Problem addressed
Asset-backed commercial paper money
market mutual fund liquidity facility
Troubled Asset Relief Program (TARP 1)
Troubled Asset Relief Progran (TARP 2)
Fair value accounting

## Policy Strategies and Clarity

Two alternative approaches to monetary policy have been suggested and one of them has been used in other countries. They are

- Inflation rate targeting
- Taylor rule

Inflation Rate Targeting A monetary policy strategy in which the central bank makes a public commitment to achieve an explicit inflation target and explain how its policy actions will achieve it is called inflation rate targeting. Australia, Canada, New Zealand, Sweden, the United Kingdom, and the European Union have been targeting inflation since the 1990s. Inflation targeting focuses the public debate on what monetary policy can achieve and the best contribution it can make to attaining full employment and sustained growth. The central fact is that monetary policy is about managing inflation expectations. An explicit inflation target that is taken seriously andtoward which policy actions are aimed and explained is a sensible way to manage those expectations.
It is when the going gets tough that inflation targeting has the greatest benefit. It is difficult to imagine a serious inflation-targeting central bank permit- ting inflation to take off in the way that it did during the 1970s. And it is difficult to imagine deflation and ongoing recession such as Japan has endured for the past 10 years if monetary policy is guided by an explicit inflation target.

Taylor Rule One way to pursue an inflation target is to set the policy interest rate (for the Fed, the federal funds rate) by using a rule or formula. The most famous and most studied interest rate rule is the Taylor rule described in Economics inAction.
Supporters of the Taylor rule argue that in computer simulations, the rule works well and limits fluctuations in inflation and output. By using such a rule, monetary policy contributes toward lessening uncertainty-the opposite of current monetary policy. In financial markets, labor markets, and markets for goods and services, people make long-term commitments. So markets work best when plans are based on correctly anticipated inflation. A wellunderstood monetary policy helps to create an environment in which inflation is easier to forecast and manage.
The debates on inflation targeting and the Taylor rule will continue!

## ECONOMICS IN ACTION

## The Taylor Rule

The Taylor rule is a formula for setting the federal funds rate. Calling the federal funds rate $F F R$, the inflation rate $I N F$, and the output gap GAP (all percentages), the Taylor rule formula is

$$
F F R=2+I N F+0.5(I N F-2)+0.5 G A P
$$

In words, the Taylor rule sets the federal funds rate at 2 percent a year plus the inflation rate plus one half of the deviation of inflation from 2 percent a year, plus one half of the output gap.

Stanford University economist John B. Taylor, who devised this rule, says inflation and real GDP would fluctuate much less if the FOMC were to use it-the Taylor rule beats the FOMC's historical performance.

The Taylor rule implies that the Fed caused the boom and bust of the past decade. The federal funds rate was 1.5 percentage points (on average) too low from 2001 through 2005, which fuelled the boom; and the rate was 0.5 percentage points (on average) too high in 2006 and 2007, which triggered the bust.

In the conditions of 2009, the Taylor rule delivered a negative interest rate, a situation that wouldn't have arisen if the Taylor rule had been followed.

## ECONOMIC ANALYSIS

- The recovery from the 2008-2009 recession has been slow, and a high unemployment rate and a recessionary gap have persisted.
- In September 2014, the Fed was sufficiently concerned about the slow pace of the recovery to maintain its "forward guidance" on interest rates by continuing its pledge to keep a very low interest rate for a "considerable time."
- Everyone agreed that in 2014 unemployment remained a problem but inflation was well inside the comfort zone.
- But not everyone agreed with the Fed's assessment of the situation and some wanted a more flexible approach and a willingness to raise interest rates earlier.
- The Fed's commitment to ongoing stimulus from a near zero interest rate was based on a view that the output gap remained large.
- But there is great uncertainty about the size of the output gap.
- Figure 1 shows two views of the gap. CBO is the official view of the Congressional Budget Office and most likely the view of the FOMC majority. $S F$ is an estimate by economists at the San Francisco Fed and most likely the view of the two FOMC dissenters.
- While there is disagreement about when to start raising interest rates, there is no disagreement that they are going up in 2015, 2016, and 2017.
- Figure 2 shows the forecasts of the 8 FOMC members and 9 other Fed board members and regional Fed presidents. (The data behind Fig. 2 are published in the FOMC minutes for its July 2014 meeting.)
- Whether the Fed will get the economy back to full employment without triggering a new outburst of inflation depends on the true size of the output gap and the speed with which a future rise in interest rates keeps aggregate demand at a non-inflationary level.
- Figure 3 illustrates the Fed's challenge. In mid-2014, real GDP was $\$ 16$ trillion and the price level was 108 at the intersection of $A D_{0}$ and $S A S$.
- The CBO says potential GDP was $\$ 16.7$ trillion, with a recessionary gap of $\$ 0.7$ trillion on $L A S_{C B O}$. The San Francisco Fed says potential GDP was $\$ 16.1$ trillion, with a recessionary gap of $\$ 0.1$ trillion on $L A S_{S F}$.
- If continued low interest rates into 2015 increases aggregate demand to $A D_{1}$, real GDP will increase to $\$ 16.7$ trillion and the price level will rise to 110 (a low inflation rate).
- What happens next depends on who is correct about potential GDP. If the CBO is correct, full employment is restored and inflation remains subdued.


Figure 1 Two Views of the Output Gap


Figure 2 Seventeen Views on the Interest Rate


Figure 3 Two Possible Outcomes

- If the San Francisco Fed economists" view is correct, an inflationary gap will open and a demand-pull inflation will begin.


## ENDS


[^0]:    *In the formula, the Greek letter delta ( $\Delta$ ) stands for "change in" and $\% \Delta$ stands for "percentage change in."

