b. What are the firm's profit-maximizing output and price? What is its profit?
c. What would the equilibrium price and quantity be in a competitive industry?
d. What would the social gain be if this monopolist were forced to produce and price at the competitive equilibrium? Who would gain and lose as a result?

6. A firm has two factories, for which costs are given by:

$$\text{Factory \#1: } C_1(Q_1) = 10Q_1^2$$

$$\text{Factory \#2: } C_2(Q_2) = 20Q_2^2$$

The firm faces the following demand curve:

$$P = 700 - 5Q$$

where Q is total output—i.e., $Q = Q_1 + Q_2$.
a. On a diagram, draw the marginal cost curves for the two factories, the average and marginal revenue curves, and the total marginal cost curve (i.e., the marginal cost of producing $Q = Q_1 + Q_2$). Indicate the profit-maximizing output for each factory, total output, and price.
b. Calculate the values of Q_1, Q_2, Q, and P that maximize profit.
c. Suppose labor costs increase in Factory 1 but not in Factory 2. How should the firm adjust (i.e., raise, lower, or leave unchanged) the following: Output in Factory 1? Output in Factory 2? Total output? Price?

7. A drug company has a monopoly on a new patented medicine. The product can be made in either of two plants. The costs of production for the two plants are $MC_1 = 20 + 2Q_1$ and $MC_2 = 10 + 5Q_2$. The firm's estimate of demand for the product is $P = 20 - 3(Q_1 + Q_2)$. How much should the firm plan to produce in each plant? At what price should it plan to sell the product?

8. One of the more important antitrust cases of this century involved the Aluminum Company of America (Alcoa) in 1945. At that time, Alcoa controlled about 90 percent of primary aluminum production in the United States, and the company had been accused of monopolizing the aluminum market. In its defense, Alcoa argued that although it indeed controlled a large fraction of the primary market, secondary aluminum (i.e., aluminum produced from the recycling of scrap) accounted for roughly 30 percent of the total supply of aluminum, and many competitive firms were engaged in recycling. Therefore, Alcoa argued, it did not have much monopoly power.
a. Provide a clear argument *in favor* of Alcoa's position.
b. Provide a clear argument *against* Alcoa's position.
c. The 1945 decision by Judge Learned Hand has been called "one of the most celebrated judicial opinions of our time." Do you know what Judge Hand's ruling was?

9. A monopolist faces the demand curve $P = 11 - Q$, where P is measured in dollars per unit and Q in thousands of units. The monopolist has a constant average cost of \$6 per unit.
a. Draw the average and marginal revenue curves and the average and marginal cost curves. What are the monopolist's profit-maximizing price and quantity? What is the resulting profit? Calculate the firm's degree of monopoly power using the Lerner index.
b. A government regulatory agency sets a price ceiling of \$7 per unit. What quantity will be produced, and what will the firm's profit be? What happens to the degree of monopoly power?
c. What price ceiling yields the largest level of output? What is that level of output? What is the firm's degree of monopoly power at this price?

10. Michelle's Monopoly Mutant Turtles (MMMT) has the exclusive right to sell Mutant Turtle t-shirts in the United States. The demand for these t-shirts is $Q = 10{,}000/P^2$. The firm's short-run cost is $SRTC = 2000 + 5Q$, and its long-run cost is $LRTC = 6Q$.
a. What price should MMMT charge to maximize profit in the short run? What quantity does it sell, and how much profit does it make? Would it be better off shutting down in the short run?
b. What price should MMMT charge in the long run? What quantity does it sell and how much profit does it make? Would it be better off shutting down in the long run?
c. Can we expect MMMT to have lower marginal cost in the short run than in the long run? Explain why.

*11. You produce widgets for sale in a perfectly competitive market at a market price of \$10 per widget. Your widgets are manufactured in two plants, one in Massachusetts and the other in Connecticut. Because of labor problems in Connecticut, you are forced to raise wages there, so that marginal costs in that plant increase. In response to this, should you shift production and produce more in your Massachusetts plant?

12. The employment of teaching assistants (TAs) by major universities can be characterized as a monopsony. Suppose the demand for TAs is $W = 30{,}000 - 125n$, where W is the wage (as an annual salary) and n is the number of TAs hired. The supply of TAs is given by $W = 1000 + 75n$.
a. If the university takes advantage of its monopsonist position, how many TAs will it hire? What wage will it pay?
b. If, instead, the university faced an infinite supply of TAs at the annual wage level of \$10,000, how many TAs would it hire?

*13. Dayna's Doorstops, Inc. (DD), is a monopolist in the doorstop industry. Its cost is $C = 100 - 5Q + Q^2$, and demand is $P = 55 - 2Q$.

a. What price should DD set to maximize profit? What output does the firm produce? How much profit and consumer surplus does DD generate?
b. What would output be if DD acted like a perfect competitor and set $MC = P$? What profit and consumer surplus would then be generated?
c. What is the deadweight loss from monopoly power in part (a)?
d. Suppose the government, concerned about the high price of doorstops, sets a maximum price at \$27. How does this affect price, quantity, consumer surplus, and DD's profit? What is the resulting deadweight loss?
e. Now suppose the government sets the maximum price at \$23. How does this decision affect price, quantity, consumer surplus, DD's profit, and deadweight loss?
f. Finally, consider a maximum price of \$12. What will this do to quantity, consumer surplus, profit, and deadweight loss?

*14. There are 10 households in Lake Wobegon, Minnesota, each with a demand for electricity of $Q = 50 - P$. Lake Wobegon Electric's (LWE) cost of producing electricity is $TC = 500 + Q$.
a. If the regulators of LWE want to make sure that there is no deadweight loss in this market, what price will they force LWE to charge? What will output be in that case? Calculate consumer surplus and LWE's profit with that price.
b. If regulators want to ensure that LWE doesn't lose money, what is the lowest price they can impose? Calculate output, consumer surplus, and profit. Is there any deadweight loss?
c. Kristina knows that deadweight loss is something that this small town can do without. She suggests that each household be required to pay a fixed amount just to receive any electricity at all, and then a per-unit charge for electricity. Then LWE can break even while charging the price calculated in part (a). What fixed amount would each household have to pay for Kristina's plan to work? Why can you be sure that no household will choose instead to refuse the payment and go without electricity?

15. A monopolist faces the following demand curve:

$$Q = 144/P^2$$

where Q is the quantity demanded and P is price. Its *average variable* cost is

$$AVC = Q^{1/2}$$

and its *fixed cost* is 5.
a. What are its profit-maximizing price and quantity? What is the resulting profit?
b. Suppose the government regulates the price to be no greater than \$4 per unit. How much will the monopolist produce? What will its profit be?
c. Suppose the government wants to set a ceiling price that induces the monopolist to produce the largest possible output. What price will accomplish this goal?

CHAPTER 11

Pricing with Market Power

Chapter Outline

List of Examples

As we explained in Chapter 10, market power is quite common. Many industries have only a few producers, so that each producer has some monopoly power. And many firms, as buyers of raw materials, labor, or specialized capital goods, have some monopsony power in the markets for these factor inputs. The problem faced by the managers of these firms is how to use their market power most effectively. They must decide how to set prices, choose quantities of factor inputs, and determine output in both the short and long run to maximize the firm's profit.

Managers of firms with market power have a harder job than those who manage perfectly competitive firms. A firm that is perfectly competitive in output markets has no influence over market price. As a result, its managers need worry only about the cost side of the firm's operations, choosing output so that price is equal to marginal cost. But the managers of a firm with monopoly power must also worry about the characteristics of demand. Even if they set a single price for the firm's output, they must obtain at least a rough estimate of the elasticity of demand to determine what that price (and corresponding output level) should be. Furthermore, firms can often do much better by using a more complicated pricing strategy—for example, charging different prices to different customers. To design such pricing strategies, managers need ingenuity and even more information about demand.

This chapter explains how firms with market power set prices. We begin with the basic objective of every pricing strategy: capturing consumer surplus and converting it into additional profit for the firm. Then we discuss how this goal can be achieved using *price discrimination*. Here different prices are charged to different customers, sometimes for the same product and sometimes for small variations in the product. Because price discrimination is widely practiced in one form or another, it is important to understand how it works.

Next, we discuss the *two-part tariff*. Here customers must pay in advance for the right to purchase units of the good at a later time (and at additional cost). The classic example of this is an amusement park, where customers pay a fee to enter and then additional fees for each ride they go on. Although amusement parks may seem like a rather specialized market, there are many other examples of two-part tariffs: the price of a

Gillette razor, which gives the owner the opportunity to purchase Gillette razor blades; the price of a Polaroid camera, which gives the owner the opportunity to purchase Polaroid film; or the monthly subscription cost of a mobile telephone, which gives users the opportunity to make phone calls from their automobiles, paying by the message unit as they do so.

We will also discuss *bundling*, a pricing strategy that simply involves tying products together and selling them as a package. For example: a personal computer that comes bundled with several software packages; a one-week vacation in Hawaii in which the airfare, rental car, and hotel are bundled and sold at a single package price; or a luxury car, in which the air conditioning, power windows, and stereo are "standard" features.

Finally, we will examine the use of *advertising* by firms with market power. As we will see, deciding how much money to spend on advertising requires information about demand and is closely related to the firm's pricing decision. We will derive a simple rule of thumb for determining the profit-maximizing advertising-to-sales ratio.

11.1 Capturing Consumer Surplus

All the pricing strategies that we will examine have one thing in common: they are means of capturing consumer surplus and transferring it to the producer. You can see this more clearly in Figure 11.1. Suppose the firm sold all its output

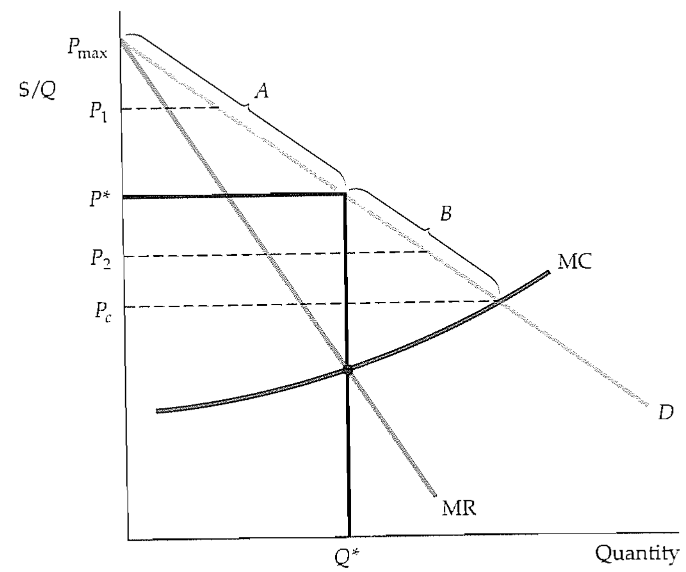

FIGURE 11.1 Capturing Consumer Surplus

If a firm can charge only one price for all its customers, that price will be P^* and the quantity produced will be Q^*. Ideally, the firm would like to charge a higher price to consumers willing to pay more than P^*, thereby capturing some of the consumer surplus under region *A* of the demand curve. The firm would also like to sell to consumers willing to pay prices lower than P^*, but only if doing so does not entail lowering the price to other consumers. In that way, the firm could also capture some of the surplus under region *B* of the demand curve.

at a single price. To maximize profit, it would pick a price P^* and corresponding output Q^* at the intersection of its marginal cost and marginal revenue curves. Although the firm would then be profitable, its managers might still wonder if they could make it even more profitable.

They know that some customers (in region *A* of the demand curve) would pay more than P^*. But raising price would mean losing some customers, selling less, and earning smaller profits. Similarly, other potential customers are not buying the firm's product because they will not pay a price as high as P^*. Many of them, however, would pay prices higher than the firm's marginal cost. (These customers are in region *B* of the demand curve.) By lowering its price, the firm could sell to some of these customers. Unfortunately, it would then earn less revenue from its existing customers, and again profits would shrink.

How can the firm capture the consumer surplus (or at least part of it) from its customers in region *A*, and perhaps also sell profitably to some of its potential customers in region *B*? Charging a single price clearly will not do the trick. However, the firm might charge different prices to different customers, according to where the customers are along the demand curve. For example, some customers in the upper end of region *A* would be charged the higher price P_1, some in region *B* would be charged the lower price P_2, and some in between would be charged P^*. This is the basis of **price discrimination:** charging different prices to different customers. The problem, of course, is to identify the different customers, and to get them to pay different prices. We will see how this can be done in the next section.

price discrimination Practice of charging different prices to different consumers for similar goods.

The other pricing techniques that we will discuss in this chapter—two-part tariffs and bundling—also expand the range of a firm's market to include more customers and to capture more consumer surplus. In each case, we will examine both the amount by which the firm's profit can be increased and the effect on consumer welfare. (As we will see, when there is a high degree of monopoly power, these pricing techniques can sometimes make both consumers and the producer better off.) We turn first to price discrimination.

11.2 Price Discrimination

Price discrimination can take three broad forms, which we call first-, second-, and third-degree price discrimination. We will examine them in turn.

First-Degree Price Discrimination

Ideally, a firm would like to charge a different price to each of its customers. If it could, it would charge each customer the maximum price that the customer is willing to pay for each unit bought. We call this maximum price the customer's **reservation price.** The practice of charging each customer his or her reservation price is called perfect **first-degree price discrimination.**[1] Let's see how it affects the firm's profit.

reservation price Maximum price that a customer is willing to pay for a good.

first-degree price discrimination Practice of charging each customer her reservation price.

First, we need to know the profit that the firm earns when it charges only the single price P^* in Figure 11.2. To find out, we can add the profit on each incremental unit produced and sold, up to the total quantity Q^*. This incremental profit is the marginal revenue less the marginal cost for each unit. In Figure 11.2,

[1] We are assuming that each customer buys one unit of the good. If a customer bought more than one unit, the firm would have to charge different prices for each of the units.

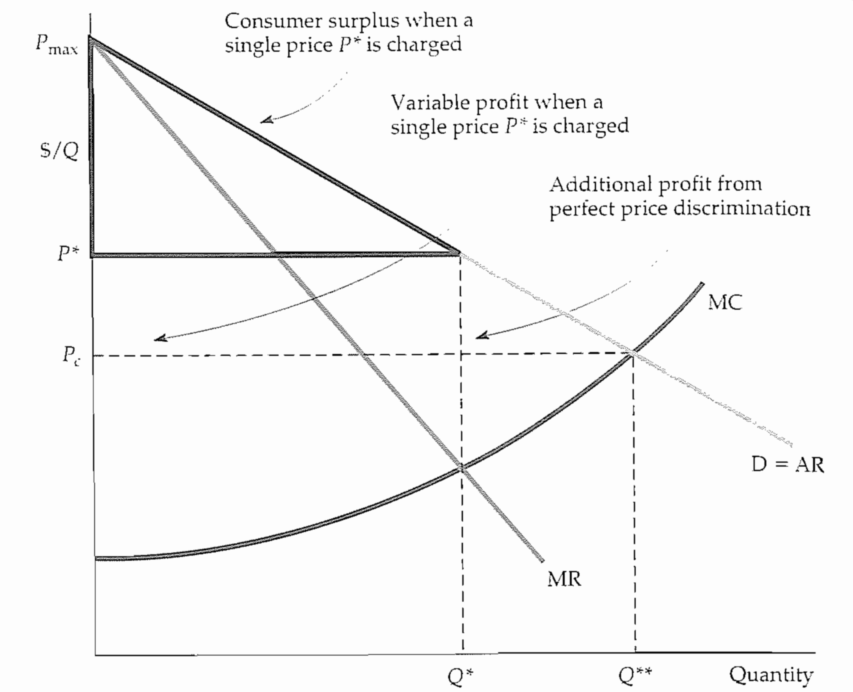

FIGURE 11.2 Additional Profit from Perfect First-Degree Price Discrimination

Because the firm charges each consumer her reservation price, it is profitable to expand output to Q^{**}. When only a single price, P^* is charged, the firm's variable profit is the area between the marginal revenue and marginal cost curves. With perfect price discrimination, this profit expands to the area between the demand curve and the marginal cost curve.

In §8.1, we explain that a firm's profit-maximizing output is at the point at which marginal revenue is equal to marginal cost.

variable profit Sum of profits on each incremental unit produced by a firm; i.e., profit ignoring fixed costs.

this marginal revenue is highest and marginal cost lowest for the first unit. For each additional unit, marginal revenue falls and marginal cost rises. Thus the firm produces the total output Q^*, at which point marginal revenue and marginal cost are equal.

If we add up the profits on each incremental unit produced, we obtain the firm's **variable profit:** the firm's profit, ignoring its fixed costs. In Figure 11.2, variable profit is given by the yellow-shaded area between the marginal revenue and marginal cost curves.[2] Consumer surplus, which is the area between the average revenue curve and the price P^* that customers pay, is outlined as a black triangle.

Now, what happens if the firm can perfectly price discriminate? Because each consumer is charged exactly what he or she is willing to pay, the marginal revenue curve is no longer relevant to the firm's output decision. Instead, the incremental revenue earned from each additional unit sold is simply the price paid for that unit; it is therefore given by the demand curve.

[2] Recall from Chapter 10 that because total profit π is the difference between total revenue R and total cost C, incremental profit is just $\Delta\pi = \Delta R - \Delta C = MR - MC$. Variable profit is found by summing all the $\Delta\pi$s, and thus it is the area between the MR and MC curves. This ignores fixed costs, which are independent of the firm's output and pricing decisions. Hence, total profit equals variable profit minus fixed cost.

Since price discrimination does not affect the firm's cost structure, the cost of each additional unit is again given by the firm's marginal cost curve. Therefore, *the additional profit from producing and selling an incremental unit is now the difference between demand and marginal cost*. As long as demand exceeds marginal cost, the firm can increase its profit by expanding production. It will do so until it produces a total output Q^{**}. At Q^{**}, demand is equal to marginal cost, and producing any more reduces profit.

Variable profit is now given by the area between the demand and marginal cost curves.[3] Observe from Figure 11.2 how the firm's profit has increased. (The additional profit resulting from price discrimination is shown by the purple shaded area.) Note also that because every customer is being charged the maximum amount that he or she is willing to pay, all consumer surplus has been captured by the firm.

In practice, perfect first-degree price discrimination is almost never possible. First, it is usually impractical to charge each and every customer a different price (unless there are only a few customers). Second, a firm usually does not know the reservation price of each customer. Even if it could ask how much each customer would be willing to pay, it probably would not receive honest answers. After all, it is in the customers' interest to claim that they would pay very little.

Sometimes, however, firms can discriminate imperfectly by charging a few different prices based on estimates of customers' reservation prices. This practice is often used by professionals, such as doctors, lawyers, accountants, or architects, who know their clients reasonably well. In such cases, the client's willingness to pay can be assessed and fees set accordingly. For example, a doctor may offer a reduced fee to a low-income patient whose willingness to pay or insurance coverage is low, but charge higher fees to upper-income or better-insured patients. And an accountant, having just completed a client's tax returns, is in an excellent position to estimate how much the client is willing to pay for the service.

Another example is a car salesperson, who typically works with a 15-percent profit margin. The salesperson can give part of this margin away to the customer by making a "deal," or can insist that the customer pay the full sticker price. A good salesperson knows how to size up customers and determine whether they will look elsewhere for a car if they don't receive a sizable discount (from the salesperson's point of view, a small profit is better than no sale and no profit), but the customer in a hurry is offered little or no discount. In other words, a successful car salesperson knows how to price discriminate!

Still another example is college and university tuition. Colleges don't charge different tuition rates to different students in the same degree programs. Instead, they offer financial aid, in the form of a scholarship or subsidized loan, which reduces the *net* tuition that the student must pay. By requiring those who seek aid to disclose information about family income and wealth, colleges can link the amount of aid to ability (and hence willingness) to pay. Thus students who are financially well off pay more for their education, while students who are less well off pay less.

[3] Incremental profit is again $\Delta\pi = \Delta R - \Delta C$, but ΔR is given by the price to each customer (i.e., the average revenue curve), so $\Delta\pi = AR - MC$. Variable profit is the sum of these $\Delta\pi$s and is given by the area between the AR and MC curves.

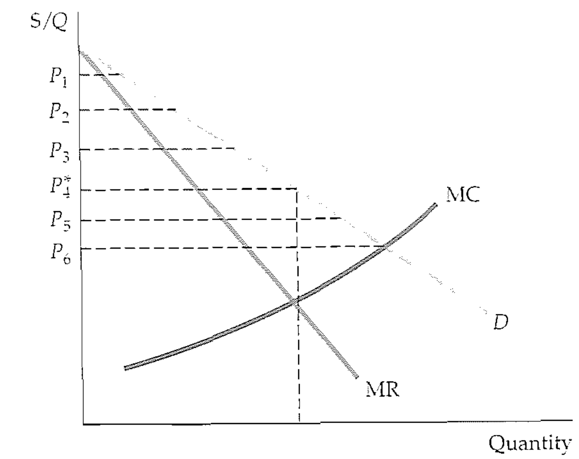

FIGURE 11.3 First-Degree Price Discrimination in Practice

Firms usually don't know the reservation price of every consumer, but sometimes reservation prices can be roughly identified. Here, six different prices are charged. The firm earns higher profits, but some consumers may also benefit. With a single price P_4^*, there are fewer consumers. The consumers who now pay P_5 or P_6 enjoy a surplus.

Figure 11.3 illustrates this kind of imperfect first-degree price discrimination. Here, if only a single price were charged, it would be P_4^*. Instead, six different prices are charged, the lowest of which, P_6, is at about the point where marginal cost intersects the demand curve. Note that those customers who would not have been willing to pay a price of P_4^* or greater are actually better off in this situation—they are now in the market and may be enjoying at least some consumer surplus. In fact, if price discrimination brings enough new customers into the market, consumer welfare can increase to the point that both the producer and consumers are better off.

Second-Degree Price Discrimination

In some markets, as each consumer purchases many units of a good over any given period, his or her demand declines with the number of units purchased. Examples include water, heating fuel, and electricity. Consumers may each purchase a few hundred kilowatt-hours of electricity a month, but their willingness to pay declines with increasing consumption. The first 100 kilowatt-hours may be worth a lot to the consumer—operating a refrigerator and providing for minimal lighting. Conservation becomes easier with the additional units and may be worthwhile if the price is high. In this situation, a firm can discriminate according to the quantity consumed. This is called **second-degree price discrimination**, and it works by charging different prices for different quantities of the same good or service.

second-degree price discrimination Practice of charging different prices per unit for different quantities of the same good or service.

Quantity discounts are an example of second-degree price discrimination. A single roll of Kodak film might be priced at \$5, while a box containing four rolls of the same film might be priced at \$14, making the average price per roll \$3.50. Similarly, the price per ounce for breakfast cereal is likely to be smaller for the 24-ounce box than for the 16-ounce box.

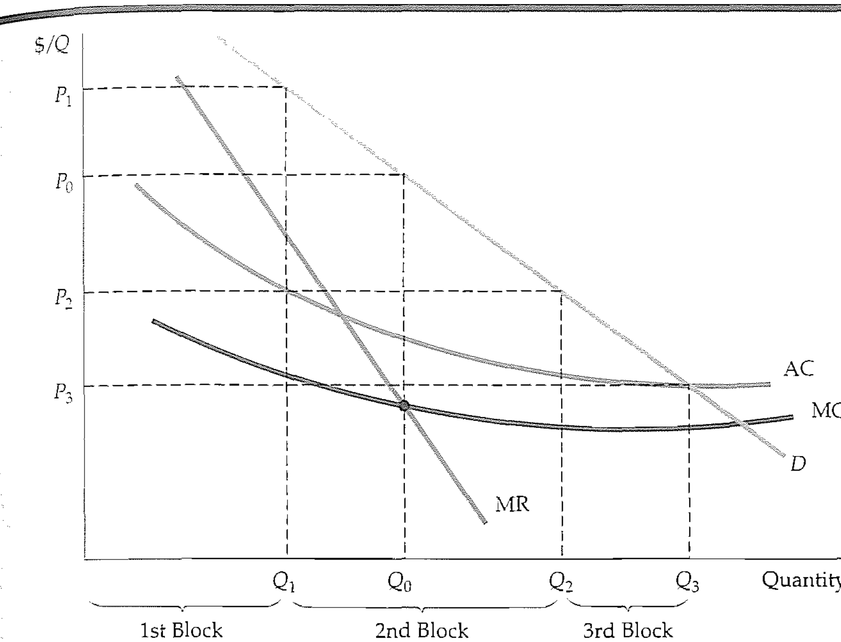

FIGURE 11.4 Second-Degree Price Discrimination

Different prices are charged for different quantities, or "blocks," of the same good. Here, there are three blocks, with corresponding prices P_1, P_2, and P_3. There are also economies of scale, and average and marginal costs are declining. Second-degree price discrimination can then make consumers better off by expanding output and lowering cost.

Another example of second-degree price discrimination is *block pricing* by electric power companies, natural gas utilities, and municipal water companies. With **block pricing**, the consumer is charged different prices for different quantities or "blocks" of a good. If scale economies cause average and marginal costs to decline, the government agency that controls rates may encourage block pricing. Because it leads to expanded output and greater scale economies, this policy can increase consumer welfare while allowing for greater profit to the company: While prices are reduced overall, the savings from the lower unit cost still permits the company to increase its profit.

block pricing Practice of charging different prices for different quantities or "blocks" of a good.

Figure 11.4 illustrates second-degree price discrimination for a firm with declining average and marginal costs. If a single price were charged, it would be P_0, and the quantity produced would be Q_0. Instead, three different prices are charged, based on the quantities purchased. The first block of sales is priced at P_1, the second at P_2, and the third at P_3.

Third-Degree Price Discrimination

A well-known liquor company has what seems to be a strange pricing practice. The company produces a vodka that it advertises as one of the smoothest and best-tasting available. This vodka is called "Three Star Golden Crown," and it is sold for about \$16 a bottle.[4] However, the company also takes some of this same

[4] We have changed the names to protect the innocent.

vodka and bottles it under the name "Old Sloshbucket," which is sold for about $8 a bottle. Why does it do this? Has the president of the company been spending too much time near the vats?

third-degree price discrimination Practice of dividing consumers into two or more groups with separate demand curves and charging different prices to each group.

Perhaps, but this liquor company is also practicing **third-degree price discrimination**, and it does it because the practice is profitable. This form of price discrimination divides consumers into two or more groups with separate demand curves for each group. It is the most prevalent form of price discrimination, and examples abound: regular versus "special" airline fares; premium versus nonpremium brands of liquor, canned food or frozen vegetables; discounts to students and senior citizens; and so on.

Creating Consumer Groups In each case, some characteristic is used to divide consumers into distinct groups. For many goods, for example, students and senior citizens are usually willing to pay less on average than the rest of the population (because their incomes are lower), and identity can be readily established (via a college ID or driver's license). Likewise, to separate vacationers from business travelers (whose companies are usually willing to pay higher fares), airlines can put restrictions on special low-fare tickets, such as requiring advance purchase or a Saturday night stay. With the liquor company, or the premium versus nonpremium (e.g., supermarket label) brand of food, the label itself divides consumers; many consumers are willing to pay more for a name brand even though the nonpremium brand is identical or nearly identical (and is in fact sometimes manufactured by the same company that produced the premium brand).

If third-degree price discrimination is feasible, how should the firm decide what price to charge each group of consumers? Let's think about this in two steps.

1. We know that however much is produced, total output should be divided between the groups of customers so that marginal revenues for each group are equal. Otherwise, the firm would not be maximizing profit. For example, if there are two groups of customers and the marginal revenue for the first group, MR_1, exceeds the marginal revenue for the second group, MR_2, the firm could clearly do better by shifting output from the second group to the first. It would do this by lowering the price to the first group and raising the price to the second group. Thus whatever the two prices, they must be such that the marginal revenues for the different groups are equal.
2. We know that *total* output must be such that the marginal revenue for each group of consumers is equal to the marginal cost of production. Again, if this were not the case, the firm could increase its profit by raising or lowering total output (and lowering or raising its prices to both groups). For example, suppose that marginal revenues were the same for each group of consumers but that marginal revenue exceeded the marginal cost of production. The firm could then make a greater profit by increasing its total output. It would lower its prices to both groups of consumers, so that marginal revenues for each group fell (but were still equal to each other) and approached marginal cost (which would increase as total output increased).

Let's look at this problem algebraically. Let P_1 be the price charged to the first group of consumers, P_2 the price charged to the second group, and $C(Q_T)$ the total cost of producing output $Q_T = Q_1 + Q_2$. In this case, total profit is given by

$$\pi = P_1Q_1 + P_2Q_2 - C(Q_T)$$

The firm should increase its sales to each group of consumers, Q_1 and Q_2, until the incremental profit from the last unit sold is zero. First, we set incremental profit for sales to the first group of consumers equal to zero:

$$\frac{\Delta\pi}{\Delta Q_1} = \frac{\Delta(P_1Q_1)}{\Delta Q_1} - \frac{\Delta C}{\Delta Q_1} = 0$$

Here, $\Delta(P_1Q_1)/\Delta Q_1$ is the incremental revenue from an extra unit of sales to the first group of consumers (i.e., MR_1). The next term, $\Delta C/\Delta Q_1$, is the incremental cost of producing this extra unit—i.e., marginal cost, MC. We thus have

$$MR_1 = MC$$

Similarly, for the second group of consumers, we must have

$$MR_2 = MC$$

Putting these relations together, we see that prices and output must be set so that

$$MR_1 = MR_2 = MC \qquad \textbf{(11.1)}$$

Again, marginal revenue must be equal across groups of consumers and must equal marginal cost.

Determining Relative Prices Managers may find it easier to think in terms of the relative prices that should be charged to each group of consumers and to relate these prices to the elasticities of demand. Recall from Section 10.1 that we can write marginal revenue in terms of the elasticity of demand:

In our discussion of a rule of thumb for pricing in §10.1, we explained that a profit-maximizing firm chooses an output at which its marginal revenue is equal to the price of the product plus the ratio of the price to the price elasticity of demand.

$$MR = P(1 + 1/E_d)$$

Thus $MR_1 = P_1(1 + 1/E_1)$ and $MR_2 = P_2(1 + 1/E_2)$, where E_1 and E_2 are the elasticities of demand for the firm's sales in the first and second markets, respectively. Now equating MR_1 and MR_2 as in equation (11.1) gives the following relationship that must hold for the prices:

$$\frac{P_1}{P_2} = \frac{(1 + 1/E_2)}{(1 + 1/E_1)} \qquad \textbf{(11.2)}$$

As you would expect, the higher price will be charged to consumers with the lower demand elasticity. For example, if the elasticity of demand for consumers in group 1 is -2 and the elasticity for consumers in group 2 is -4, we will have $P_1/P_2 = (1 - 1/4)/(1 - 1/2) = (3/4)/(1/2) = 1.5$. In other words, the price charged to the first group of consumers should be 1.5 times as high as the price charged to the second group.

Figure 11.5 illustrates third-degree price discrimination. Note that the demand curve D_1 for the first group of consumers is less elastic than the curve for the second group; the price charged to the first group is likewise higher. The total quantity produced, $Q_T = Q_1 + Q_2$, is found by summing the marginal revenue curves MR_1 and MR_2 horizontally, which yields the dashed curve MR_T, and finding its intersection with the marginal cost curve. Because MC must equal

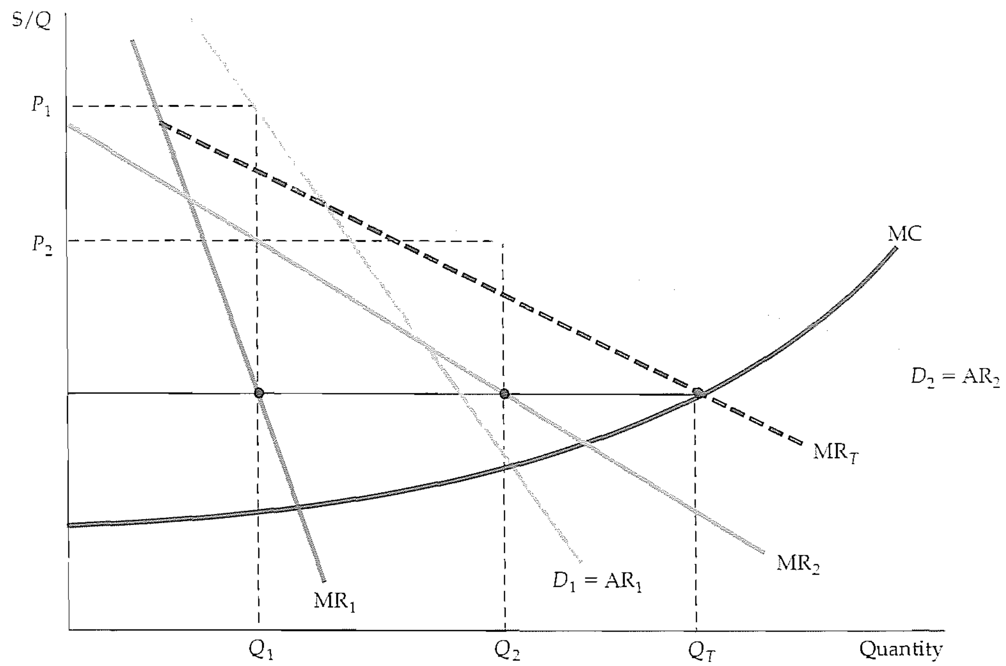

FIGURE 11.5 Third-Degree Price Discrimination

Consumers are divided into two groups, with separate demand curves for each group. The optimal prices and quantities are such that the marginal revenue from each group is the same and equal to marginal cost. Here group 1, with demand curve D_1, is charged P_1, and group 2, with the more elastic demand curve D_2, is charged the lower price P_2. Marginal cost depends on the total quantity produced Q_T. Note that Q_1 and Q_2 are chosen so that $MR_1 = MR_2 = MC$.

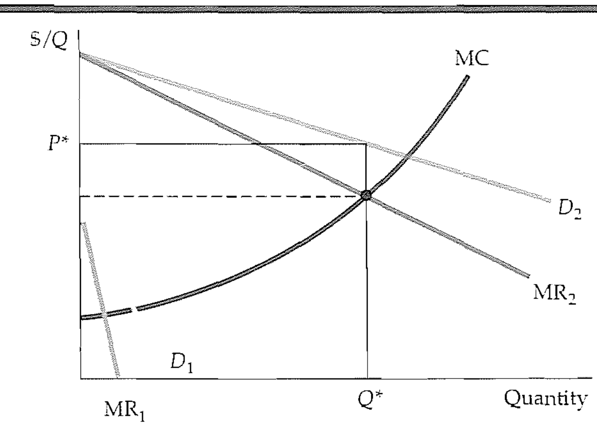

FIGURE 11.6 No Sales to Smaller Market

Even if third-degree price discrimination is feasible, it does not always pay to sell to both groups of consumers if marginal cost is rising. Here, the first group of consumers, with demand D_1, are not willing to pay much for the product. It is unprofitable to sell to them because the price would have to be too low to compensate for the resulting increase in marginal cost.

MR_1 and MR_2, we can draw a horizontal line leftward from this intersection to find the quantities Q_1 and Q_2.

It may not always be worthwhile for the firm to try to sell to more than one group of consumers. In particular, if demand is small for the second group and marginal cost is rising steeply, the increased cost of producing and selling to this group may outweigh the increase in revenue. In Figure 11.6, therefore, the firm is better off charging a single price P^* and selling only to the larger group of consumers: The additional cost of serving the smaller market would outweigh the additional revenue that might come from the smaller market.

EXAMPLE 11.1 The Economics of Coupons and Rebates

The producers of processed foods and related consumer goods often issue coupons that let customers buy products at discounts. These coupons are usually distributed as part of an advertisement for the product. They may appear in newspapers or magazines or in promotional mailings. For example, a coupon for a particular breakfast cereal might be worth 25 cents toward the purchase of a box of the cereal. Why do firms issue these coupons? Why not just lower the price of the product and thereby save the costs of printing and collecting the coupons?

Coupons provide a means of price discrimination. Studies show that only about 20 to 30 percent of all consumers regularly bother to clip, save, and use coupons. These consumers tend to be more sensitive to price than those who ignore coupons. They generally have more price-elastic demands and lower reservation prices. By issuing coupons, therefore, a cereal company can separate its customers into two groups and, in effect, charge the more price-sensitive customers a lower price than the other customers.

Rebate programs work the same way. For example, Kodak ran a program in which a consumer could mail in a form together with the proof of purchase of three rolls of film and receive a rebate of $1.50. Why not just lower the price of film by 50 cents a roll? Because only those consumers with relatively price-sensitive demands bother to send in the materials and request rebates. Again, the program is a means of price discrimination.

Can consumers really be divided into distinct groups in this way? Table 11.1 shows the results of a statistical study in which, for a variety of products, price elasticities of demand were estimated for users and nonusers of coupons.[5] This study confirms that users of coupons tend to have more price-sensitive demands. It also shows the extent to which the elasticities differ for the two groups of consumers and how the difference varies from one product to another.

[5] The study is by Chakravarthi Narasimhan, "A Price Discrimination Theory of Coupons," *Marketing Science* (Spring 1984). A recent study of coupons for breakfast cereals finds that contrary to the predictions of the price-discrimination model, shelf prices for cereals tend to be lower during periods when coupons are more widely available. This might occur because couponing spurs more price competition among cereal manufacturers. See Aviv Nevo and Catherine Wolfram, "Prices and Coupons for Breakfast Cereals," National Bureau of Economic Research Working Paper No. 6932, February 1999.

TABLE 11.1 Price Elasticities of Demand for Users versus Nonusers of Coupons

	PRICE ELASTICITY	
Product	Nonusers	Users
Toilet tissue	−0.60	−0.66
Stuffing/dressing	−0.71	−0.96
Shampoo	−0.84	−1.04
Cooking/salad oil	−1.22	−1.32
Dry mix dinners	−0.88	−1.09
Cake mix	−0.21	−0.43
Cat food	−0.49	−1.13
Frozen entrees	−0.60	−0.95
Gelatin	−0.97	−1.25
Spaghetti sauce	−1.65	−1.81
Creme rinse/conditioner	−0.82	−1.12
Soups	−1.05	−1.22
Hot dogs	−0.59	−0.77

By themselves, these elasticity estimates do not tell a firm what price to set and how large a discount to offer because they pertain to *market demand*, not to the demand for the firm's particular brand. For example, Table 11.1 indicates that the elasticity of demand for cake mix is −0.21 for nonusers of coupons and −0.43 for users. But the elasticity of demand for any of the eight or ten major brands of cake mix on the market will be far larger than either of these numbers—about eight or ten times as large, as a rule of thumb.[6] So for any one brand of cake mix, say, Pillsbury, the elasticity of demand for users of coupons might be about −4, versus about −2 for nonusers. From equation (11.2), therefore, we can determine that the price to nonusers of coupons should be about 1.5 times the price to users. In other words, if a box of cake mix sells for $1.50, the company should offer coupons that give a 50-cent discount.

EXAMPLE 11.2 Airline Fares

Travelers are often amazed at the variety of fares available for a round-trip flight from New York to Los Angeles. For example, the first-class fare was recently above $3000; the regular (unrestricted) economy fare was about $1800; and special discount fares (often requiring the purchase of a ticket two weeks in advance and/or a Saturday night stayover) could be bought for as little as $400. Although first-class service is not the same as economy service with a minimum stay requirement, the difference would not seem to warrant a price that is four times as high. Why do airlines set such fares?

[6] This rule of thumb follows if interfirm competition can be described by the Cournot model, which we discuss in Chapter 12.

TABLE 11.2 Elasticities of Demand for Air Travel

	FARE CATEGORY		
Elasticity	First-Class	Unrestricted Coach	Discount
Price	−0.3	−0.4	−0.9
Income	1.2	1.2	1.8

These fares provide a profitable form of price discrimination. The gains from discriminating are large because different types of customers, with very different elasticities of demand, purchase these different types of tickets. Table 11.2 shows price (and income) elasticities of demand for three categories of service within the United States: first-class, unrestricted coach, and discount tickets. (A discounted ticket often has restrictions and may be partly nonrefundable.)

Note that the demand for discounted fares is about two or three times as price elastic as first-class or unrestricted coach service. Why the difference? While discounted tickets are usually used by families and other leisure travelers, first-class and unrestricted coach tickets are more often bought by business travelers, who have little choice about when they travel and whose companies pick up the tab. Of course, these elasticities pertain to market demand, and with several airlines competing for customers, the elasticities of demand for each airline will be larger. But the *relative* sizes of elasticities across the three categories of service should be about the same. When elasticities of demand differ so widely, it should not be surprising that airlines set such different fares for different categories of service.

Airline price discrimination has become increasingly sophisticated in the United States. A wide variety of fares is available, depending on how far in advance the ticket is bought, the percentage of the fare that is refundable if the trip is changed or cancelled, and whether the trip includes a weekend stay.[7] The objective of the airlines has been to discriminate more finely among travelers with different reservation prices. As one industry executive puts it, "You don't want to sell a seat to a guy for $69 when he is willing to pay $400."[8] At the same time, an airline would rather sell a seat for $69 than leave it empty.

[7] Airlines also allocate the number of seats on each flight that will be available for each fare category. The allocation is based on the total demand and mix of passengers expected for each flight, and can change as the departure of the flight nears and estimates of demand and passenger mix change.

[8] "The Art of Devising Air Fares," *New York Times*, March 4, 1987.

11.3 Intertemporal Price Discrimination and Peak-Load Pricing

intertemporal price discrimination Practice of separating consumers with different demand functions into different groups by charging different prices at different points in time.

peak-load pricing Practice of charging higher prices during peak periods when capacity constraints cause marginal costs to be high.

Two other closely related forms of price discrimination are important and widely practiced. The first of these is **intertemporal price discrimination:** separating consumers with different demand functions into different groups by charging different prices at different points in time. The second is **peak-load pricing:** charging higher prices during peak periods when capacity constraints cause marginal costs to be high. Both of these strategies involve charging different prices at different times, but the reasons for doing so are somewhat different in each case. We will take each in turn.

Intertemporal Price Discrimination

The objective of intertemporal price discrimination is to divide consumers into high-demand and low-demand groups by charging a price that is high at first but falls later. To see how this strategy works, think about how an electronics company might price new, technologically advanced equipment, such as videocassette recorders during the 1970s, compact disc players in the early 1980s, and, more recently, DVD systems. In Figure 11.7, D_1 is the (inelastic) demand curve for a small group of consumers who value the product highly and do not want to wait to buy it (e.g., stereo buffs who value high-quality sound and want the latest equipment). D_2 is the demand curve for the broader group of consumers who are more willing to forgo the product if the price is too high. The strategy, then, is to initially offer the product at the high price P_1, selling mostly to consumers on demand curve D_1. Later, after this first group of consumers has bought the product, the price is lowered to P_2, and sales are made to the larger group of consumers on demand curve D_2.[9]

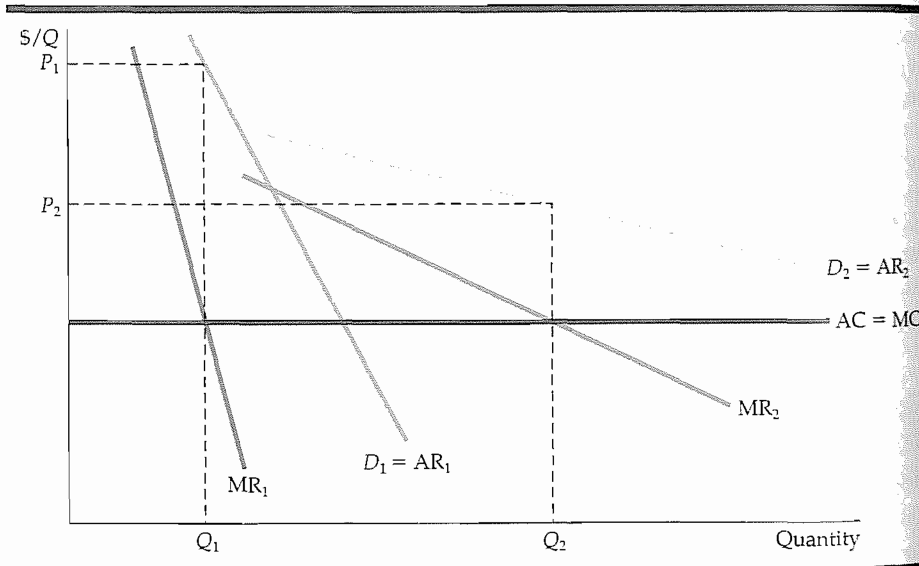

FIGURE 11.7 Intertemporal Price Discrimination

Consumers are divided into groups by changing the price over time. Initially, the price is high. The firm captures surplus from consumers who have a high demand for the good and who are unwilling to wait to buy it. Later, the price is reduced to appeal to the mass market.

There are other examples of intertemporal price discrimination. One involves charging a high price for a first-run movie and then lowering the price after the movie has been out a year. Another, practiced almost universally by publishers, is to charge a high price for the hardcover edition of a book and then to release the paperback version at a much lower price about a year later. Many people think that the lower price of the paperback is due to a much lower cost of production, but this is not true. Once a book has been edited and typeset, the marginal cost of printing an additional copy, whether hardcover or paperback, is quite low, perhaps a dollar or so. The paperback version is sold for much less not because it is much cheaper to print but because high-demand consumers have already purchased the hardbound edition. The remaining consumers—paperback buyers—generally have more elastic demands.

Peak-Load Pricing

Peak-load pricing also involves charging different prices at different points in time. Rather than capturing consumer surplus, however, the objective is to increase economic efficiency by charging consumers prices that are close to marginal cost.

For some goods and services, demand peaks at particular times—for roads and tunnels during commuter rush hours, for electricity during late summer afternoons, and for ski resorts and amusement parks on weekends. Marginal cost is also high during these peak periods because of capacity constraints. Prices should thus be higher during peak periods.

This is illustrated in Figure 11.8, where D_1 is the demand curve for the peak period and D_2 the demand curve for the nonpeak period. The firm sets marginal revenue equal to marginal cost for each period, obtaining the high price P_1 for the peak period and the lower price P_2 for the nonpeak period, selling corresponding quantities Q_1 and Q_2. This strategy increases the firm's profit above what it would be if it charged one price for all periods. It is also more efficient: The sum of producer and consumer surplus is greater because prices are closer to marginal cost.

In §9.2, we explain that economic efficiency means that aggregate consumer and producer surplus is maximized.

The efficiency gain from peak-load pricing is important. If the firm were a regulated monopolist (e.g., an electric utility), the regulatory agency should set the prices P_1 and P_2 at the points where the *demand* curves, D_1 and D_2, intersect the marginal cost curve, rather than where the marginal revenue curves intersect marginal cost. In that case, consumers realize the entire efficiency gain.

Note that peak-load pricing is different from third-degree price discrimination. With third-degree price discrimination, marginal revenue must be equal for each group of consumers and equal to marginal cost. Why? Because the costs of serving the different groups are not independent. For example, with unrestricted versus discounted air fares, increasing the number of seats sold at discounted fares affects the cost of selling unrestricted tickets—marginal cost rises rapidly as the airplane fills up. But this is not so with peak-load pricing (and for that

[9] The prices of new electronic products also come down over time because costs fall as producers start to achieve greater scale economies and move down the learning curve. But even if costs did not fall, producers can make more money by first setting high prices and then reducing them over time, thereby discriminating and capturing consumer surplus.

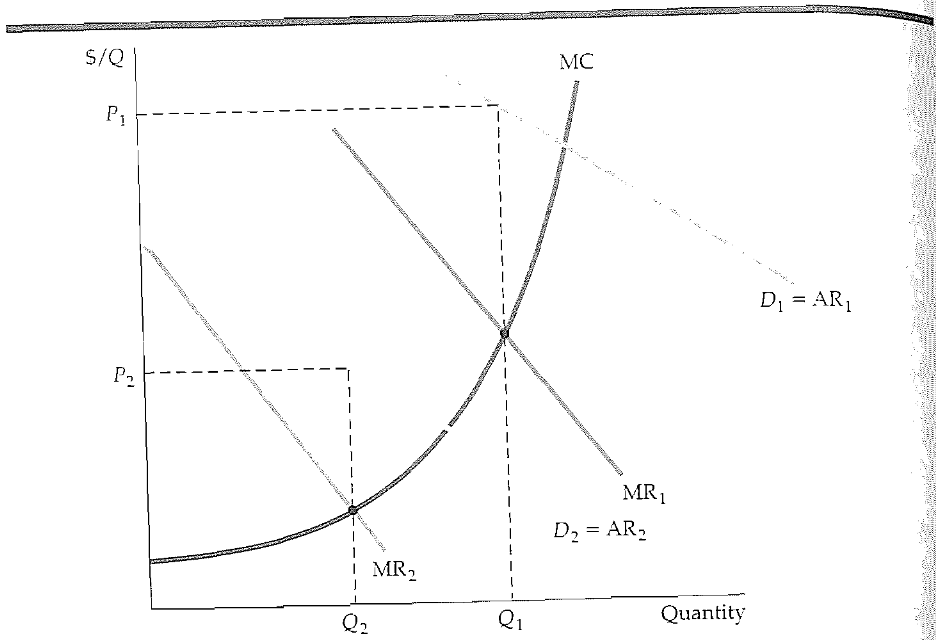

FIGURE 11.8 Peak-Load Pricing

Demands for some goods and services increase sharply during particular times of the day or year. Charging a higher price P_1 during the peak periods is more profitable for the firm than charging a single price at all times. It is also more efficient because marginal cost is higher during peak periods.

matter, with most instances of intertemporal price discrimination). Selling more tickets for the ski lifts or amusement park on a weekday does not significantly raise the cost of selling tickets on the weekend. Similarly, selling more electricity during the off-peak period will not significantly increase the cost of selling electricity during the peak period. As a result, price and sales in each period can be determined independently by setting marginal cost equal to marginal revenue for each period.

Movie theaters, which charge more for evening shows than for matinees, are another example. For most movie theaters, the marginal cost of serving customers during the matinee is independent of the marginal cost during the evening. The owner of a movie theater can determine the optimal prices for the evening and matinee shows independently, using estimates of demand in each period along with estimates of marginal cost.

EXAMPLE 11.3 How to Price a Best-Selling Novel

Publishing both hardbound and paperback editions of a book allows publishers to price discriminate. As they do with most goods, consumers differ considerably in their willingness to pay for books. For example, some consumers want to buy a new bestseller as soon as it is released, even if the price is \$25. Other consumers, however, will wait a year until the book is available in paperback for \$10. But how does a publisher decide that \$25 is the right price for the new hardbound edition and \$10 is the right price for the paperback edition? And how long should it wait before bringing out the paperback edition?

The key is to divide consumers into two groups, so that those who are willing to pay a high price do so and only those unwilling to pay a high price wait and buy the paperback. This means that significant time must be allowed to pass before the paperback is released. If consumers know that the paperback will be available within a few months, they will have little incentive to buy the hardbound edition.[10] On the other hand, if the publisher waits too long to bring out the paperback edition, interest will wane and the market will dry up. As a result, publishers typically wait 12 to 18 months before releasing paperback editions.

What about price? Setting the price of the hardbound edition is difficult because, except for a few authors whose books always seem to sell, publishers have little data with which to estimate demand for a book that is about to be published. Often, they can judge only from the past sales of similar books. But usually only aggregate data are available for each category of book. Most new novels, therefore, are released at similar prices. It is clear, however, that those consumers willing to wait for the paperback edition have demands that are far more elastic than those of bibliophiles. It is not surprising, then, that paperback editions sell for so much less than hardbound ones.[11]

11.4 The Two-Part Tariff

two-part tariff Form of pricing in which consumers are charged both an entry and a usage fee.

The **two-part tariff** is related to price discrimination and provides another means of extracting consumer surplus. It requires consumers to pay a fee up front for the right to buy a product. Consumers then pay an additional fee for each unit of the product they wish to consume. The classic example of this is an amusement park.[12] You pay an admission fee to enter, and you also pay a certain amount for each ride. The owner of the park must decide whether to charge a high entrance fee and a low price for the rides or, alternatively, to admit people for free but charge high prices for the rides.

The two-part tariff has been applied in many settings: tennis and golf clubs (you pay an annual membership fee plus a fee for each use of a court or round of golf); the rental of large mainframe computers (a flat monthly fee plus a fee for each unit of processing time consumed); telephone service (a monthly hook-up fee plus a fee for message units). The strategy also applies to the sale of Polaroid cameras (you pay for the camera, which lets you productively consume the film, which you pay for by the package) and safety razors (you pay for the razor, which lets you consume the blades that fit only that brand of razor).

[10] Some consumers will buy the hardbound edition even if the paperback is already available because it is more durable and more attractive on a bookshelf. This must be taken into account when setting prices, but it is of secondary importance compared with intertemporal price discrimination.

[11] Hardbound and paperback editions are often published by different companies. The author's agent auctions the rights to the two editions, but the contract for the paperback specifies a delay to protect the sales of the hardbound edition. The principle still applies, however. The length of the delay and the prices of the two editions are chosen to price discriminate intertemporally.

[12] This pricing strategy was first analyzed by Walter Oi, "A Disneyland Dilemma: Two-Part Tariffs for a Mickey Mouse Monopoly," *Quarterly Journal of Economics* (February 1971): 77–96.

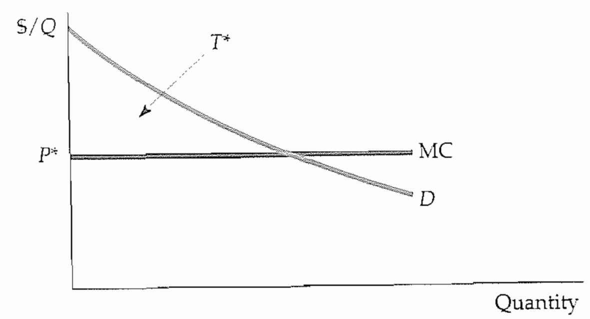

FIGURE 11.9 Two-Part Tariff with a Single Consumer

The consumer has demand curve D. The firm maximizes profit by setting usage fee P equal to marginal cost and entry fee T equal to the entire surplus of the consumer.

The problem for the firm is how to set the *entry fee* (which we denote by T) versus the *usage fee* (which we denote by P). Assuming that the firm has some market power, should it set a high entry fee and low usage fee, or vice versa? To see how a firm can solve this problem, we need to understand the basic principles involved.

Single Consumer Let us begin with an artificial but simple case. Suppose there is only one consumer in the market (or many consumers with identical demand curves). Suppose also that the firm knows this consumer's demand curve. Now, remember that the firm wants to capture as much consumer surplus as possible. In this case, the solution is straightforward: Set the usage fee P equal to marginal cost and the entry fee T equal to the total consumer surplus for each consumer. Thus, in Figure 11.9, the consumer pays T^* (or a bit less) to use the product, and $P^* = MC$ per unit consumed. With the fees set in this way, the firm captures *all* the consumer surplus as its profit.

Two Consumers Now, suppose there are two different consumers (or two groups of identical consumers). The firm, however, can set only *one* entry fee and one usage fee. It would thus no longer want to set the usage fee equal to marginal cost. If it did, it could make the entry fee no larger than the consumer surplus of the consumer with the smaller demand (or else it would lose that consumer), and this would not yield a maximum profit. Instead, the firm should set the usage fee *above* marginal cost and then set the entry fee equal to the remaining consumer surplus of the consumer with the smaller demand.

Figure 11.10 illustrates this. With the optimal usage fee at P^* greater than MC, the firm's profit is $2T^* + (P^* - MC)(Q_1 + Q_2)$. (There are two consumers, and each pays T^*.) You can verify that this profit is more than twice the area of triangle ABC, the consumer surplus of the consumer with the smaller demand when $P = MC$. To determine the exact values of P^* and T^*, the firm would need to know (in addition to its marginal cost) the demand curves D_1 and D_2. It would then write down its profit as a function of P and T and choose the two prices that maximize this function. (See Exercise 10 for an example of how to do this.)

Many Consumers Most firms, however, face a variety of consumers with different demands. Unfortunately, there is no simple formula to calculate the optimal two-part tariff in this case, and some trial and error might be required.

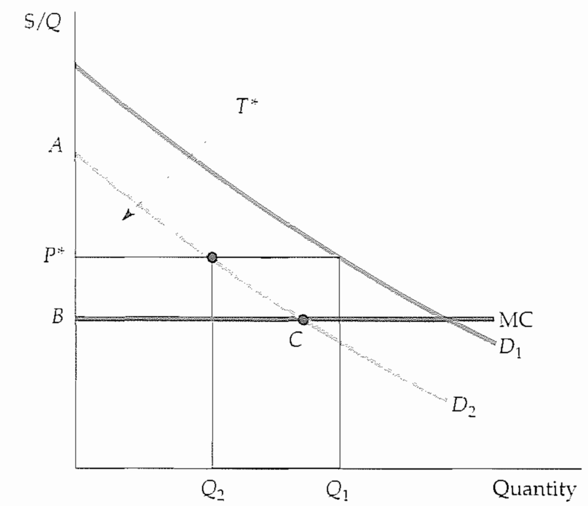

FIGURE 11.10 Two-Part Tariff with Two Consumers

The profit-maximizing usage fee P^* will exceed marginal cost. The entry fee T^* is equal to the surplus of the consumer with the smaller demand. The resulting profit is $2T^* + (P^* - MC)(Q_1 + Q_2)$. Note that this profit is larger than twice the area of triangle ABC.

But there is always a trade-off: A lower entry fee means more entrants and thus more profit from sales of the item. However, as the entry fee becomes smaller and the number of entrants larger, the profit derived from the entry fee will fall. The problem, then, is to pick an entry fee that results in the optimum number of entrants—that is, the fee that allows for maximum profit. In principle, we can do this by starting with a price for sales of the item P, finding the optimum entry fee T, and then estimating the resulting profit. The price P is then changed, and the corresponding entry fee calculated, along with the new profit level. By iterating this way, we can approach the optimal two-part tariff.

Figure 11.11 illustrates this principle. The firm's profit π is divided into two components, each of which is plotted as a function of the entry fee T, assuming a fixed sales price P. The first component, π_a, is the profit from the entry fee and is equal to the revenue $n(T)T$, where $n(T)$ is the number of entrants. (Note that a high T implies a small n.) Initially, as T is increased from zero, revenue $n(T)T$ rises. Eventually, however, further increases in T will make n so small that $n(T)T$ falls. The second component, π_s, is the profit from sales of the item itself at price P and is equal to $(P - MC)Q$, where Q is the rate at which entrants purchase the item. Q will be larger the larger the number of entrants n. Thus π_s falls when T is increased because a higher T reduces n.

Starting with a number for P, we determine the optimal (profit-maximizing) T^*. We then change P, find a new T^*, and determine whether profit is now higher or lower. This procedure is repeated until profit has been maximized.

Obviously, more data are needed to design an optimal two-part tariff than to choose a single price. Knowing marginal cost and the aggregate demand curve is not enough. It is impossible (in most cases) to determine the demand curve of every consumer, but one would at least like to know by how much individual demands differ from one another. If consumers' demands for your product are

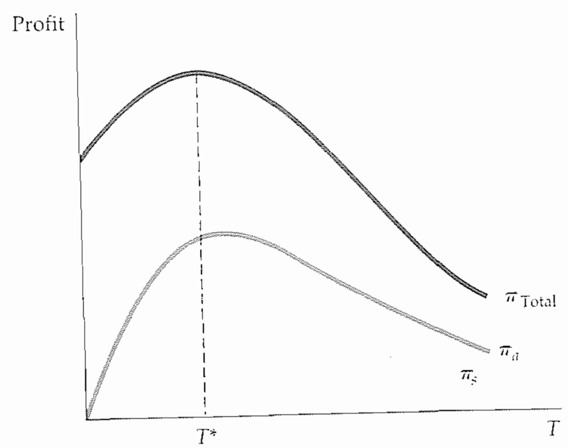

FIGURE 11.11 Two-Part Tariff with Many Different Consumers

Total profit π is the sum of the profit from the entry fee π_a and the profit from sales π_s. Both π_a and π_s depend on T, the entry fee. Therefore:

$$\pi = \pi_a + \pi_s = n(T)T + (P - \text{MC})Q(n)$$

where n is the number of entrants, which depends on the entry fee T, and Q is the rate of sales, which is greater the larger is n. Here T^* is the profit-maximizing entry fee, given P. To calculate optimum values for P and T, we can start with a number for P, find the optimum T, and then estimate the resulting profit. P is then changed and the corresponding T is recalculated, along with the new profit level.

fairly similar, you would want to charge a price P that is close to marginal cost and make the entry fee T large. This is the ideal situation from the firm's point of view because most of the consumer surplus could then be captured. On the other hand, if consumers have different demands for your product, you would probably want to set P substantially above marginal cost and charge a lower entry fee T. In that case, however, the two-part tariff is a much less effective means of capturing consumer surplus; setting a single price may do almost as well.

At Disneyland in California and Disneyworld in Florida, the strategy is to charge a high entry fee and charge nothing for the rides. This policy makes sense because consumers have reasonably similar demands for Disney vacations. Most people visiting the parks plan daily budgets (including expenditures for food and beverages) that, for the majority of consumers, do not differ very much.

Firms are perpetually searching for innovative pricing strategies, and a few have devised and introduced a two-part tariff with a "twist"—the entry fee T entitles the customer to a certain number of free units. For example, if you buy a Gillette razor, several blades are usually included in the package. And the monthly lease fee for a mainframe computer usually includes some free usage before usage is charged. This twist lets the firm set a higher entry fee T without losing as many small customers. Because these small customers might pay little or nothing for usage under this scheme, the higher entry fee will capture their surplus without driving them out of the market, while also capturing more of the surplus of the large customers.

EXAMPLE 11.4 Polaroid Cameras

In 1971, Polaroid introduced its new SX-70 camera. This camera was sold, not leased, to individual consumers. Nevertheless, because film was sold separately, Polaroid could apply a two-part tariff to the pricing of the SX-70. Let's see how this pricing strategy gave Polaroid greater profits than would have been possible if its camera had used ordinary roll film, and how Polaroid might have determined the optimal prices for each part of its two-part tariff. Some time later, Kodak entered the market with a competing self-developing film and camera. We will also consider the effect of Kodak's entry into the market on Polaroid's prices and profits.

Why did the pricing of Polaroid's cameras and film involve a two-part tariff? Because Polaroid had a monopoly on both its camera and the film, only Polaroid film could be used in the camera. Consumers bought the camera and film to take instant pictures: The camera was the "entry fee" that provided access to the consumption of instant pictures, which was what consumers ultimately demanded.[13] In this sense, the price of the camera was like the entry fee at an amusement park. However, while the marginal cost of allowing someone entry into the park is close to zero, the marginal cost of producing a camera is significantly above zero, and thus had to be taken into account when designing the two-part tariff.

It was important that Polaroid have a monopoly on the film as well as the camera. If the camera had used ordinary roll film, competitive forces would have pushed the price of film close to its marginal cost. If all consumers had identical demands, Polaroid could still have captured all the consumer surplus by setting a high price for the camera (equal to the surplus of each consumer). But in practice, consumers were heterogeneous, and the optimal two-part tariff required a price for the film well above marginal cost. (In fact Polaroid got—and still gets—most of its profits from film rather than cameras.) Polaroid needed its monopoly on the film to maintain this high price.

How should Polaroid have selected its prices for the camera and film? It could have begun with some analytical spadework. Its profit is given by

$$\pi = PQ + nT - C_1(Q) - C_2(n)$$

where P is the price of the film, T the price of the camera, Q the quantity of film sold, n the number of cameras sold, and $C_1(Q)$ and $C_2(n)$ the costs of producing film and cameras, respectively.

Polaroid wanted to maximize its profit π, taking into account that Q and n depend on P and T. Given a heterogeneous base of potential consumers, this dependence on P and T might only have been guessed at initially, drawing on knowledge of related products. Later, a better understanding of demand and of how Q and n depend on P and T might have been possible as the firm accumulated data from its sales experience. Knowledge of C_1 and C_2 may have been

[13] We are simplifying here. In fact, some consumers obtain utility just from owning the camera, even if they take few or no pictures. Adults, like children, enjoy new toys and can obtain pleasure from the mere possession of a technologically innovative good.

easier to come by, perhaps from engineering and statistical studies (as discussed in Chapter 7).

Given some initial guesses or estimates for $Q(P)$, $n(T)$, $C_1(Q)$, and $C_2(n)$, Polaroid could have calculated the profit-maximizing prices P and T. It could also have determined how sensitive these prices were to uncertainty over demand and cost. This could have provided a guideline for trial-and-error pricing experiments. Over time these experiments would also have told Polaroid more about demand and cost, so that it could refine its two-part tariff accordingly.[14]

Did the entry of Kodak with a competing instant camera and film mean that Polaroid lost its ability to use a two-part tariff to extract consumer surplus? No—only Polaroid film could be used in Polaroid cameras, and Polaroid still had some monopoly power to exploit. However, its monopoly power was reduced, the amount of consumer surplus that could potentially be extracted was smaller, and prices had to be changed. With demand now more elastic, Polaroid would have wanted to reduce the price of its cameras significantly (and indeed it did). In 1984, the courts ruled that Kodak's camera and film involved a patent infringement, and Kodak was forced to withdraw from the instant-picture market in 1985. Polaroid took advantage of this situation by introducing new cameras and films to appeal to different consumers.

In 1996, Polaroid's One Step cameras sold for \$35 to \$60 and used Polaroid 600 film, which was priced at about \$14 per pack of 10 pictures. Polaroid's higher-end Spectra cameras sold for above \$100 and used Spectra film, priced at about \$13 per pack. These film prices were well above marginal cost, reflecting the considerable heterogeneity of consumer demands. In 1999 Polaroid introduced its I-Zone camera and film, which takes matchbook-size pictures. The camera was priced at \$25 and the film at \$7 per pack.

EXAMPLE 11.5 Pricing Cellular Phone Service

Most telephone service is priced using a two-part tariff: a monthly access fee, which may include some free minutes, plus a per-minute charge for additional minutes of usage. This is also true for cellular phone service, which grew explosively during the 1990s, both in the United States and around the world. In the case of cellular service, providers have taken the two-part tariff and turned it into an art form.

In most parts of the United States, consumers can choose between two or more cellular providers that offer local service within the region. In the Boston area, for example, consumers can choose between Bell Atlantic and Cellular One. The service area might have a radius of 50 or 100 miles. If a consumer places calls outside that service area, the call is picked up by a different provider and the consumer pays so-called "roaming" charges. Alternatively, a consumer can choose cellular service with a national provider such as AT&T or Sprint. With these providers, the service area is most of the United States, so there are few or no "roaming" charges.

Most consumers, therefore, have at least three or four cellular providers to choose from. Providers compete among themselves for customers, but each has some market power. This market power arises in part from oligopolistic pricing and output decisions, as will be explained in Chapters 12 and 13. Market power also arises because consumers face switching costs: When they sign up for a cellular plan, they must typically make a commitment to stay with it for at least one year.

Because providers have market power, they must think carefully about profit-maximizing pricing strategies. The two-part tariff provides an ideal means by which cellular providers can capture consumer surplus and turn it into profit.

Table 11.3 shows cellular rate plans (for 1999) for the digital services offered by two providers. The first, Bell Atlantic, offers local service in the Boston area. The second, AT&T, is a national provider.

Note that each provider offers several different plans. The least expensive Bell Atlantic plan has a monthly access charge of just \$19.99; it includes 20 minutes of free air time and a charge of 35 cents per minute beyond the free 20 minutes. Other Bell Atlantic plans have higher monthly access charges but offer larger amounts of free monthly minutes and lower per-minute charges for additional minutes. The most expensive plan has a monthly access charge of \$199.99 but offers 2500 free minutes and charges only 15 cents per additional minute. AT&T likewise has several different plans, although the variation in prices is not as great as with Bell Atlantic.

Why do these cellular phone providers offer several different plans? Why don't they simply offer a single two-part tariff with a monthly access charge and a per-minute usage charge? Offering several different plans allows

[14] Setting prices for a product such as a Polaroid camera is clearly not a simple matter. We have ignored the *dynamic* behavior of cost and demand: namely, how production costs fall as the firm moves down its learning curve and how demand changes over time as the market becomes saturated.

TABLE 11.3 Cellular Rate Plans (1999)

Plan	Monthly Access Charge	Airtime Minutes Included	Additional Minutes
A. BELL ATLANTIC DIGITAL CHOICE			
DC20	\$19.99	20	\$.35
DC90	29.99	90	.30
DC500	49.99	500	.25
DC1000	89.99	1000	.20
DC2000	149.99	2000	.20
DC2500	199.99	2500	.15
B. AT&T DIGITAL ONE RATE			
600	\$89.99	600	\$.25
1000	119.99	1000	.25
1400	149.99	1400	.25

companies to combine third-degree price discrimination with the two-part tariff. The plans are structured so that consumers sort themselves into groups based on their plan choices. A different two-part tariff is then applied to each group.

To see how this works, consider some of the Bell Atlantic plans. The least expensive plan, DC20, is best suited for someone who uses a cell phone only occasionally and wants to spend as little as possible on the service. The most expensive plan, DC2500, is best suited for a very heavy cellular user, perhaps a salesperson who makes calls from a car throughout the day and wants to minimize per-minute cost. Other plans, such as DC500 or DC1000, are better suited for consumers with moderate needs.

Consumers will choose a plan that best matches their needs. Thus they will sort themselves into groups, and the consumers in each group will be relatively homogenous in terms of demands for cellular service. Remember that the two-part tariff works best when consumers have identical or very similar demands. (Recall from Figure 11.9 that with identical consumers, the two-part tariff can be used to capture *all* consumer surplus.) Creating a situation in which consumers sort themselves into groups in this way makes best use of the two-part tariff.

*11.5 Bundling

You have probably seen the 1939 film, *Gone with the Wind*. It is a classic that is nearly as popular now as it was then.[15] Yet we would guess that you have not seen *Getting Gertie's Garter*, a flop that the same film company (Loews) also produced in 1939. And we would also guess that you did not know that these two films were priced in what was then an unusual and innovative way.[16]

Movie theaters that leased *Gone with the Wind* also had to lease *Getting Gertie's Garter*. (Movie theaters pay the film companies or their distributors a daily or weekly fee for the films they lease.) In other words, these two films were **bundled**—i.e., sold as a package. Why would the film company do this?

bundling Practice of selling two or more products as a package.

You might think that the answer is obvious: *Gone with the Wind* was a great film and *Gertie* was a lousy film, so bundling the two forced movie theaters to lease *Gertie*. But this answer doesn't make economic sense. Suppose a theater's reservation price (the maximum price it will pay) for *Gone with the Wind* is $12,000 per week, and its reservation price for *Gertie* is $3000 per week. Then the most it would pay for *both* films is $15,000, whether it takes the films individually or as a package.

Bundling makes sense when *customers have heterogeneous demands* and when the firm cannot price discriminate. With films, different movie theaters serve different groups of patrons and therefore different theaters may face different demands for films. For example, different theaters might appeal to different age groups, who in turn have different relative film preferences.

[15] Adjusted for inflation, *Gone with the Wind* was also the largest grossing film of all time. *Titanic*, released in 1997, made $601 million. *Gone with the Wind* grossed $81.5 million in 1939 dollars, which is equivalent to $941 million in 1997 dollars.

[16] For those readers who claim to know all this, our final trivia question is: Who played the role of Gertie in *Getting Gertie's Garter*?

To see how a film company can use this heterogeneity to its advantage, suppose that there are *two* movie theaters and that their reservation prices for our two films are as follows:

	GONE WITH THE WIND	GETTING GERTIE'S GARTER
Theater *A*	$12,000	$3000
Theater *B*	$10,000	$4000

If the films are rented separately, the maximum price that could be charged for *Wind* is $10,000 because charging more would exclude Theater *B*. Similarly, the maximum price that could be charged for *Gertie* is $3000. Charging these two prices would yield $13,000 from each theater, for a total of $26,000 in revenue. But suppose the films are *bundled*. Theater *A* values the *pair* of films at $15,000 ($12,000 + $3000), and Theater *B* values the pair at $14,000 ($10,000 + $4000). Therefore, we can charge each theater $14,000 for the pair of films and earn a total revenue of $28,000. Clearly, we can earn more revenue ($2000 more) by bundling the films.

Relative Valuations

Why is bundling more profitable than selling the films separately? Because (in this example) the *relative valuations* of the two films are reversed. In other words, although both theaters would pay much more for *Wind* than for *Gertie*, Theater *A* would pay more than Theater *B* for *Wind* ($12,000 vs. $10,000), while Theater *B* would pay more than Theater *A* for *Gertie* ($4000 vs. $3000). In technical terms, we say that the demands are *negatively correlated*—the customer willing to pay the most for *Wind* is willing to pay the least for *Gertie*. To see why this is critical, suppose demands were *positively correlated*—that is, Theater *A* would pay more for *both* films:

	GONE WITH THE WIND	GETTING GERTIE'S GARTER
Theater *A*	$12,000	$4000
Theater *B*	$10,000	$3000

The most that Theater *A* would pay for the pair of films is now $16,000, but the most that Theater *B* would pay is only $13,000. Thus if we bundled the films, the maximum price that could be charged for the package is $13,000, yielding a total revenue of $26,000, the same as by selling the films separately.

Now, suppose a firm is selling two different goods to many consumers. To analyze the possible advantages of bundling, we will use a simple diagram to describe the preferences of the consumers in terms of their reservation prices and their consumption decisions given the prices charged. In Figure 11.12 the horizontal axis is r_1, which is the reservation price of a consumer for good 1, and the vertical axis is r_2, which is the reservation price for good 2. The figure shows the reservation prices for three consumers. Consumer *A* is willing to pay up to $3.25 for good 1 and up to $6 for good 2; consumer *B* is willing to pay up to $8.25

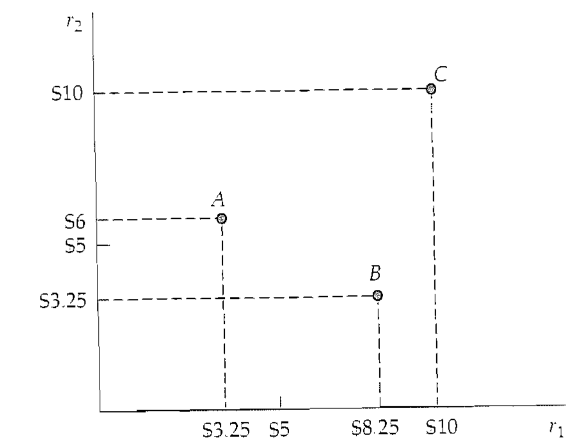

FIGURE 11.12 Reservation Prices

Reservation prices r_1 and r_2 for two goods are shown for three consumers, labeled *A*, *B*, and *C*. Consumer *A* is willing to pay up to \$3.25 for good 1 and up to \$6 for good 2.

for good 1 and up to \$3.25 for good 2; and consumer *C* is willing to pay up to \$10 for each of the goods. In general, the reservation prices for any number of consumers can be plotted this way.

Suppose that there are many consumers and that the products are sold separately, at prices P_1 and P_2, respectively. Figure 11.13 shows how consumers can be divided into groups. Consumers in region I of the graph have reservation prices that are above the prices being charged for each of the goods, so they will

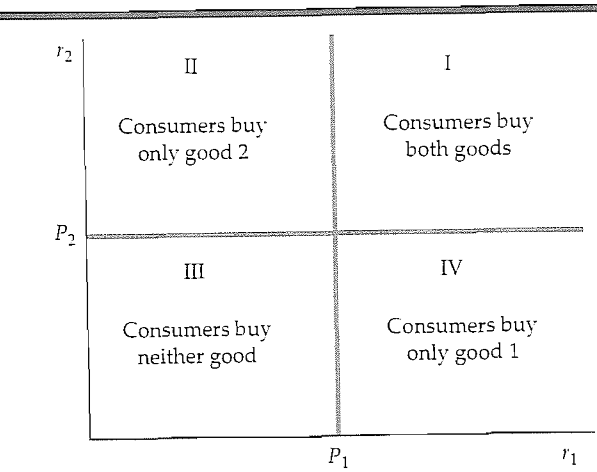

FIGURE 11.13 Consumption Decisions When Products Are Sold Separately

The reservation prices of consumers in region I exceed the prices P_1 and P_2 for the two goods, so these consumers buy both goods. Consumers in regions II and IV buy only one of the goods, and consumers in region III buy neither good.

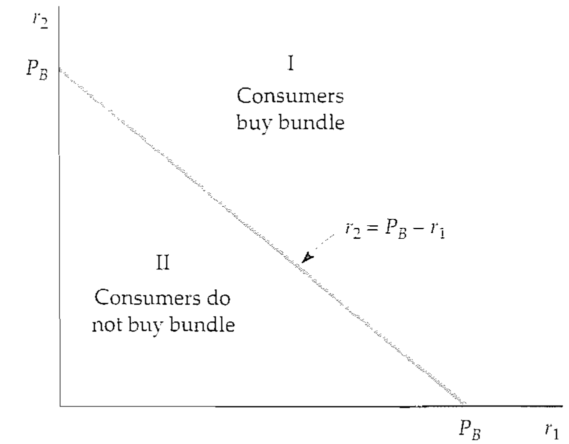

FIGURE 11.14 Consumption Decisions When Products Are Bundled

Consumers compare the *sum* of their reservation prices, $r_1 + r_2$, with the price of the bundle P_B. They buy the bundle only if $r_1 + r_2$ is at least as large as P_B.

buy both goods. Consumers in region II have a reservation price for good 2 that is above P_2, but a reservation price for good 1 that is below P_1; they will buy only good 2. Similarly, consumers in region IV will buy only good 1. Finally, consumers in region III have reservation prices below the prices charged for each of the goods, and so will buy neither.

Now suppose the goods are sold only as a bundle, for a total price of P_B. We can then divide the graph into two regions, as in Figure 11.14. Any given consumer will buy the bundle only if its price is less than or equal to the sum of that consumer's reservation prices for the two goods. The dividing line is therefore the equation $P_B = r_1 + r_2$ or, equivalently, $r_2 = P_B - r_1$. Consumers in region I have reservation prices that add up to more than P_B, so they will buy the bundle. Consumers in region II, who have reservation prices that add up to less than P_B, will not buy the bundle.

Depending on the prices, some of the consumers in region II of Figure 11.14 might have bought one of the goods if they had been sold separately. These consumers are lost to the firm, however, when it sells the goods only as a bundle. The firm, then, must determine whether it can do better by bundling.

In general, the effectiveness of bundling depends on the extent to which demands are negatively correlated. In other words, it works best when consumers who have a high reservation price for good 1 have a low reservation price for good 2, and vice versa. Figure 11.15 shows two extremes. In part (a), each point represents the two reservation prices of a consumer. Note that the demands for the two goods are perfectly positively correlated—consumers with a high reservation price for good 1 also have a high reservation price for good 2. If the firm bundles and charges a price $P_B = P_1 + P_2$, it will make the same profit that it would make by selling the goods separately at prices P_1 and P_2. In part (b), on the other hand, demands are perfectly negatively correlated—a higher reservation price for good 2 implies a proportionately lower one for good 1. In this case, bundling is the ideal strategy. By charging the price P_B shown in the figure, the firm can capture *all* the consumer surplus.

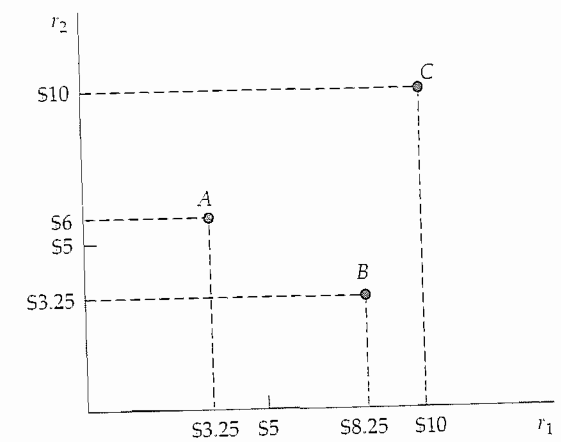

FIGURE 11.12 Reservation Prices

Reservation prices r_1 and r_2 for two goods are shown for three consumers, labeled A, B, and C. Consumer A is willing to pay up to $3.25 for good 1 and up to $6 for good 2.

for good 1 and up to $3.25 for good 2; and consumer C is willing to pay up to $10 for each of the goods. In general, the reservation prices for any number of consumers can be plotted this way.

Suppose that there are many consumers and that the products are sold separately, at prices P_1 and P_2, respectively. Figure 11.13 shows how consumers can be divided into groups. Consumers in region I of the graph have reservation prices that are above the prices being charged for each of the goods, so they will buy both goods. Consumers in region II have a reservation price for good 2 that is above P_2, but a reservation price for good 1 that is below P_1; they will buy only good 2. Similarly, consumers in region IV will buy only good 1. Finally, consumers in region III have reservation prices below the prices charged for each of the goods, and so will buy neither.

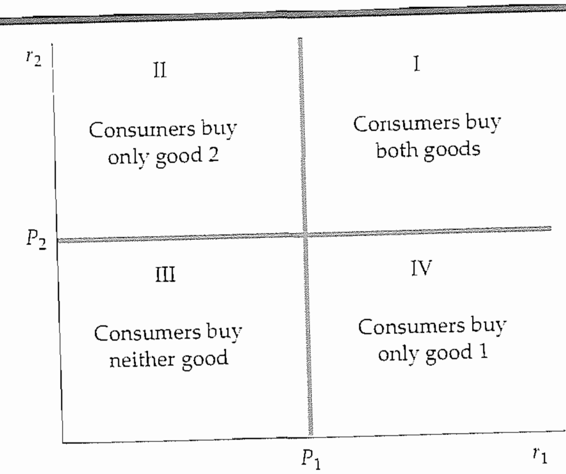

FIGURE 11.13 Consumption Decisions When Products Are Sold Separately

The reservation prices of consumers in region I exceed the prices P_1 and P_2 for the two goods, so these consumers buy both goods. Consumers in regions II and IV buy only one of the goods, and consumers in region III buy neither good.

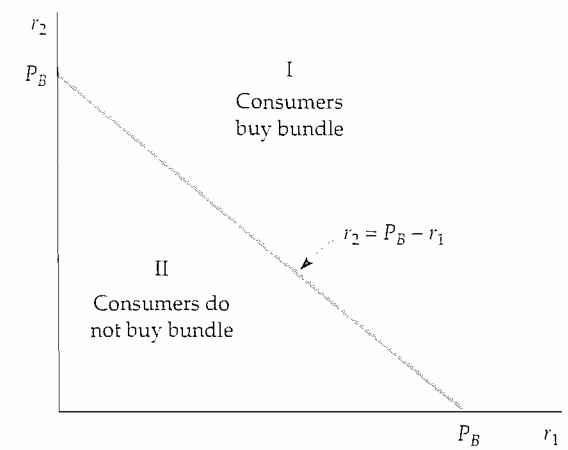

FIGURE 11.14 Consumption Decisions When Products Are Bundled

Consumers compare the *sum* of their reservation prices, $r_1 + r_2$, with the price of the bundle P_B. They buy the bundle only if $r_1 + r_2$ is at least as large as P_B.

Now suppose the goods are sold only as a bundle, for a total price of P_B. We can then divide the graph into two regions, as in Figure 11.14. Any given consumer will buy the bundle only if its price is less than or equal to the sum of that consumer's reservation prices for the two goods. The dividing line is therefore the equation $P_B = r_1 + r_2$ or, equivalently, $r_2 = P_B - r_1$. Consumers in region I have reservation prices that add up to more than P_B, so they will buy the bundle. Consumers in region II, who have reservation prices that add up to less than P_B, will not buy the bundle.

Depending on the prices, some of the consumers in region II of Figure 11.14 might have bought one of the goods if they had been sold separately. These consumers are lost to the firm, however, when it sells the goods only as a bundle. The firm, then, must determine whether it can do better by bundling.

In general, the effectiveness of bundling depends on the extent to which demands are negatively correlated. In other words, it works best when consumers who have a high reservation price for good 1 have a low reservation price for good 2, and vice versa. Figure 11.15 shows two extremes. In part (a), each point represents the two reservation prices of a consumer. Note that the demands for the two goods are perfectly positively correlated—consumers with a high reservation price for good 1 also have a high reservation price for good 2. If the firm bundles and charges a price $P_B = P_1 + P_2$, it will make the same profit that it would make by selling the goods separately at prices P_1 and P_2. In part (b), on the other hand, demands are perfectly negatively correlated—a higher reservation price for good 2 implies a proportionately lower one for good 1. In this case, bundling is the ideal strategy. By charging the price P_B shown in the figure, the firm can capture *all* the consumer surplus.

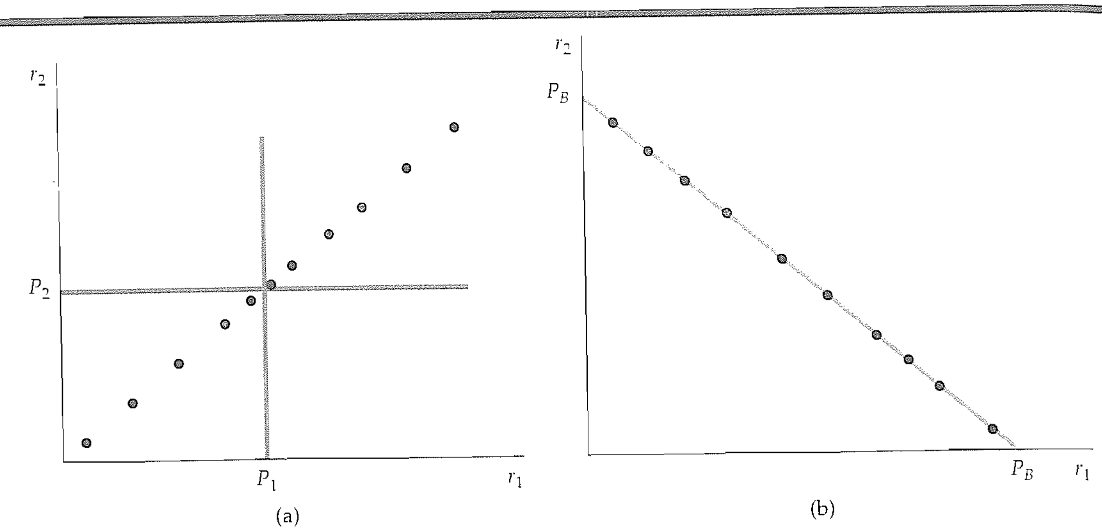

FIGURE 11.15 Reservation Prices

In **(a)**, demands are perfectly positively correlated, so the firm does not gain by bundling. It would earn the same profit by selling the goods separately. In **(b)**, demands are perfectly negatively correlated. Bundling is the ideal strategy—all the consumer surplus can be extracted.

Figure 11.16, which shows the movie example that we introduced at the beginning of this section, illustrates how the demands of the two movie theaters are negatively correlated. (Theater A will pay relatively more for *Gone with the Wind*, but Theater B will pay relatively more for *Getting Gertie's Garter*.) This makes it more profitable to rent the films as a bundle, priced at $14,000.

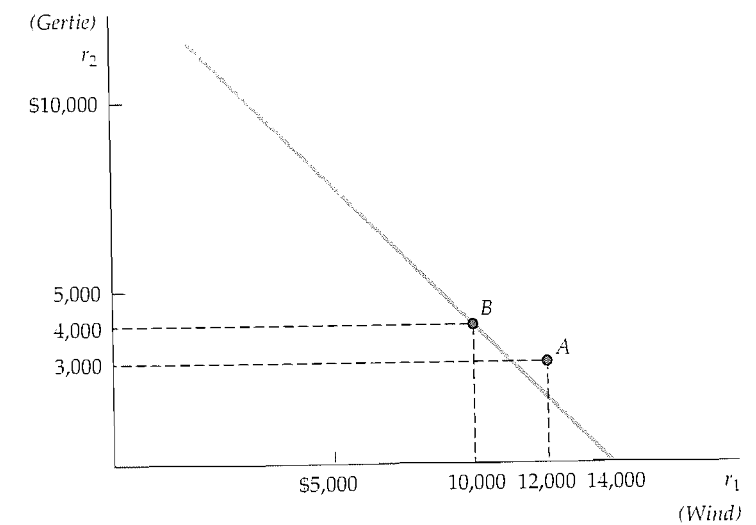

FIGURE 11.16 Movie Example

Consumers A and B are two movie theaters. The diagram shows their reservation prices for the films *Gone with the Wind* and *Getting Gertie's Garter*. Since the demands are negatively correlated, bundling pays.

Mixed Bundling

So far, we have assumed that the firm has two options: to sell the goods either separately or as a bundle. But there is a third option, called **mixed bundling.** As the name suggests, the firm offers its products *both* separately and as a bundle, with a package price below the sum of the individual prices. (We use the term **pure bundling** to refer to the strategy of selling the products *only* as a bundle.) Mixed bundling is often the ideal strategy when demands are only somewhat negatively correlated and/or when marginal production costs are significant. (Thus far, we have assumed that marginal production costs are zero.)

mixed bundling Practice of selling two or more goods both as a package and individually.

pure bundling Practice of selling products only as a package.

In Figure 11.17, mixed bundling is the most profitable strategy. Although demands are perfectly negatively correlated, there are significant marginal costs. (The marginal cost of producing good 1 is $20, and the marginal cost of producing good 2 is $30.) We have four consumers, labeled A through D. Now, let's compare three strategies:

1. Selling the goods separately at prices $P_1 = \$50$ and $P_2 = \$90$
2. Selling the goods only as a bundle at a price of $100
3. Mixed bundling, whereby the goods are offered separately at prices $P_1 = P_2 = \$89.95$, or as a bundle at a price of $100.

Table 11.4 shows these three strategies and the resulting profits. (You can try other prices for P_1, P_2, and P_B to verify that those given in the table maximize

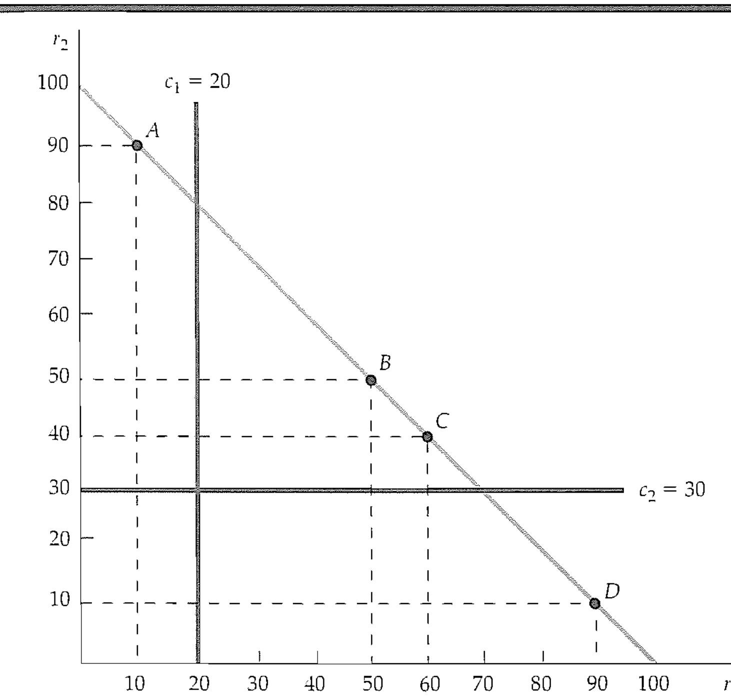

FIGURE 11.17 Mixed versus Pure Bundling

With positive marginal costs, mixed bundling may be more profitable than pure bundling. Consumer A has a reservation price for good 1 that is below marginal cost c_1, and consumer D has a reservation price for good 2 that is below marginal cost c_2. With mixed bundling, consumer A is induced to buy only good 2, and consumer D is induced to buy only good 1, thus reducing the firm's cost.

TABLE 11.4 Bundling Example

	P_1	P_2	P_B	PROFIT
Sell separately	\$50	\$90	—	\$150
Pure bundling	—	—	\$100	\$200
Mixed bundling	\$89.95	\$89.95	\$100	\$229.90

profit for each strategy.) When the goods are sold separately, only consumers *B*, *C*, and *D* buy good 1, and only consumer *A* buys good 2; total profit is 3(\$50 − \$20) + 1(\$90 − \$30) = \$150. With pure bundling, all four consumers buy the bundle for \$100, so that total profit is 4(\$100 − \$20 − \$30) = \$200. As we should expect, pure bundling is better than selling the goods separately because consumers' demands are negatively correlated. But what about mixed bundling? Now consumer *D* buys only good 1 for \$89.95, consumer *A* buys only good 2 for \$89.95, and consumers *B* and *C* buy the bundle for \$100. Total profit is now (\$89.95 − \$20) + (\$89.95 − \$30) + 2(\$100 − \$20 − \$30) = \$229.90.

In this case, mixed bundling is the most profitable strategy, even though demands are perfectly negatively correlated (i.e., all four consumers have reservation prices on the line $r_2 = 100 - r_1$). Why? For each good, marginal production cost exceeds the reservation price of one consumer. For example, consumer *A* has a reservation price of \$90 for good 2 but a reservation price of only \$10 for good 1. Since the cost of producing a unit of good 1 is \$20, the firm would prefer that consumer *A* buy only good 2, not the bundle. It can achieve this by offering good 2 separately for a price just below consumer *A*'s reservation price, while also offering the bundle at a price acceptable to consumers *B* and *C*.

Mixed bundling would *not* be the preferred strategy in this example if marginal costs were zero, because then there would be no benefit in excluding consumer *A* from buying good 1 and consumer *D* from buying good 2. We leave it to you to demonstrate this (see Exercise 12).[17]

If marginal costs are zero, mixed bundling can still be more profitable than pure bundling if consumers' demands are not perfectly negatively correlated. (Recall that in Figure 11.17, the reservation prices of the four consumers are perfectly negatively correlated.) This is illustrated by Figure 11.18, in which we have modified the example of Figure 11.17. In Figure 11.18, marginal costs are zero, but the reservation prices for consumers *B* and *C* are now higher. Let's once again compare three strategies: selling the two goods separately, pure bundling, and mixed bundling.

Table 11.5 shows the optimal prices and the resulting profits for each strategy. (Once again, you should try other prices for P_1, P_2, and P_B to verify that those given in the table maximize profit for each strategy.) When the goods are sold separately, only consumers *C* and *D* buy good 1, and only consumers *A* and *B* buy good 2; total profit is thus \$320. With pure bundling, all four consumers buy the bundle for \$100, so that total profit is \$400. As expected, pure bundling is better than selling the goods separately because consumers' demands are negatively correlated. But mixed bundling is better still. With mixed bundling, consumer *A* buys only good 2, consumer *D* buys only good 1, and consumers *B* and *C* buy the bundle at a price of \$120. Total profit is now \$420.

Why does mixed bundling give higher profits than pure bundling even though marginal costs are zero? The reason is that demands are not perfectly negatively correlated: The two consumers who have high demands for both goods (*B* and *C*) are willing to pay more for the bundle than are consumers *A* and *D*. Hence, with mixed bundling, we can increase the price of the bundle (from \$100 to \$120), sell this bundle to two consumers, and charge the remaining consumers \$90 for a single good.

[17] Sometimes a firm with monopoly power will find it profitable to bundle its product with the product of another firm; see Richard L. Schmalensee, "Commodity Bundling by Single-Product Monopolies," *Journal of Law and Economics* 25 (April 1982): 67–71. Bundling can also be profitable when the products are substitutes or complements. See Arthur Lewbel, "Bundling of Substitutes or Complements," *International Journal of Industrial Organization* 3 (1985): 101–107.

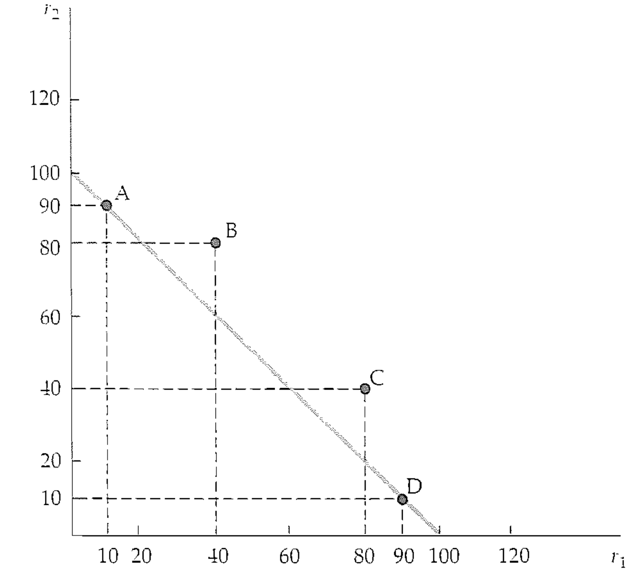

FIGURE 11.18 Mixed Bundling with Zero Marginal Costs

If marginal costs are zero, mixed bundling is still more profitable than pure bundling if consumers' demands are not perfectly negatively correlated. In this example, consumers *B* and *C* are willing to pay \$20 more for the bundle than are consumers *A* and *D*. With pure bundling, the price of the bundle is \$100. With mixed bundling, the price of the bundle can be increased to \$120 and consumers *A* and *D* can still be charged \$90 for a single good.

Bundling in Practice

Bundling is a widely used pricing strategy. When you buy a new car, for example, you can purchase such options as power windows, power seats, or a sunroof separately, or you can purchase a "luxury package" in which these options are bundled. Manufacturers of luxury cars (such as Lexus, BMW, or Infiniti) tend

TABLE 11.5 Mixed Bundling with Zero Marginal Costs

	P_1	P_2	P_B	PROFIT
Sell separately	\$80	\$80	—	\$320
Pure bundling	—	—	\$100	\$400
Mixed bundling	\$90	\$90	\$120	\$420

to include such "options" as standard equipment; this is pure bundling. For more moderately priced cars, however, these items are optional but are usually offered as part of a bundle. Automobile companies must decide which items to include in such bundles and how to price them.

Another example is vacation travel. If you plan a vacation to Europe, you might make your own hotel reservations, buy an airplane ticket, and order a rental car. Alternatively, you might buy a vacation package in which the airfare, land arrangements, hotel, and even meals are all bundled together.

Still another example is cable television. Cable operators typically offer a basic service for a low monthly fee, plus individual "premium" channels, such as Cinemax, Home Box Office, and the Disney Channel on an individual basis for additional monthly fees. However, they also offer packages in which two or more premium channels are sold as a bundle. Bundling cable channels is profitable because demands are negatively correlated. How do we know that? Given that there are only 24 hours in a day, the time a consumer spends watching HBO is time that cannot be spent watching the Disney Channel. Thus consumers with high reservation prices for some channels will have relatively low reservation prices for others.

How can a company decide whether to bundle its products, and determine the profit-maximizing prices? Most companies do not know their customers' reservation prices. However, by conducting market surveys, they may be able to estimate the distribution of reservation prices, and then use this information to design a pricing strategy.

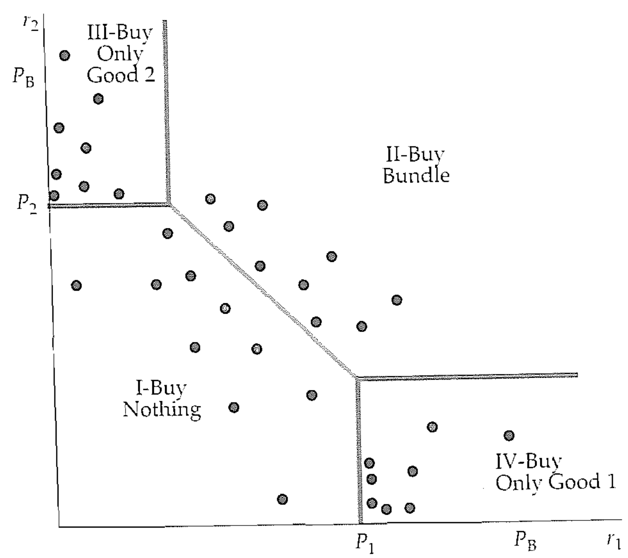

FIGURE 11.19 Mixed Bundling in Practice

The dots in this figure are estimates of reservation prices for a representative sample of consumers. A company could first choose a price for the bundle, P_B, such that a diagonal line connecting these prices passes roughly midway through the dots. The company could then try individual prices P_1 and P_2. Given P_1, P_2, and P_B, profits can be calculated for this sample of consumers. One can then raise or lower P_1, P_2, and P_B and see whether this leads to higher profits. This is done repeatedly until total profit is roughly maximized.

This is illustrated in Figure 11.19. The dots are estimates of reservation prices for a representative sample of consumers (obtained, say, from a market survey). The company might first choose a price for the bundle, P_B, such that a diagonal line connecting these prices passes roughly midway through the dots in the figure. It could then try individual prices P_1 and P_2. Given P_1, P_2, and P_B, we can separate consumers into four regions, as shown in the figure. Consumers in Region I buy nothing (because $r_1 < P_1$, $r_2 < P_2$, and $r_1 + r_2 < P_B$). Consumers in Region II buy the bundle (because $r_1 + r_2 > P_B$). Consumers in Region III buy only good 2 (because $r_2 > P_2$ but $r_1 < P_B - P_2$). Likewise, consumers in Region IV buy only good 1. Given this distribution, we can calculate the resulting profits. We can then raise or lower P_1, P_2, and P_B and see whether doing so leads to higher profits. This can be done repeatedly (on a computer) until prices are found that roughly maximize total profit.

EXAMPLE 11.6 The Complete Dinner versus à la Carte: A Restaurant's Pricing Problem

Many restaurants offer both complete dinners and à la carte menus. Why? Most customers go out to eat knowing roughly how much they are willing to spend for dinner (and choose the restaurant accordingly). Diners, however, have different preferences. For example, some value appetizers highly but could happily skip dessert. Others attach little value to the appetizer but regard dessert as essential. And some customers attach moderate values to both appetizers and desserts. What pricing strategy lets the restaurant capture as much consumer surplus as possible from these heterogeneous customers? The answer, of course, is mixed bundling.

For a restaurant, mixed bundling means offering both complete dinners (the appetizer, main course, and dessert come as a package) and an à la carte menu (the customer buys the appetizer, main course, and dessert separately). This strategy allows the à la carte menu to be priced to capture consumer surplus from customers who value some dishes much more highly than others. (Such customers would correspond to consumers *A* and *D* in Figure 11.17.) At the same time, the complete dinner retains those customers who have lower variations in their reservation prices for different dishes (e.g., customers who attach moderate values to both appetizers and desserts).

For example, if the restaurant expects to attract customers willing to spend about $20 for dinner, it might charge about $5 for appetizers, about $14 for a typical main dish, and about $4 for dessert. It could also offer a complete dinner, which includes an appetizer, main course, and dessert, for $20. Then, the customer who loves dessert but couldn't care less about an appetizer will order only the main dish and dessert, and spend $18 (saving the restaurant the cost of preparing an appetizer). At the same time, another customer who attaches a moderate value (say, $3 or $3.50) to both the appetizer and dessert will buy the complete dinner.

You don't have to go to an expensive French restaurant to experience mixed bundling. Table 11.6 shows the prices of individual items at a Boston-area McDonald's, as well as the prices of "super meals" that include meat or fish items along with a large order of French fries and a large soda. Note that you can buy a Big Mac, a large fries, and a large soda separately for a total of $5.47, or you can buy them as a bundle for $4.19. You say you don't care for fries?

TABLE 11.6 Mixed Bundling at McDonald's (1999)

INDIVIDUAL ITEM	PRICE	MEAL (INCLUDES SODA AND FRIES)	UNBUNDLED PRICE	PRICE OF BUNDLE	SAVING
Grilled Chicken	$2.79	Grilled Chicken	$5.87	$4.78	$1.09
Filet-O-Fish	$2.09	Filet-O-Fish	$5.17	$4.38	$0.79
Cheeseburger	$0.99	Two Cheeseburgers	$5.06	$3.78	$1.28
Double Cheeseburger	$1.95	Double Cheeseburger	$5.03	$3.78	$1.25
Big Mac	$2.39	Big Mac	$5.47	$4.19	$1.28
Quarter Pounder	$2.39	Quarter Pounder	$5.47	$4.19	$1.28
Large French Fries	$1.79				
Large Soda	$1.29				

Then just buy the Big Mac and large soda separately, for a total of $3.68, which is $.51 less than the price of the bundle.

Unfortunately, for consumers, perhaps, creative pricing is sometimes more important than creative cooking for the financial success of a restaurant. Successful restaurateurs know their customers' demand characteristics and use that knowledge to design a pricing strategy that extracts as much consumer surplus as possible.

Tying

tying Practice of requiring a customer to purchase one good in order to purchase another.

Tying is a general term that refers to any requirement that products be bought or sold in some combination. Pure bundling is a common form of tying, but tying can also take other forms. For example, suppose a firm sells a product (such as a copying machine) that requires the consumption of a secondary product (such as paper). The consumer who buys the first product is also required to buy the secondary product from the same company. This requirement is usually imposed through a contract. Note that this is different from the examples of bundling discussed earlier. In those examples, the consumer might have been happy to buy just one of the products. In this case, however, the first product is useless without access to the secondary product.

Why might firms use this kind of pricing practice? One of the main benefits of tying is that it often allows a firm to *meter demand* and thereby practice price discrimination more effectively. During the 1950s, for example, when Xerox had a monopoly on copying machines but not on paper, customers who leased Xerox copiers also had to buy Xerox paper. This allowed Xerox to meter consumption (customers who used a machine intensively bought more paper), and thereby apply a two-part tariff to the pricing of its machines. Also during the 1950s, IBM required customers who leased its mainframe computers to use paper computer cards made only by IBM. By pricing cards well above marginal cost, IBM was effectively charging higher prices for computer usage to customers with larger demands.[18]

[18] However, antitrust actions forced IBM to discontinue this pricing practice

Tying can also be used to extend a firm's market power. As we discussed in Example 10.6, in 1998 the Department of Justice brought suit against Microsoft, claiming that the company had tied its Internet Explorer Web browser to its Windows 98 operating system in order to maintain its monopoly power in the market for PC operating systems.

Tying can have other uses. An important one is to protect customer goodwill connected with a brand name. This is why franchises are often required to purchase inputs from the franchiser. For example, Mobil Oil requires its service stations to sell only Mobil motor oil, Mobil batteries, and so on. Similarly, until recently, a McDonald's franchisee had to purchase all materials and supplies—from the hamburgers to the paper cups—from McDonald's, thus ensuring product uniformity and protecting the brand name.[19]

*11.6 Advertising

We have seen how firms can utilize their market power when making pricing decisions. Pricing is important for a firm, but most firms with market power have another important decision to make: how much to advertise. In this section, we will see how firms with market power can make profit-maximizing advertising decisions, and how those decisions depend on the characteristics of demand for the firm's product.[20]

For simplicity, we will assume that the firm sets only one price for its product. We will also assume that having done sufficient market research, it knows how its quantity demanded depends on both its price P *and* its advertising expenditures in dollars A; that is, it knows $Q(P,A)$. Figure 11.20 shows the firm's demand and cost curves with and without advertising. AR and MR are the firm's average and marginal revenue curves when it does not advertise, and AC and MC are its average and marginal cost curves. It produces a quantity Q_0, where MR = MC, and receives a price P_0. Its profit per unit is the difference between P_0 and average cost, so its total profit π_0 is given by the gray-shaded rectangle.

In §7.2, marginal cost—the increase in cost that results from producing one extra unit of output—is distinguished from average cost—the cost per unit of output.

Now suppose the firm advertises. This causes its demand curve to shift out and to the right; the new average and marginal revenue curves are given by AR' and MR'. Advertising is a fixed cost, so the firm's average cost curve rises (to AC'). Marginal cost, however, remains the same. With advertising, the firm produces Q_1 (where MR' = MC) and receives a price P_1. Its total profit π_1, given by the purple-shaded rectangle, is now much larger.

While the firm in Figure 11.20 is clearly better off advertising, the figure does not help us determine *how much* advertising it should do. It must choose its price P and advertising expenditure A to maximize profit, which is now given by:

$$\pi = PQ(P,A) - C(Q) - A$$

[19] In some cases, the courts have ruled that tying is not necessary to protect customer goodwill and is anticompetitive. Today, a McDonald's franchisee can buy supplies from any McDonald's approved source. For a discussion of some of the antitrust issues involved in franchise tying, see Benjamin Klein and Lester F. Saft, "The Law and Economics of Franchise Tying Contracts," *Journal of Law and Economics* 28 (May 1985): 345–61

[20] A perfectly competitive firm has little reason to advertise, since by definition it can sell as much as it produces at a market price that it takes as given. That is why it would be unusual to see a producer of corn or soybeans advertise.

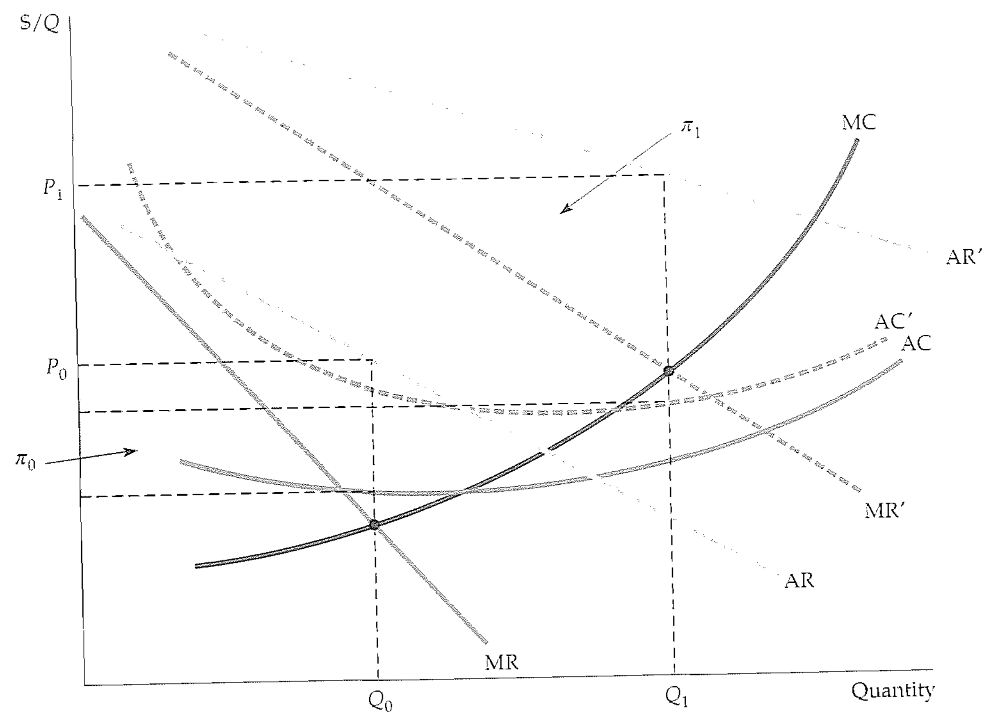

FIGURE 11.20 Effects of Advertising

AR and MR are average and marginal revenue when the firm doesn't advertise, and AC and MC are average and marginal cost. The firm produces Q_0 and receives a price P_0. Its total profit π_0 is given by the gray-shaded rectangle. If the firm advertises, its average and marginal revenue curves shift to the right. Average cost rises (to AC') but marginal cost remains the same. The firm now produces Q_1 (where MR' = MC), and receives a price P_1. Its total profit, π_1, is now larger.

Given a price, more advertising will result in more sales and thus more revenue. But what is the firm's profit-maximizing advertising expenditure? You might be tempted to say that the firm should increase its advertising expenditures until the last dollar of advertising just brings forth an additional dollar of revenue—that is, until the marginal revenue from advertising, $\Delta(P,Q)/\Delta A$, is just equal to 1. But as Figure 11.20 shows, this reasoning omits an important element. Remember that *advertising leads to increased output* (in the figure, output increased from Q_0 to Q_1). But increased output in turn means increased production costs, and this must be taken into account when comparing the costs and benefits of an extra dollar of advertising.

The correct decision is to increase advertising until the marginal revenue from an additional dollar of advertising, MR_{Ads}, just equals the *full* marginal cost of that advertising. That full marginal cost is the sum of the dollar spent directly on the advertising and the marginal production cost resulting from the increased sales that advertising brings about. Thus the firm should advertise up to the point that

$$MR_{Ads} = P\frac{\Delta Q}{\Delta A} = 1 + MC\frac{\Delta Q}{\Delta A} \quad (11.3)$$

$$= \textit{full} \text{ marginal cost of advertising}$$

This rule is often ignored by managers, who justify advertising budgets by comparing the expected benefits (i.e., added sales) only with the cost of the advertising. But additional sales mean increased production costs that must also be taken into account.[21]

A Rule of Thumb for Advertising

Like the rule MR = MC, equation (11.3) is sometimes difficult to apply in practice. In Chapter 10, we saw that MR = MC implies the following rule of thumb for pricing: $(P - MC)/P = -1/E_P$, where E_P is the firm's price elasticity of demand. We can combine this rule of thumb for pricing with equation (11.3) to obtain a rule of thumb for advertising.

In equation (10.1), we offer a rule of thumb for pricing for a profit-maximizing firm—the markup over marginal cost as a percentage of price should equal minus the inverse of the price elasticity of demand.

First, rewrite equation (11.3) as follows:

$$(P - MC)\frac{\Delta Q}{\Delta A} = 1$$

Now multiply both sides of this equation by A/PQ, the **advertising-to-sales ratio:**

advertising-to-sales ratio Ratio of a firm's advertising expenditures to its sales.

$$\frac{P - MC}{P}\left[\frac{A}{Q}\frac{\Delta Q}{\Delta A}\right] = \frac{A}{PQ}$$

The term in brackets, $(A/Q)(\Delta Q/\Delta A)$, is the **advertising elasticity of demand:** the percentage change in the quantity demanded that results from a 1-percent increase in advertising expenditures. We will denote this elasticity by E_A. Because $(P - MC)/P$ must equal $-1/E_P$, we can rewrite this equation as follows:

advertising elasticity of demand Percentage change in quantity demanded resulting from a 1-percent increase in advertising expenditures.

$$A/PQ = -(E_A/E_P) \quad (11.4)$$

Equation (11.4) is a rule of thumb for advertising. It says that to maximize profit, the firm's advertising-to-sales ratio should be equal to minus the ratio of the advertising and price elasticities of demand. Given information (from, say, market research studies) on these two elasticities, the firm can use this rule to check that its advertising budget is not too small or too large.

To put this rule into perspective, assume that a firm is generating sales revenue of \$1 million per year while allocating only \$10,000 (1 percent of its revenues) to advertising. The firm knows that its advertising elasticity of demand is .2, so that a doubling of its advertising budget from \$10,000 to \$20,000 should increase sales by 20 percent. The firm also knows that the price elasticity of demand for its product is −4. Should it increase its advertising budget, knowing that with a price elasticity of demand of −4, its markup of price over marginal cost is substantial? The answer is yes; equation (11.4) tells us that the firm's advertising-to-sales ratio should be $-(.2/-4) = 5$ percent, so the firm should increase its advertising budget from \$10,000 to \$50,000.

This rule makes intuitive sense. It says firms should advertise a lot if (i) demand is very sensitive to advertising (E_A is large), or (ii) demand is not very price elastic (E_P is small). Although (i) is obvious, why should firms advertise

[21] To derive this result using calculus, differentiate $\pi(Q,A)$ with respect to A, and set the derivative equal to zero:

$$\partial\pi/\partial A = P(\partial Q)/\partial A) - MC(\partial Q/\partial A) - 1 = 0$$

Rearranging gives equation (11.3).

more when the price elasticity of demand is small? A small elasticity of demand implies a large markup of price over marginal cost. Therefore, the marginal profit from each extra unit sold is high. In this case, if advertising can help sell a few more units, it will be worth its cost.[22]

EXAMPLE 11.7 Advertising in Practice

In Example 10.2, we looked at the use of markup pricing by supermarkets, convenience stores, and makers of designer jeans. We saw in each case how the markup of price over marginal cost depended on the firm's price elasticity of demand. Now let's see why these firms, as well as producers of other goods, advertise as much (or as little) as they do.

First, supermarkets. We said that the price elasticity of demand for a typical supermarket is around −10. To determine the advertising-to-sales ratio, we also need to know the advertising elasticity of demand. This number can vary considerably depending on what part of the country the supermarket is located in and whether it is in a city, suburb, or rural area. A reasonable range, however, would be 0.1 to 0.3. Substituting these numbers into equation (11.4), we find that the manager of a typical supermarket should have an advertising budget of around 1 to 3 percent of sales—which is indeed what many supermarkets spend on advertising.

Convenience stores have lower price elasticities of demand (around −5), but their advertising-to-sales ratios are usually less than those for supermarkets (and are often zero). Why? Because convenience stores mostly serve customers who live nearby; they may need a few items late at night or may simply not want to drive to the supermarket. These customers already know about the convenience store and are unlikely to change their buying habits if the store advertises. Thus E_A is very small, and advertising is not worthwhile.

Advertising is quite important for makers of designer jeans, who will have advertising-to-sales ratios as high as 10 or 20 percent. Advertising helps to make consumers aware of the label and gives it an aura and image. We said that price elasticities of demand in the range of −3 to −4 are typical for the major labels, and advertising elasticities of demand can range from .3 to as high as 1. So, these levels of advertising would seem to make sense.

Laundry detergents have among the highest advertising-to-sales ratios of all products, sometimes exceeding 30 percent, even though demand for any one brand is at least as price elastic as it is for designer jeans. What justifies all the advertising? A very large advertising elasticity. The demand for any one brand of laundry detergent depends crucially on advertising; without it, consumers would have little basis for selecting that particular brand.[23]

[22] Advertising often affects the price elasticity of demand, and this must be taken into account. For some products, advertising broadens the market by attracting a large range of customers, or by creating a bandwagon effect. This is likely to make demand more price elastic than it would have been otherwise (But E_A is likely to be large, so that advertising will still be worthwhile.) Sometimes advertising is used to differentiate a product from others (by creating an image, allure, or brand identification), making the product's demand less price elastic than it would otherwise be

[23] For an overview of statistical approaches to estimating the advertising elasticity of demand, see Ernst R. Berndt, *The Practice of Econometrics* (Reading, MA: Addison-Wesley, 1990), ch. 8.

TABLE 11.7 1993 Sales and Advertising Expenditures for Leading Brands of Over-the-Counter Drugs (in millions of dollars)

	SALES	ADVERTISING	RATIO (%)
Pain Medications			
Tylenol	855	143.8	17
Advil	360	91.7	26
Bayer	170	43.8	26
Excedrin	130	26.7	21
Antacids			
Alka-Seltzer	160	52.2	33
Mylanta	135	32.8	24
Tums	135	27.6	20
Cold Remedies (decongestants)			
Benadryl	130	30.9	24
Sudafed	115	28.6	25
Cough Medicine			
Vicks	350	26.6	8
Robitussin	205	37.7	19
Halls	130	17.4	13

Source: New York Times, September 27, 1994.

Finally, Table 11.7 shows sales, advertising expenditures, and the ratio of the two for leading brands of over-the-counter drugs. Observe that overall, the ratios are quite high. As with laundry detergents, the advertising elasticity for name-brand drugs is very high. Alka-Seltzer, Mylanta, and Tums, for instance, are all antacids that do much the same thing. Sales depend on consumer identification with a particular brand, which requires advertising.

SUMMARY

1. Firms with market power are in an enviable position because they have the potential to earn large profits. Realizing that potential, however, may depend critically on pricing strategy. Even if the firm sets a single price, it needs an estimate of the elasticity of demand for its output. More complicated strategies, which can involve setting several different prices, require even more information about demand.
2. A pricing strategy aims to enlarge the customer base that the firm can sell to and capture as much consumer surplus as possible. There are a number of ways to do this, and they usually involve setting more than a single price.
3. Ideally, the firm would like to price discriminate perfectly—i.e., to charge each customer his or her reservation price. In practice, this is almost always impossible. On the other hand, various forms of imperfect price discrimination are often used to increase profits.
4. The two-part tariff is another means of capturing consumer surplus. Customers must pay an "entry" fee that allows them to buy the good at a per-unit price.

The two-part tariff is most effective when customer demands are relatively homogeneous.

5. When demands are heterogeneous and negatively correlated, bundling can increase profits. With pure bundling, two or more different goods are sold only as a package. With mixed bundling, the customer can buy the goods individually or as a package. Mixed bundling can be more profitable than pure bundling if marginal costs are significant or if demands are not perfectly negatively correlated.
6. Bundling is a special case of tying, a requirement that products be bought or sold in some combination. Tying can be used to meter demand or to protect customer goodwill associated with a brand name.
7. Advertising can further increase profits. The profit-maximizing advertising-to-sales ratio is equal in magnitude to the ratio of the advertising and price elasticities of demand.

QUESTIONS FOR REVIEW

1. Suppose a firm can practice perfect, first-degree price discrimination. What is the lowest price it will charge, and what will its total output be?
2. How does a car salesperson practice price discrimination? How does the ability to discriminate correctly affect his or her earnings?
3. Electric utilities often practice second-degree price discrimination. Why might this improve consumer welfare?
4. Give some examples of third-degree price discrimination. Can third-degree price discrimination be effective if the different groups of consumers have different levels of demand but the same price elasticities?
5. Show why optimal, third-degree price discrimination requires that marginal revenue for each group of consumers equals marginal cost. Use this condition to explain how a firm should change its prices and total output if the demand curve for one group of consumers shifted outward, so that marginal revenue for that group increased.
6. When pricing automobiles, American car companies typically charge a much higher percentage markup over cost for "luxury operation" items (such as leather trim, etc.) than for the car itself or for more "basic" options such as power steering and automatic transmission. Explain why.
7. How is peak-load pricing a form of price discrimination? Can it make consumers better off? Give an example.
8. How can a firm determine an optimal two-part tariff if it has two customers with different demand curves? (Assume that it knows the demand curves.)
9. Why is the pricing of a Gillette safety razor a form of a two-part tariff? Must Gillette be a monopoly producer of its blades as well as its razors? Suppose you were advising Gillette on how to determine the two parts of the tariff. What procedure would you suggest?
10. Why did Loews bundle *Gone with the Wind* and *Getting Gertie's Garter*? What characteristic of demands is needed for bundling to increase profits?
11. How does mixed bundling differ from pure bundling? Under what conditions is mixed bundling preferable to pure bundling? Why do many restaurants practice mixed bundling (by offering a complete dinner as well as an à la carte menu) instead of pure bundling?
12. How does tying differ from bundling? Why might a firm want to practice tying?
13. Why is it incorrect to advertise up to the point that the last dollar of advertising expenditures generates another dollar of sales? What is the correct rule for the marginal advertising dollar?
14. How can a firm check that its advertising-to-sales ratio is not too high or too low? What information does it need?

EXERCISES

1. Price discrimination requires the ability to sort customers and the ability to prevent arbitrage. Explain how the following can function as price discrimination schemes and discuss both sorting and arbitrage:
   a. Requiring airline travelers to spend at least one Saturday night away from home to qualify for a low fare.
   b. Insisting on delivering cement to buyers and basing prices on buyers' locations.
   c. Selling food processors along with coupons that can be sent to the manufacturer to obtain a $10 rebate.
   d. Offering temporary price cuts on bathroom tissue.
   e. Charging high-income patients more than low-income patients for plastic surgery.

2. If the demand for drive-in movies is more elastic for couples than for single individuals, it will be optimal for theaters to charge one admission fee for the driver of the car and an extra fee for passengers. True or false? Explain.
3. In Example 11.1, we saw how producers of processed foods and related consumer goods use coupons as a means of price discrimination. Although coupons are widely used in the United States, that is not the case in other countries. In Germany, coupons are illegal.
   a. Does prohibiting the use of coupons in Germany make German *consumers* better off or worse off?
   b. Does prohibiting the use of coupons make German *producers* better off or worse off?
4. Suppose that BMW can produce any quantity of cars at a constant marginal cost equal to $15,000 and a fixed cost of $20 million. You are asked to advise the CEO as to what prices and quantities BMW should set for sales in Europe and in the United States. The demand for BMWs in each market is given by

$$Q_E = 18{,}000 - 400P_E$$

and

$$Q_U = 5500 - 100P_U$$

where the subscript E denotes Europe, the subscript U denotes the United States, and all prices and costs are in thousands of dollars. Assume that BMW can restrict U.S. sales to authorized BMW dealers only.
   a. What quantity of BMWs should the firm sell in each market, and what will the price be in each market? What will the total profit be?
   b. If BMW were forced to charge the same price in each market, what would be the quantity sold in each market, the equilibrium price, and the company's profit?
5. A monopolist is deciding how to allocate output between two markets. The two markets are separated geographically (East Coast and Midwest). Demand and marginal revenue for the two markets are

$$P_1 = 15 - Q_1 \qquad MR_1 = 15 - 2Q_1$$

$$P_2 = 25 - 2Q_2 \qquad MR_2 = 25 - 4Q_2$$

The monopolist's total cost is $C = 5 + 3(Q_1 + Q_2)$. What are price, output, profits, marginal revenues, and deadweight loss (i) if the monopolist can price discriminate? (ii) if the law prohibits charging different prices in the two regions?
*6. Elizabeth Airlines (EA) flies only one route: Chicago-Honolulu. The demand for each flight on this route is $Q = 500 - P$. EA's cost of running each flight is $30,000 plus $100 per passenger.
   a. What is the profit-maximizing price EA will charge? How many people will be on each flight? What is EA's profit for each flight?
   b. EA learns that the fixed costs per flight are in fact $41,000 instead of $30,000. Will the airline stay in business long? Illustrate your answer using a graph of the demand curve that EA faces, EA's average cost curve when fixed costs are $30,000, and EA's average cost curve when fixed costs are $41,000.
   c. Wait! EA finds out that two different types of people fly to Honolulu. Type A is business people with a demand of $Q_A = 260 - 0.4P$. Type B is students whose total demand is $Q_B = 240 - 0.6P$. The students are easy to spot, so EA decides to charge them different prices. Graph each of these demand curves and their horizontal sum. What price does EA charge the students? What price does it charge other customers? How many of each type are on each flight?
   d. What would EA's profit be for each flight? Would the airline stay in business? Calculate the consumer surplus of each consumer group. What is the total consumer surplus?
   e. Before EA started price discriminating, how much consumer surplus was the Type A demand getting from air travel to Honolulu? Type B? Why did total consumer surplus decline with price discrimination, even though total quantity sold remained unchanged?
7. Many retail video stores offer two alternative plans for renting films:
   - *A two-part tariff:* Pay an annual membership fee (e.g., $40) and then pay a small fee for the daily rental of each film (e.g., $2 per film per day).
   - *A straight rental fee:* Pay no membership fee, but pay a higher daily rental fee (e.g., $4 per film per day).

   What is the logic behind the two-part tariff in this case? Why offer the customer a choice of two plans rather than simply a two-part tariff?
8. Sal's satellite company broadcasts TV to subscribers in Los Angeles and New York. The demand functions for each of these two groups are

$$Q_{NY} = 50 - (1/3)P_{NY}$$

$$Q_{LA} = 80 - (2/3)P_{LA}$$

where Q is in thousands of subscriptions per year and P is the subscription price per year. The cost of providing Q units of service is given by

$$C = 1000 + 30Q$$

where $Q = Q_{NY} + Q_{LA}$.
   a. What are the profit-maximizing prices and quantities for the New York and Los Angeles markets?
   b. As a consequence of a new satellite that the Pentagon recently deployed, people in Los Angeles receive Sal's New York broadcasts and people in New York receive Sal's Los Angeles broadcasts. As a result, anyone in New York or Los Angeles can receive Sal's broadcasts by subscribing in either city. Thus Sal can charge only a single price. What

price should he charge, and what quantities will he sell in New York and Los Angeles?

c. In which of the above situations, (a) or (b), is Sal better off? In terms of consumer surplus, which situation do people in New York prefer and which do people in Los Angeles prefer? Why?

*9. You are an executive for Super Computer, Inc. (SC), which rents out super computers. SC receives a fixed rental payment per time period in exchange for the right to unlimited computing at a rate of P cents per second. SC has two types of potential customers of equal number—10 businesses and 10 academic institutions. Each business customer has the demand function $Q = 10 - P$, where Q is in millions of seconds per month; each academic institution has the demand $Q = 8 - P$. The marginal cost to SC of additional computing is 2 cents per second, regardless of volume.

a. Suppose that you could separate business and academic customers. What rental fee and usage fee would you charge each group? What are your profits?

b. Suppose you were unable to keep the two types of customers separate and charged a zero rental fee. What usage fee maximizes your profits? What are your profits?

c. Suppose you set up one two-part tariff—that is, you set one rental and one usage fee that both business and academic customers pay. What usage and rental fees will you set? What are your profits? Explain why price is not equal to marginal cost.

10. As the owner of the only tennis club in an isolated wealthy community, you must decide on membership dues and fees for court time. There are two types of tennis players. "Serious" players have demand

$$Q_1 = 6 - P$$

where Q_1 is court hours per week and P is the fee per hour for each individual player. There are also "occasional" players with demand

$$Q_2 = 3 - (1/2)P$$

Assume that there are 1000 players of each type. Because you have plenty of courts, the marginal cost of court time is zero. You have fixed costs of $5000 per week. Serious and occasional players look alike, so you must charge them the same prices.

a. Suppose that to maintain a "professional" atmosphere, you want to limit membership to serious players. How should you set the *annual* membership dues and court fees (assume 52 weeks per year) to maximize profits, keeping in mind the constraint that only serious players choose to join? What are profits (per week)?

b. A friend tells you that you could make greater profits by encouraging both types of players to join. Is your friend right? What annual dues and court fees would maximize weekly profits? What would these profits be?

c. Suppose that over the years, young, upwardly mobile professionals move to your community, all of whom are serious players. You believe there are now 3000 serious players and 1000 occasional players. Is it still profitable to cater to the occasional player? What are the profit-maximizing annual dues and court fees? What are profits per week?

11. Look again at Figure 11.12 (p. 394), which shows the reservation prices of three consumers for two goods. Assuming that marginal production cost is zero for both goods, can the producer make the most money by selling the goods separately, by using pure bundling, or by using mixed bundling? What prices should be charged?

12. Look again at Figure 11.17 (p. 397). Suppose the marginal costs c_1 and c_2 were zero. Show that in this case, pure bundling, not mixed bundling, is the most profitable pricing strategy. What price should be charged for the bundle? What will the firm's profit be?

13. Some years ago, an article appeared in the *New York Times* about IBM's pricing policy. The previous day, IBM had announced major price cuts on most of its small and medium-sized computers. The article said:

> IBM probably has no choice but to cut prices periodically to get its customers to purchase more and lease less. If they succeed, this could make life more difficult for IBM's major competitors. Outright purchases of computers are needed for ever larger IBM revenues and profits, says Morgan Stanley's Ulric Weil in his new book, *Information Systems in the '80's*. Mr. Weil declares that IBM cannot revert to an emphasis on leasing.

a. Provide a brief but clear argument in *support* of the claim that IBM should try "to get its customers to purchase more and lease less."

b. Provide a brief but clear argument *against* this claim.

c. What factors determine whether leasing or selling is preferable for a company like IBM? Explain briefly.

14. You are selling two goods, 1 and 2, to a market consisting of three consumers with reservation prices as follows:

	RESERVATION PRICE ($)	
Consumer	For 1	For 2
A	10	70
B	40	40
C	70	10

The unit cost of each product is $20.

a. Compute the optimal prices and profits for (i) selling the goods separately, (ii) pure bundling, and (iii) mixed bundling.

b. Which strategy is most profitable? Why?

15. Your firm produces two products, the demands for which are independent. Both products are produced at zero marginal cost. You face four consumers (or groups of consumers) with the following reservation prices:

Consumer	Good 1 ($)	Good 2 ($)
A	30	90
B	40	60
C	60	40
D	90	30

a. Consider three alternative pricing strategies: (i) selling the goods separately; (ii) pure bundling; (iii) mixed bundling. For *each strategy*, determine the optimal prices to be charged and the resulting profits. Which strategy is best?

b. Now suppose that the production of each good entails a marginal cost of $35. How does this information change your answers to (a)? Why is the optimal strategy now different?

16. A cable TV company offers, in addition to its basic service, two products: a Sports Channel (Product 1) and a Movie Channel (Product 2). Subscribers to the basic service can subscribe to these additional services individually at the monthly prices P_1 and P_2, respectively, or they can buy the two as a bundle for the price P_B, where $P_B < P_1 + P_2$. They can also forgo the additional services and simply buy the basic service. The company's marginal cost for these additional services is *zero*. Through market research, the cable company has estimated the reservation prices for these two services for a representative group of consumers in the company's service area. These reservation prices are plotted (as x's) in Figure 11.21, as are the prices P_1, P_2, and P_B that the cable company is currently charging. The graph is divided into regions I, II, III, and IV.

a. Which products, if any, will be purchased by the consumers in region I? In region II? In region III? In region IV? Explain briefly.

b. Note that the reservation prices for the Sports Channel and the Movie Channel, as drawn in the figure, are negatively correlated. Why would you, or why would you not, expect consumers' reservation prices for cable TV channels to be negatively correlated?

c. The company's vice president has said: "Because the marginal cost of providing an additional channel is zero, mixed bundling offers no advantage over pure bundling. Our profits would be just as high if we offered the Sports Channel and the Movie Channel together as a bundle, and only as a bundle." Do you agree or disagree? Explain why.

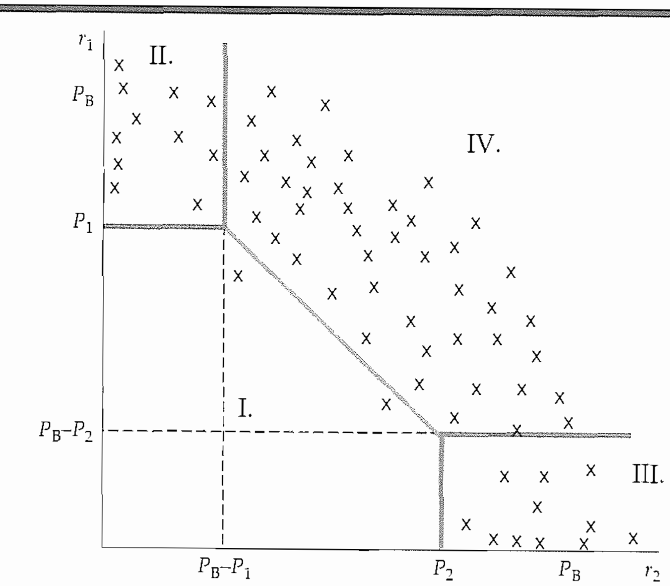

FIGURE 11.21 Figure for Exercise 16

d. Suppose the cable company continues to use mixed bundling to sell these two services. Based on the distribution of reservation prices shown in Figure 11.21, do you think the cable company should alter any of the prices it is now charging? If so, how?

17. Consider a firm with monopoly power that faces the demand curve

$$P = 100 - 3Q + 4A^{1/2}$$

and has the total cost function

$$C = 4Q^2 + 10Q + A$$

where A is the level of advertising expenditures, and P and Q are price and output.

a. Find the values of A, Q, and P that maximize the firm's profit.

b. Calculate the Lerner index of monopoly power, $L = (P - MC)/P$, for this firm at its profit-maximizing levels of A, Q, and P.

APPENDIX TO CHAPTER 11

Transfer Pricing in the Integrated Firm

So far we have studied the firm's pricing decision assuming that it sells its output in an *outside market*, i.e., to consumers or to other firms. Many firms, however, are *vertically integrated*—they contain several divisions, with some divisions producing parts and components that other divisions use to produce the finished product.[1] For example, each of the major U.S. automobile companies has "upstream" divisions that produce engines, brakes, radiators, and other components that the "downstream" divisions use to produce the finished products. *Transfer pricing* refers to the valuation of these parts and components within the firm. **Transfer prices** are internal prices at which the parts and components from upstream divisions are "sold" to downstream divisions. Transfer prices must be chosen correctly because they are the signals that divisional managers use to determine output levels.

transfer prices Internal prices at which parts and components from upstream divisions are "sold" to downstream divisions within a firm.

This appendix shows how a profit-maximizing firm chooses its transfer prices and divisional output levels. We will also examine other issues raised by vertical integration. For example, suppose a computer firm's upstream division produces memory chips used by a downstream division to produce the final product. If other firms also produce these chips, should our firm obtain all its chips from the upstream division, or should it also buy some on the outside market? Should the upstream division produce more chips than the downstream division needs and sell the excess in the market? How should the firm coordinate its upstream and downstream divisions? In particular, can we design incentives for the divisions that help the firm to maximize its profits?

We begin with the simplest case: There is no outside market for the output of the upstream division—i.e., the upstream division produces a good that is neither produced nor used by any other firm. Next we consider what happens when there is an outside market for the upstream division's output.

Transfer Pricing When There Is No Outside Market

Consider a firm with three divisions: Two upstream divisions produce inputs to a downstream processing division. The two upstream divisions produce quantities Q_1 and Q_2 and have total costs $C_1(Q_1)$ and $C_2(Q_2)$. The downstream division produces a quantity Q using the production function

$$Q = f(K, L, Q_1, Q_2)$$

where K and L are capital and labor inputs, and Q_1 and Q_2 are the intermediate inputs from the upstream divisions. Excluding the costs of the inputs Q_1 and Q_2, the downstream division has a total production cost $C_d(Q)$. Total revenue from sales of the final product is $R(Q)$.

[1] A firm is *horizontally integrated* when it has several divisions that produce the same or closely related products. Many firms are both vertically and horizontally integrated.

We assume there are *no outside markets* for the intermediate inputs Q_1 and Q_2. (They can be used only by the downstream division.) Then the firm has two problems:

1. What quantities Q_1, Q_2, and Q maximize its profit?
2. Is there an incentive scheme that will decentralize the firm's management? In particular, is there a set of transfer prices P_1 and P_2, so that *if each division maximizes its own divisional profit, the profit of the overall firm will also be maximized*?

To solve these problems, note that the firm's total profit is

$$\pi(Q) = R(Q) - C_d(Q) - C_1(Q_1) - C_2(Q_2) \quad \text{(A11.1)}$$

In §10.1, we explain that a firm maximizes its profit at the output at which marginal revenue is equal to marginal cost.

Now, what is the level of Q_1 that maximizes this profit? It is the level at which *the cost of the last unit of Q_1 is just equal to the additional revenue it brings to the firm*. The cost of producing one extra unit of Q_1 is the marginal cost $\Delta C_1/\Delta Q_1 = MC_1$. How much extra revenue results from that one extra unit? An extra unit of Q_1 allows the firm to produce more final output Q of an amount $\Delta Q/\Delta Q_1 = MP_1$, the marginal product of Q_1. An extra unit of final output results in additional revenue $\Delta R/\Delta Q = MR$, but it also results in additional cost to the downstream division, of an amount $\Delta C_d/\Delta Q = MC_d$. Thus the *net marginal revenue* NMR_1 that the firm earns from an extra unit of Q_1 is $(MR - MC_d)MP_1$. Setting this equal to the marginal cost of the unit, we obtain the following rule for profit maximization:[2]

$$NMR_1 = (MR - MC_d)MP_1 = MC_1 \quad \text{(A11.2)}$$

Going through the same steps for the second intermediate input gives

$$NMR_2 = (MR - MC_d)MP_2 = MC_2 \quad \text{(A11.3)}$$

Note from equations (A11.2) and (A11.3) that it is *incorrect* to determine the firm's final output level Q by setting marginal revenue equal to marginal cost for the downstream division—i.e., by setting $MR = MC_d$. Doing so ignores the cost of producing the intermediate input. (MR exceeds MC_d because this cost is positive.) Also, note that equations (A11.2) and (A11.3) are standard conditions of marginal analysis: The output of each upstream division should be such that its marginal cost is equal to its marginal contribution to the profit of the overall firm.

Now, what transfer prices P_1 and P_2 should be "charged" to the downstream division for its use of the intermediate inputs? Remember that if each of the three divisions uses these transfer prices to maximize its own divisional profit, the profit of the overall firm should be maximized. The two upstream divisions will maximize their divisional profits, π_1 and π_2, which are given by

$$\pi_1 = P_1Q_1 - C_1(Q_1)$$

[2] Using calculus, we can obtain this rule by differentiating equation (A11.1) with respect to Q_1:

$$d\pi/dQ_1 = (dR/dQ)(\partial Q/\partial Q_1) - (dC_d/dQ)(\partial Q/\partial Q_1) - dC_1/dQ_1$$
$$= (MR - MC_d)MP_1 - MC_1$$

Setting $d\pi/dQ = 0$ to maximize profit gives equation (A11.2).

and

$$\pi_2 = P_2Q_2 - C_2(Q_2)$$

Since the upstream divisions take P_1 and P_2 as given, they will choose Q_1 and Q_2 so that $P_1 = MC_1$ and $P_2 = MC_2$. Similarly, the downstream division will maximize

$$\pi(Q) = R(Q) - C_d(Q) - P_1Q_1 - P_2Q_2$$

Since the downstream division also takes P_1 and P_2 as given, it will choose Q_1 and Q_2 so that

$$(MR - MC_d)MP_1 = NMR_1 = P_1 \quad \text{(A11.4)}$$

and

$$(MR - MC_d)MP_2 = NMR_2 = P_2 \quad \text{(A11.5)}$$

Note that by setting the transfer prices equal to the respective marginal costs ($P_1 = MC_1$ and $P_2 = MC_2$), the profit-maximizing conditions given by equations (A11.2) and (A11.3) will be satisfied. We therefore have a simple solution to the transfer pricing problem: *Set each transfer price equal to the marginal cost of the respective upstream division*. Then when each division is required to maximize its own profit, the quantities Q_1 and Q_2 that the upstream divisions will want to produce will be the same quantities that the downstream division will want to "buy," and they will maximize the firm's total profit.

We can illustrate this graphically with the following example. Race Car Motors, Inc., has two divisions. The upstream Engine Division produces engines, and the downstream Assembly Division puts together automobiles, using one engine (and a few other parts) in each car. In Figure A11.1, the average revenue curve AR is Race Car Motors' demand curve for cars. (Note that the firm has monopoly power in the automobile market.) MC_A is the marginal cost of assembling automobiles, *given the engines* (i.e., it does not include the cost of the engines). Since the car requires one engine, the marginal product of the engines is one. Therefore, the curve labeled $MR - MC_A$ is also the net marginal revenue curve for engines:

$$NMR_E = (MR - MC_A)MP_E = MR - MC_A$$

The profit-maximizing number of engines (and number of cars) is given by the intersection of the net marginal revenue curve NMR_E with the marginal cost curve for engines MC_E. Having determined the number of cars it will produce, and knowing its divisional cost functions, the management of Race Car Motors can now set the transfer price P_E that correctly values the engines used to produce its cars. This is the transfer price that should be used to calculate divisional profit (and year-end bonuses for divisional managers).

Transfer Pricing with a Competitive Outside Market

Now suppose there is a *competitive* outside market for the intermediate good produced by an upstream division. Since the outside market is competitive, there is a single market price at which one can buy or sell the good. Therefore, *the marginal*

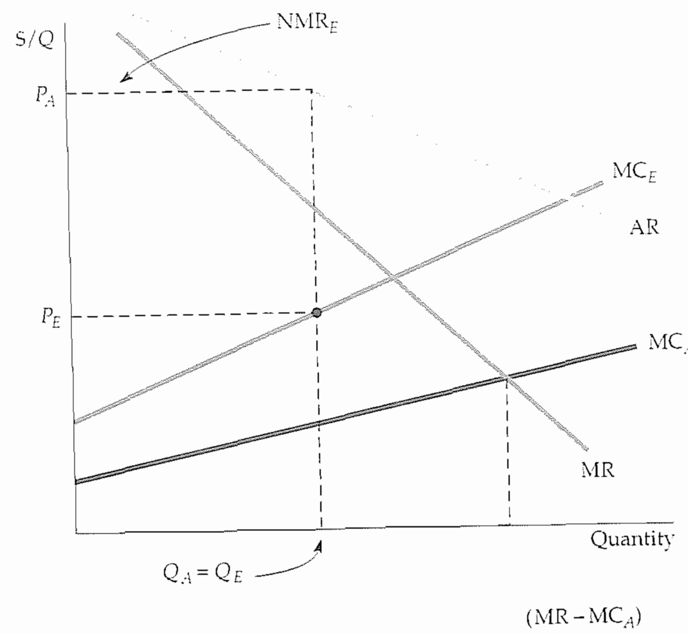

FIGURE A11.1 Race Car Motors, Inc.

The firm's upstream division should produce a quantity of engines Q_E that equates its marginal cost of engine production MC_E with the downstream division's net marginal revenue of engines NMR_E. Since the firm uses one engine in every car, NMR_E is the difference between the marginal revenue from selling cars and the marginal cost of assembling them, i.e., $MR - MC_A$. The optimal transfer price for engines P_E equals the marginal cost of producing them. Finished cars are sold at price P_A.

cost of the intermediate good is simply the market price. Because the optimal transfer price must equal marginal cost, it must also equal the competitive market price.

To see this, suppose there is a competitive market for the engines that Race Car Motors produces. If the market price is low, Race Car Motors may want to buy some or all of its engines in the market; if it is high, it may want to sell engines in the market. Figure A11.2 illustrates the first case. For quantities below $Q_{E,1}$, the upstream division's marginal cost of producing engines MC_E is below the market price $P_{E,M}$; for quantities above $Q_{E,1}$, it is above the market price. The firm should obtain engines at the least cost, so the marginal cost of engines MC_E^* will be the upstream division's marginal cost for quantities up to $Q_{E,1}$ and the market price for quantities above $Q_{E,1}$. Note that Race Car Motors uses more engines and produces more cars than it would have had there been no outside engine market. The downstream division now buys $Q_{E,2}$ engines and produces an equal number of automobiles. However, it "buys" only $Q_{E,1}$ of these engines from the upstream division and the rest on the open market.

It might seem strange that Race Car Motors must go into the open market to buy engines that it can make itself. If it made all of its own engines, however, its marginal cost of producing them would exceed the competitive market price. Although the profit of the upstream division would be higher, *the total profit of the firm would be lower.*

Figure A11.3 shows the case where Race Car Motors *sells* engines in the outside market. Now the competitive market price $P_{E,M}$ is above the transfer price

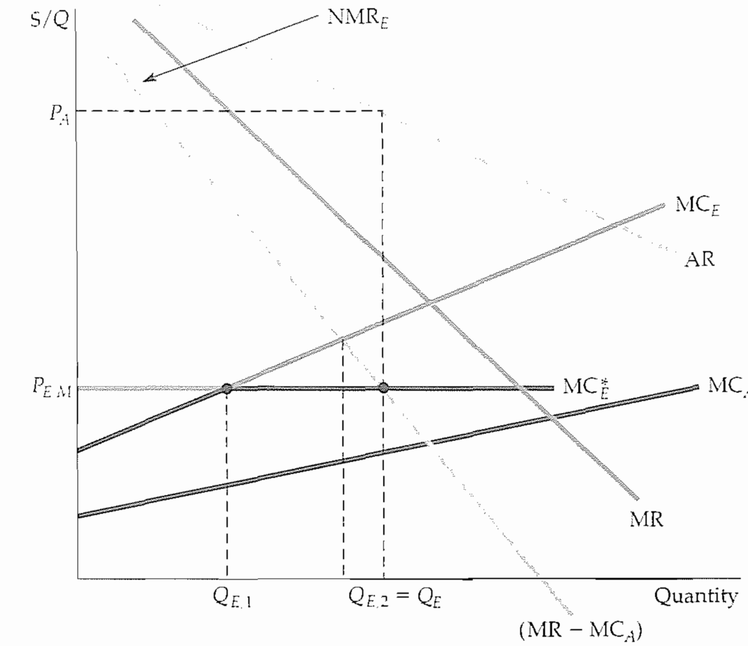

FIGURE A11.2 Buying Engines in a Competitive Outside Market

Race Car Motors' marginal cost of engines MC_E^* is the upstream division's marginal cost for quantities up to $Q_{E,1}$ and the market price $P_{E,M}$ for quantities above $Q_{E,1}$. The downstream division should use a total of $Q_{E,2}$ engines to produce an equal number of cars; then the marginal cost of engines equals net marginal revenue. $Q_{E,2} - Q_{E,1}$ of these engines are bought in the outside market. The upstream division "pays" the downstream division the transfer price $P_{E,M}$ for the remaining $Q_{E,1}$ engines.

that the firm would have set had there been no outside market. In this case, although the upstream Engine Division produces $Q_{E,1}$ engines, only $Q_{E,2}$ engines are used by the downstream division to produce automobiles. The rest are sold in the outside market at the price $P_{E,M}$.

Note that compared with a situation in which there is no outside engine market, Race Car Motors is producing more engines but fewer cars. Why not produce this larger number of engines but use all of them to produce more cars? Because the engines are too valuable. On the margin, the net revenue that can be earned from selling them in the outside market is higher than the net revenue from using them to build additional cars.

Transfer Pricing with a Noncompetitive Outside Market

Now suppose there is an outside market for the output of the upstream division, but that market is not competitive—the firm has monopoly power. The same principles apply, but we must be careful when measuring net marginal revenue.

Suppose the engine produced by the upstream Engine Division is a special one that only Race Car Motors can make. There is, however, an outside market for this engine. Race Car Motors, therefore, can be a monopoly supplier to that market while also producing engines for its own use. What is the optimal transfer

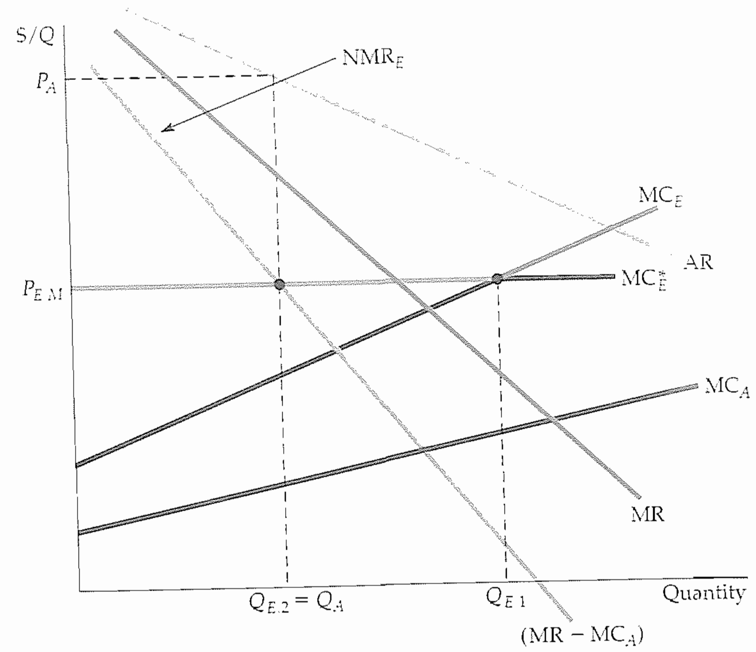

FIGURE A11.3 Selling Engines in a Competitive Outside Market

The optimal transfer price for Race Car Motors is again the market price $P_{E,M}$. This price is above the point at which MC_E intersects NMR_E, so the upstream division sells some of its engines in the outside market. The upstream division produces $Q_{E,1}$ engines, the quantity at which MC_E equals $P_{E,M}$. The downstream division uses only $Q_{E,2}$ of these engines, the quantity at which NMR_E equals $P_{E,M}$. Compared with Figure A11.1, in which there is no outside market, more engines but fewer cars are produced.

price for use of the engines by the downstream division, and at what price (if any) should engines be sold in the outside market?

We must find the firm's net marginal revenue from the sale of engines. In Figure A11.4, $D_{E,M}$ is the outside market demand curve for engines and $MR_{E,M}$ is the corresponding marginal revenue curve. Race Car Motors thus has two sources of marginal revenue from the production and sale of an additional engine: marginal revenue $MR_{E,M}$ from sales in the outside market and net marginal revenue $(MR - MC_A)$ from the use of the engines by the downstream division. By summing these two curves horizontally, we obtain the *total net marginal revenue curve for engines*; it is the green line labeled NMR_E.

The intersection of the marginal cost and total net marginal revenue curves gives the quantity of engines $Q_{E,1}$ that the upstream division should produce and the optimal transfer price P_E^*. Again, the optimal transfer price is equal to marginal cost. But note that only $Q_{E,2}$ of these engines are used by the downstream division to make cars. This is the quantity at which the downstream division's net marginal revenue, $MR - MC_A$, is equal to the transfer price P_E^*. The remaining engines $Q_{E,3}$ are sold in the outside market. However, they are not sold at the transfer price P_E^*. Instead the firm exercises its monopoly power and sells them at the higher price $P_{E,M}$.

Why pay the upstream division only P_E^* per engine when the firm is selling engines in the outside market at the higher price $P_{E,M}$? Because if the upstream

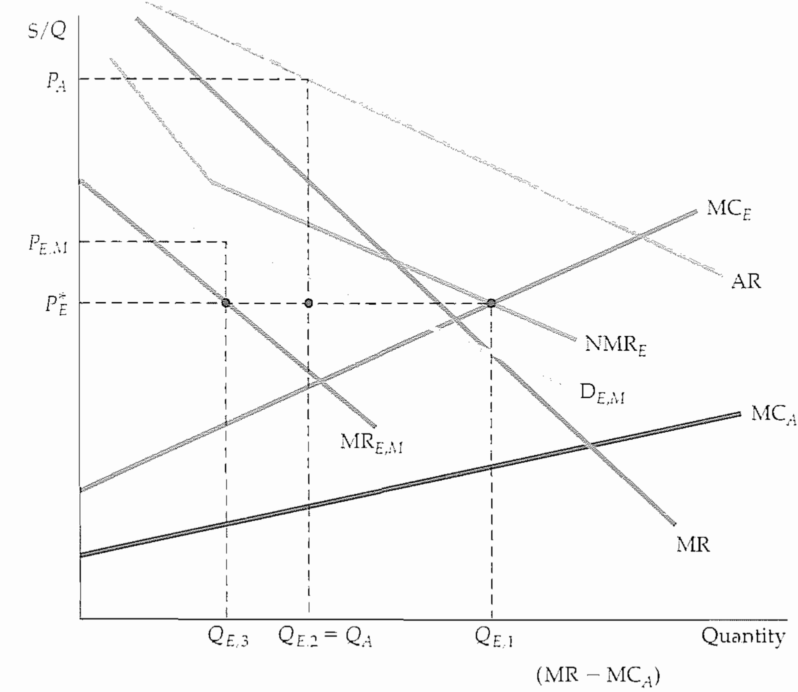

FIGURE A11.4 Race Car Motors—A Monopoly Supplier of Engines to an Outside Market

$D_{E,M}$ is the outside market demand curve for engines; $MR_{E,M}$ is the corresponding marginal revenue curve; $(MR - MC_A)$ is the net marginal revenue from the use of engines by the downstream division. The *total net marginal revenue curve for engines* NMR_E is the horizontal sum of these two marginal revenues. The optimal transfer price P_E^* and the quantity of engines that the upstream division produces, $Q_{E,1}$, are found where $MC_E = NMR_E$. $Q_{E,2}$ of these engines are used by the downstream division, the quantity at which the downstream division's net marginal revenue, $MR - MC_A$, is equal to the transfer price P_E^*. The remaining engines, $Q_{E,3}$, are sold in the outside market at the price $P_{E,M}$.

division is paid more than P_E^* (and thereby encouraged to produce more engines), the marginal cost of engines will rise and exceed the net marginal revenue from their use by the downstream division. And if the price charged in the outside market were lowered, the marginal revenue from sales in that market would fall below marginal cost. At the prices P_E^* and $P_{E,M}$, marginal revenues and marginal cost are equal:

$$MR_{E,M} = (MR - MC_A) = MC_E$$

Sometimes a vertically integrated firm can buy components in an outside market in which it has *monopsony* power. Suppose, for example, that Race Car Motors can obtain engines from its upstream Engine Division, or can purchase them *as a monopsonist* in an outside market. Although we have not illustrated this case graphically, you should be able to see that in this case the transfer price paid to the Engine Division will be *above* the price at which engines are bought in the outside market. Why "pay" the upstream division a price that is higher

In §10.5, we explain that when a buyer has monopsony power, its marginal expenditure curve lies above its average expenditure curve because the decision to buy an extra unit of the good raises the price that must be paid on all units.

than that paid in the outside market? With monopsony power, purchasing one additional engine in the outside market incurs a *marginal expenditure* that is greater than the actual price per engine paid in that market. The marginal expenditure is higher because purchasing an additional unit raises the average expenditure paid for *all* units bought in the outside market.

A Numerical Example

Suppose Race Car Motors has the following demand for its automobiles:

$$P = 20{,}000 - Q$$

Its marginal revenue is thus

$$MR = 20{,}000 - 2Q$$

The downstream division's cost of assembling cars is

$$C_A(Q) = 8000Q$$

so that the division's marginal cost is $MC_A = 8000$. The upstream division's cost of producing engines is

$$C_E(Q_E) = 2Q_E^2$$

The division's marginal cost is thus $MC_E(Q_E) = 4Q_E$.

First, suppose there is *no outside market* for the engines. How many engines and cars should the firm produce? What should be the transfer price for engines? To solve this problem, we set the net marginal revenue for engines equal to the marginal cost of producing engines. Since each car has one engine, $Q_E = Q$. The net marginal revenue of engines is thus

$$NMR_E = MR - MC_A = 12{,}000 - 2Q_E$$

Now set NMR_E equal to MC_E:

$$12{,}000 - 2Q_E = 4Q_E$$

Thus $6Q_E = 12{,}000$ and $Q_E = 2000$. The firm should therefore produce 2000 engines and 2000 cars. The optimal transfer price is the marginal cost of these 2000 engines:

$$P_E = 4Q_E = \$8000$$

Second, suppose that engines can be bought or sold for \$6000 in an *outside competitive market*. This is below the \$8000 transfer price that is optimal when there is no outside market, so the firm should buy some engines outside. Its marginal cost of engines, and the optimal transfer price, is now \$6000. Set this \$6000 marginal cost equal to the net marginal revenue of engines:

$$6000 = NMR_E = 12{,}000 - 2Q_E$$

Thus the total quantity of engines and cars is now 3000. The company now produces more cars (and sells them at a lower price) because its cost of engines is lower. Also, since the transfer price for the engines is now \$6000, the upstream Engine Division supplies only 1500 engines (because $MC_E(1500) = \$6000$). The remaining 1500 engines are bought in the outside market.

Finally, suppose Race Car Motors is the only producer of these engines but can sell them in an outside market. Demand in the outside market is

$$P_{E,M} = 10{,}000 - Q_E$$

The marginal revenue from sales in the market is therefore

$$MR_{E,M} = 10{,}000 - 2Q_E$$

To determine the optimal transfer price, we find the *total* net marginal revenue by horizontally summing $MR_{E,M}$ with the net marginal revenue from "sales" to the downstream division, $12{,}000 - 2Q_E$, as in Figure A11.4. For outputs Q_E greater than 1000, this is

$$NMR_{E,Total} = 11{,}000 - Q_E$$

Now set this equal to the marginal cost of producing engines:

$$11{,}000 - Q_E = 4Q_E$$

Therefore the total quantity of engines produced should be $Q_E = 2200$.

How many of these engines should go to the downstream division and how many to the outside market? Note that the marginal cost of producing these 2200 engines—and therefore the optimal transfer price—is $4Q_E = \$8800$. Set this price equal to the marginal revenue from sales in the outside market:

$$8800 = 10{,}000 - 2Q_E$$

or $Q_E = 600$. Therefore, 600 engines should be sold in the outside market. Finally, set this \$8800 transfer price equal to the net marginal revenue from "sales" to the downstream division:

$$8800 = 12{,}000 - 2Q_E$$

or $Q_E = 1600$. Thus 1600 engines should be supplied to the downstream division for use in the production of 1600 cars.

EXERCISES

1. Review the numerical example about Race Car Motors. Calculate the profit earned by the upstream division, the downstream division, and the firm as a whole in each of the three cases examined: (a) there is no outside market for engines; (b) there is a competitive market for engines in which the market price is \$6000; and (c) the firm is a monopoly supplier of engines to an outside market. In which case does Race Car Motors earn the most profit? In which case does the upstream division earn the most? The downstream division?

2. Ajax Computer makes a computer for climate control in office buildings. The company uses a microprocessor produced by its upstream division, along with other parts bought in outside competitive markets. The microprocessor is produced at a constant marginal cost of \$500, and the marginal cost of assembling the computer (including the cost of the other parts) by the downstream division is a constant \$700. The firm has been selling the computer for \$2000, and until now there has been no outside market for the microprocessor.

a. Suppose an outside market for the microprocessor develops and Ajax has monopoly power in that market, selling microprocessors for $1000 each. Assuming that demand for the microprocessor is unrelated to the demand for the Ajax computer, what transfer price should Ajax apply to the microprocessor for its use by the downstream computer division? Should production of computers be increased, decreased, or left unchanged? Explain briefly.

b. How would your answer to (a) change if the demands for the computer and the microprocessors were competitive; i.e., if some of the people who buy the microprocessors use them to make climate control systems of their own?

3. Reebok produces and sells running shoes. It faces a market demand schedule $P = 11 - 1.5Q_s$, where Q_s is the number of pairs of shoes sold (in thousands) and P is the price in dollars per thousand pair of shoes. Production of each pair of shoes requires 1 square yard of leather. The leather is shaped and cut by the Form Division of Reebok. The cost function for leather is

$$TC_L = 1 + Q_L + 0.5Q_L^2$$

where Q_L is the quantity of leather (in thousands of square yards) produced. Excluding leather, the cost function for running shoes is

$$TC_s = 2Q_s$$

a. What is the optimal transfer price?

b. Leather can be bought and sold in a competitive market at the price of $P_F = 1.5$. In this case, how much leather should the Form Division supply internally? How much should it supply to the outside market? Will Reebok buy any leather in the outside market? Find the optimal transfer price.

c. Now suppose the leather is unique and of extremely high quality. Therefore, the Form Division may act as a monopoly supplier to the outside market as well as a supplier to the downstream division. Suppose the outside demand for leather is given by $P = 32 - Q_L$. What is the optimal transfer price for the use of leather by the downstream division? At what price, if any, should leather be sold to the outside market? What quantity, if any, will be sold to the outside market?

4. The House Products Division of Acme Corporation manufactures and sells digital clock radios. A major component is supplied by the electronics division of Acme. The cost functions for the radio and the electronic component divisions are, respectively,

$$TC_r = 30 + 2Q_r$$

$$TC_c = 70 + 6Q_c + Q_c^2$$

Note that TC_r does not include the cost of the component. Manufacture of one radio set requires the use of one electronic component. Market studies show that the firm's demand curve for the digital clock radio is given by

$$P_r = 108 - Q_r$$

a. Assuming no outside market for the components, how many of them should be produced to maximize profits for Acme as a whole? What is the optimal transfer price?

b. If other firms are willing to purchase in the outside market the component manufactured by the electronics division (which is the only supplier of this product), what is the optimal transfer price? Why? What price should be charged in the outside market? Why? How many units will the electronics division supply internally and to the outside market? Why? (*Note*: The demand for components in the outside market is $P_c = 72 - 1.5Q_c$.)

CHAPTER 12

Monopolistic Competition and Oligopoly

Chapter Outline

List of Examples

In the last two chapters, we saw how firms with monopoly power can choose prices and output levels to maximize profit. We also saw that monopoly power does not require a firm to be a pure monopolist. In many industries, even though several firms compete each has at least some monopoly power: It has control over price and will charge a price that exceeds marginal cost.

In this chapter, we examine market structures other than pure monopoly that can give rise to monopoly power. We begin with **monopolistic competition.** A monopolistically competitive market is similar to a perfectly competitive market in two key respects: There are many firms and entry by new firms is not restricted. But it differs from perfect competition in that the product is *differentiated*: Each firm sells a brand or version of the product that differs in quality, appearance, or reputation, and each firm is the sole producer of its own brand. The amount of monopoly power the firm has depends on its success in differentiating its product from those of other firms. Examples of monopolistically competitive industries abound: toothpaste, laundry detergent, and packaged coffee are a few.

The second form of market structure we will examine is **oligopoly:** a market in which only a few firms compete with one another, and entry by new firms is impeded. The product that the firms produce might be differentiated, as with automobiles, or it might not be, as with steel. Monopoly power and profitability in oligopolistic industries depend in part on how the firms interact. For example, if the interaction is more cooperative than competitive, the firms could charge prices well above marginal cost and earn large profits.

In some oligopolistic industries, firms do cooperate, but in others firms compete aggressively, even though this means lower profits. To see why, we need to consider how oligopolistic firms decide on output and prices. These decisions are complicated because each firm must operate *strategically*—when making a decision, it must weigh the probable reactions of its competitors. To understand oligopolistic markets, we must therefore introduce some basic concepts of gaming and strategy. We develop these concepts more fully in Chapter 13.

monopolistic competition Market in which firms can enter freely, each producing its own brand or version of a differentiated product.

oligopoly Market in which only a few firms compete with one another, and entry by new firms is impeded.

cartel Market in which some or all firms explicitly collude, coordinating prices and output levels to maximize joint profits.

The third form of market structure we examine is a **cartel**. In a cartelized market, some or all firms explicitly *collude:* they coordinate their prices and output levels to maximize *joint* profits. Cartels can arise in markets that would otherwise be competitive, as with OPEC, or oligopolistic, as with the international bauxite cartel.

At first glance, a cartel may seem like a pure monopoly. After all, the firms in a cartel appear to operate as though they were parts of one big company. But a cartel differs from a monopoly in two important respects. First, since cartels rarely control the entire market, they must consider how their pricing decisions will affect noncartel production levels. Second, because the members of a cartel are *not* part of one big company, they may be tempted to "cheat" their partners by undercutting prices and grabbing bigger shares of the market. As a result, many cartels tend to be unstable and short-lived.

12.1 Monopolistic Competition

In many industries, the products are differentiated. For one reason or another, consumers view each firm's brand as different from other brands. Crest toothpaste, for example, is perceived to be different from Colgate, Aim, and a dozen other toothpastes. The difference is partly flavor, partly consistency, and partly reputation—the consumer's image (correct or incorrect) of the relative decay-preventing efficacy of Crest. As a result, some consumers (but not all) will pay more for Crest.

In §10.2, we explain that a seller of a product has some monopoly power if it can profitably charge a price greater than marginal cost.

Because Procter & Gamble is the sole producer of Crest, it has monopoly power. But its monopoly power is limited because consumers can easily substitute other brands if the price of Crest rises. Although consumers who prefer Crest will pay more for it, most of them will not pay much more. The typical Crest user might pay 25 or even 50 cents a tube more, but probably not a dollar more. For most consumers, toothpaste is toothpaste, and the differences among brands are small. Therefore, the demand curve for Crest toothpaste, though downward sloping, is fairly elastic. (A reasonable estimate of the elasticity of demand for Crest is −7.) Because of its limited monopoly power, Procter & Gamble will charge a price that is higher, but not much higher, than marginal cost. The situation is similar for Tide detergent or Scott paper towels.

The Makings of Monopolistic Competition

A monopolistically competitive market has two key characteristics:

1. Firms compete by selling differentiated products that are highly substitutable for one another but not perfect substitutes. (In other words, the cross-price elasticities of demand are large but not infinite.)
2. There is *free entry and exit:* it is relatively easy for new firms to enter the market with their own brands and for existing firms to leave if their products become unprofitable.

To see why free entry is an important requirement, let's compare the markets for toothpaste and automobiles. The toothpaste market is monopolistically competitive, but the automobile market is better characterized as an oligopoly. It is relatively easy for other firms to introduce new brands of toothpaste, and this limits the profitability of producing Crest or Colgate. If the profits were large, other firms would spend the necessary money (for development, production, advertising, and promotion) to introduce new brands of their own, which would reduce the market shares and profitability of Crest and Colgate.

The automobile market is also characterized by product differentiation. However, the large scale economies involved in production make entry by new firms difficult. Thus, until the mid-1970s, when Japanese producers became important competitors, the three major U.S. automakers had the market largely to themselves.

There are many other examples of monopolistic competition besides toothpaste. Soap, shampoo, deodorants, shaving cream, cold remedies, and many other items found in a drugstore are sold in monopolistically competitive markets. The markets for bicycles and other sporting goods are likewise monopolistically competitive. So is most retail trade, because goods are sold in many different stores that compete with one another by differentiating their services according to location, availability and expertise of salespeople, credit terms, etc. Entry is relatively easy, so if profits are high in a neighborhood because there are only a few stores, new stores will enter.

Equilibrium in the Short Run and the Long Run

As with monopoly, in monopolistic competition firms face downward-sloping demand curves. Therefore, they have monopoly power. But this does not mean that monopolistically competitive firms are likely to earn large profits. Monopolistic competition is also similar to perfect competition: Because there is free entry, the potential to earn profits will attract new firms with competing brands, driving economic profits down to zero.

In §10.1, we explain that a monopolist maximizes profit by choosing an output at which marginal revenue is equal to marginal cost.

To make this clear, let's examine the equilibrium price and output level for a monopolistically competitive firm in the short and long run. Figure 12.1(a) shows the short-run equilibrium. Because the firm's product differs from its competitors', its demand curve D_{SR} is downward sloping. (This is the *firm's* demand curve, not the market demand curve, which is more steeply sloped.) The profit-maximizing quantity Q_{SR} is found at the intersection of the marginal revenue and marginal cost curves. Because the corresponding price P_{SR} exceeds average cost, the firm earns a profit, as shown by the shaded rectangle in the figure.

Recall from §8.6 that with the possibility of entry and exit, firms will earn zero economic profit in long-run equilibrium.

In the long run, this profit will induce entry by other firms. As they introduce competing brands, our firm will lose market share and sales; its demand curve will shift down, as in Figure 12.1(b). (In the long run, the average and marginal cost curves may also shift. We have assumed for simplicity that costs do not change.) The long-run demand curve D_{LR} will be just tangent to the firm's average cost curve. Here profit maximization implies the quantity Q_{LR} and the price P_{LR}. It also implies *zero profit* because price is equal to average cost. Our firm still has monopoly power: Its long-run demand curve is downward sloping because its particular brand is still unique. But the entry and competition of other firms have driven its profit to zero.

More generally, firms may have different costs, and some brands will be more distinctive than others. In this case, firms may charge slightly different prices, and some will earn a small profit.

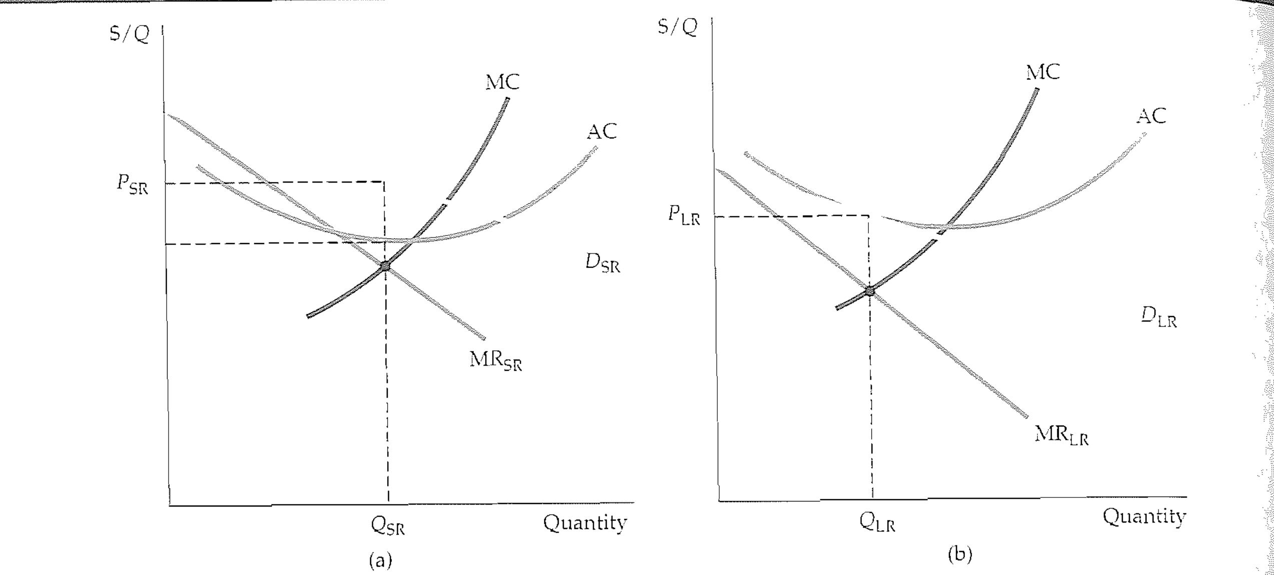

FIGURE 12.1 A Monopolistically Competitive Firm in the Short and Long Run

Because the firm is the only producer of its brand, it faces a downward-sloping demand curve: Price exceeds marginal cost and the firm has monopoly power. In the short run, described in part **(a)**, price also exceeds average cost, and the firm earns profits shown by the yellow-shaded rectangle. In the long run, these profits attract new firms with competing brands. The firm's market share falls, and its demand curve shifts downward. In long-run equilibrium, described in part **(b)**, price equals average cost, so the firm earns zero profit, even though it has monopoly power.

Monopolistic Competition and Economic Efficiency

In §9.2, we explain that competitive markets are efficient because they maximize the sum of consumers' and producers' surplus.

Perfectly competitive markets are desirable because they are economically efficient: As long as there are no externalities and nothing impedes the workings of the market, the total surplus of consumers and producers is as large as possible. Monopolistic competition is similar to competition in some respects, but is it an efficient market structure? To answer this question, let's compare the long-run equilibrium of a monopolistically competitive industry to the long-run equilibrium of a perfectly competitive industry.

Figure 12.2 shows that there are two sources of inefficiency in a monopolistically competitive industry.

1. Unlike perfect competition, with monopolistic competition the equilibrium price exceeds marginal cost. This means that the value to consumers of additional units of output exceeds the cost of producing those units. If output were expanded to the point where the demand curve intersects the marginal cost curve, total surplus could be increased by an amount equal to the yellow-shaded area in Figure 12.2(b). This should not be surprising. We saw in Chapter 10 that monopoly power creates a deadweight loss, and monopoly power exists in monopolistically competitive markets.
2. Note in Figure 12.2 that the monopolistically competitive firm operates with *excess capacity*: Its output is below that which minimizes average cost. Entry of new firms drives profits to zero in both perfectly competitive and

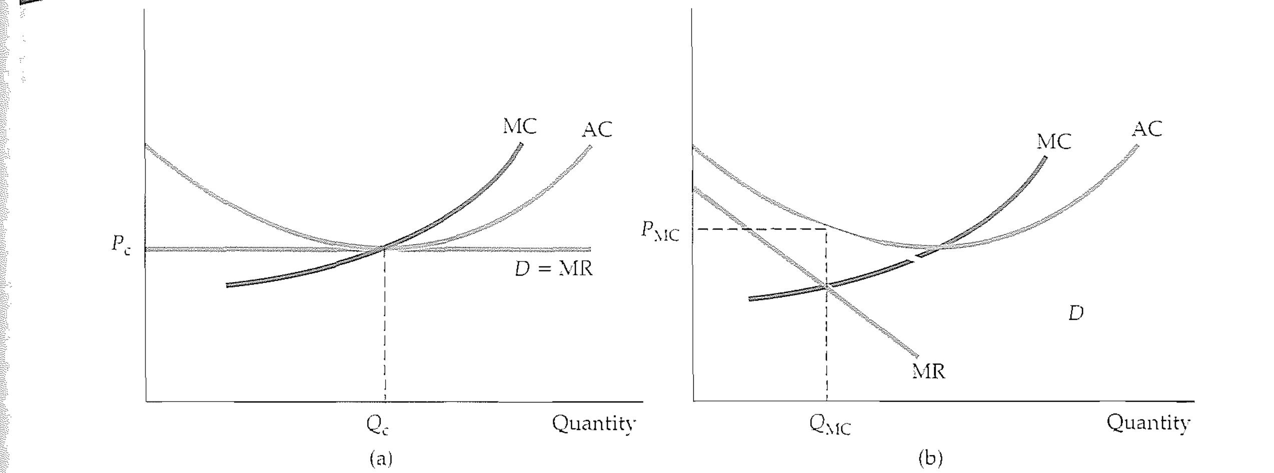

FIGURE 12.2 Comparison of Monopolistically Competitive Equilibrium and Perfectly Competitive Equilibrium

Under perfect competition, as in **(a)**, price equals marginal cost, but under monopolistic competition, price exceeds marginal cost, so there is a deadweight loss as shown by the yellow-shaded area in **(b)**. In both types of markets, entry occurs until profits are driven to zero. Under perfect competition, the demand curve facing the firm is horizontal, so the zero-profit point occurs at the point of minimum average cost. Under monopolistic competition the demand curve is downward-sloping, so the zero-profit point is to the left of the point of minimum average cost. In evaluating monopolistic competition, these inefficiencies must be balanced against the gains to consumers from product diversity.

monopolistically competitive markets. In a perfectly competitive market, each firm faces a horizontal demand curve, so the zero-profit point occurs at minimum average cost, as Figure 12.2(a) shows. In a monopolistically competitive market, however, the demand curve is downward sloping, so the zero-profit point is to the left of minimum average cost. Excess capacity is inefficient because average cost would be lower with fewer firms.

These inefficiencies make consumers worse off. Is monopolistic competition then a socially undesirable market structure that should be regulated? The answer—for two reasons—is probably no:

1. In most monopolistically competitive markets, market power is small. Usually, enough firms compete, with brands that are sufficiently substitutable for one another, so that no single firm has substantial market power. Any deadweight loss from market power will therefore be small. And because firms' demand curves will be fairly elastic, excess capacity will also be small.
2. Any inefficiency must be balanced against an important benefit that monopolistic competition provides: *product diversity*. Most consumers value the ability to choose among a wide variety of competing products and brands that differ in various ways. The gains from product diversity can be large and may easily outweigh the inefficiency costs resulting from downward-sloping demand curves.

EXAMPLE 12.1 Monopolistic Competition in the Markets for Colas and Coffee

The markets for soft drinks and coffee illustrate the characteristics of monopolistic competition. Each market has a variety of brands that differ slightly but are close substitutes for one another. Each brand of cola, for example, tastes a little different from the next. (Can you tell the difference between Coke and Pepsi? Between Coke and Royal Crown Cola?) And each brand of ground coffee has a slightly different flavor, fragrance, and caffeine content. Most consumers develop their own preferences; you might prefer Maxwell House coffee to other brands and buy it regularly. Brand loyalties, however, are usually limited. If the price of Maxwell House were to rise substantially above those of other brands, you and most other consumers who had been buying it would probably switch brands.

Just how much monopoly power does General Foods, the producer of Maxwell House, have with this brand? In other words, how elastic is the demand for Maxwell House? Most large companies carefully study product demands as part of their market research. Company estimates are usually proprietary, but one study of the demands for various brands of colas and ground coffees used a simulated shopping experiment to determine how market shares for each brand would change in response to specific changes in price.[1] Table 12.1 summarizes the results by showing the elasticities of demand for several brands.

First, note that among colas, Royal Crown is much less price elastic than Coke. Although it has a small share of the cola market, its taste is more distinctive than that of Coke, Pepsi, and other brands, so consumers who buy it have stronger brand loyalty. But because Royal Crown has more monopoly power than Coke does not mean that it is more profitable. Profits depend on fixed costs and volume, as well as price. Even if its average profit is smaller, Coke will generate more profit because it has a much larger share of the market.

Second, note that coffees as a group are more price elastic than colas. There is less brand loyalty among coffee buyers than among cola buyers because the differences among coffees are less perceptible than the differences among colas.

TABLE 12.1 Elasticities of Demand for Brands of Colas and Coffee

	BRAND	ELASTICITY OF DEMAND
Colas:	Royal Crown	−2.4
	Coke	−5.2 to −5.7
Ground coffee:	Hills Brothers	−7.1
	Maxwell House	−8.9
	Chase & Sanborn	−5.6

[1] The study was by John R. Nevin, "Laboratory Experiments for Estimating Consumer Demand: A Validation Study," *Journal of Marketing Research* 11 (August 1974): 261–68. In simulated shopping trips, consumers had to choose the brands they preferred from a variety of prepriced brands. The trips were repeated several times, with different prices each time.

Compared with different brands of colas, fewer consumers notice or care about the differences between Hills Brothers and Maxwell House coffees.

With the exception of Royal Crown, all the colas and coffees are quite price elastic. With elasticities on the order of −5 to −9, each brand has only limited monopoly power. This is typical of monopolistic competition.

12.2 Oligopoly

In oligopolistic markets, the products may or may not be differentiated. What matters is that only a few firms account for most or all of total production. In some oligopolistic markets, some or all firms earn substantial profits over the long run because *barriers to entry* make it difficult or impossible for new firms to enter. Oligopoly is a prevalent form of market structure. Examples of oligopolistic industries include automobiles, steel, aluminum, petrochemicals, electrical equipment, and computers.

Why might barriers to entry arise? We discussed some of the reasons in Chapter 10. Scale economies may make it unprofitable for more than a few firms to coexist in the market; patents or access to a technology may exclude potential competitors; and the need to spend money for name recognition and market reputation may discourage entry by new firms. These are "natural" entry barriers—they are basic to the structure of the particular market. In addition, incumbent firms may take *strategic actions* to deter entry. For example, they might threaten to flood the market and drive prices down if entry occurs, and to make the threat credible, they can construct excess production capacity.

Managing an oligopolistic firm is complicated because pricing, output, advertising, and investment decisions involve important strategic considerations. Because only a few firms are competing, each firm must carefully consider how its actions will affect its rivals, and how its rivals are likely to react.

Suppose that because of sluggish car sales, Ford is considering a 10-percent price cut to stimulate demand. It must think carefully about how GM and Chrysler will react. They might not react at all, or they might cut their prices only slightly, in which case Ford could enjoy a substantial increase in sales, largely at the expense of its competitors. Or they might match Ford's price cut, in which case all three automakers will sell more cars but might make much lower profits because of the lower prices. Another possibility is that GM and Chrysler will cut their prices by even more than Ford. They might cut price by 15 percent to punish Ford for rocking the boat, and this in turn might lead to a price war and to a drastic fall in profits for all three firms. Ford must carefully weigh all these possibilities. In fact, for almost any major economic decision a firm makes—setting price, determining production levels, undertaking a major promotion campaign, or investing in new production capacity—it must try to determine the most likely response of its competitors.

These strategic considerations can be complex. When making decisions, each firm must weigh its competitors' reactions, knowing that these competitors will also weigh *its* reactions to *their* decisions. Furthermore, decisions, reactions, reactions to reactions, and so forth are dynamic, evolving over time. When the managers of a firm evaluate the potential consequences of their decisions, they must assume that their competitors are as rational and intelligent as they are. Then, they must put themselves in their competitors' place and consider how they would react.

Equilibrium in an Oligopolistic Market

When we study a market, we usually want to determine the price and quantity that will prevail in equilibrium. For example, we saw that in a perfectly competitive market, the equilibrium price equates the quantity supplied with the quantity demanded. Then we saw that for a monopoly, an equilibrium occurs when marginal revenue equals marginal cost. Finally, when we studied monopolistic competition, we saw how a long-run equilibrium results as the entry of new firms drives profits to zero.

In these markets, each firm could take price or market demand as given and largely ignore its competitors. In an oligopolistic market, however, a firm sets price or output based partly on strategic considerations regarding the behavior of its competitors. At the same time, competitors' decisions depend on the first firm's decision. How, then, can we figure out what the market price and output will be in equilibrium, or whether there will even be an equilibrium? To answer these questions, we need an underlying principle to describe an equilibrium when firms make decisions that explicitly take each other's behavior into account.

In §8.6, we explain that long-run equilibrium occurs when no firm has an incentive to enter or exit because firms are earning zero economic profit and the quantity demanded is equal to the quantity supplied.

Remember how we described an equilibrium in competitive and monopolistic markets: *When a market is in equilibrium, firms are doing the best they can and have no reason to change their price or output.* Thus a competitive market is in equilibrium when the quantity supplied equals the quantity demanded: Each firm is doing the best it can—it is selling all that it produces and is maximizing its profit. Likewise, a monopolist is in equilibrium when marginal revenue equals marginal cost because it, too, is doing the best it can and is maximizing its profit.

Nash Equilibrium

With some modification, we can apply this same principle to an oligopolistic market. Now, however, each firm will want to do the best it can *given what its competitors are doing*. And what should the firm assume that its competitors are doing? Because the firm will do the best it can given what its competitors are doing, *it is natural to assume that these competitors will do the best they can given what that firm is doing*. Each firm, then, takes its competitors into account, and assumes that its competitors are doing likewise.

This may seem a bit abstract at first, but it is logical, and as we will see, it gives us a basis for determining an equilibrium in an oligopolistic market. The concept was first explained clearly by the mathematician John Nash in 1951, so we call the equilibrium it describes a **Nash equilibrium.** It is an important concept that we will use repeatedly:

Nash equilibrium Set of strategies or actions in which each firm does the best it can given its competitors' actions.

> *Nash Equilibrium:* Each firm is doing the best it can given what its competitors are doing.

We discuss this equilibrium concept in more detail in Chapter 13, where we show how it can be applied to a broad range of strategic problems. In this chapter, we will apply it to the analysis of oligopolistic markets.

To keep things as uncomplicated as possible, this chapter will focus largely on markets in which two firms are competing with each other. We call such a market a **duopoly.** Thus each firm has just one competitor to take into account in making its decisions. Although we focus on duopolies, our basic results will also apply to markets with more than two firms.

duopoly Market in which two firms compete with each other.

The Cournot Model

We will begin with a simple model of duopoly first introduced by the French economist Augustin Cournot in 1838. Suppose the firms produce a homogeneous good and know the market demand curve. *Each firm must decide how much to produce, and the two firms make their decisions at the same time.* When making its production decision, each firm takes its competitor into account. It knows that its competitor is *also* deciding how much to produce, and the market price will depend on the *total output* of both firms.

Recall from §8.8 that when firms produce homogeneous or identical goods, consumers consider only price when making their purchasing decisions.

The essence of the **Cournot model** is that *each firm treats the output level of its competitor as fixed and then decides how much to produce.* To see how this works, let's consider the output decision of Firm 1. Suppose Firm 1 thinks that Firm 2 will produce nothing. In that case, Firm 1's demand curve is the market demand curve. In Figure 12.3 this is shown as $D_1(0)$, which means the demand curve for Firm 1, assuming Firm 2 produces zero. Figure 12.3 also shows the corresponding marginal revenue curve $MR_1(0)$. We have assumed that Firm 1's marginal cost MC_1 is constant. As shown in the figure, Firm 1's profit-maximizing output is 50 units, the point where $MR_1(0)$ intersects MC_1. So if Firm 2 produces zero, Firm 1 should produce 50.

Cournot model Oligopoly model in which firms produce a homogeneous good, each firm treats the output of its competitors as fixed, and all firms decide simultaneously how much to produce.

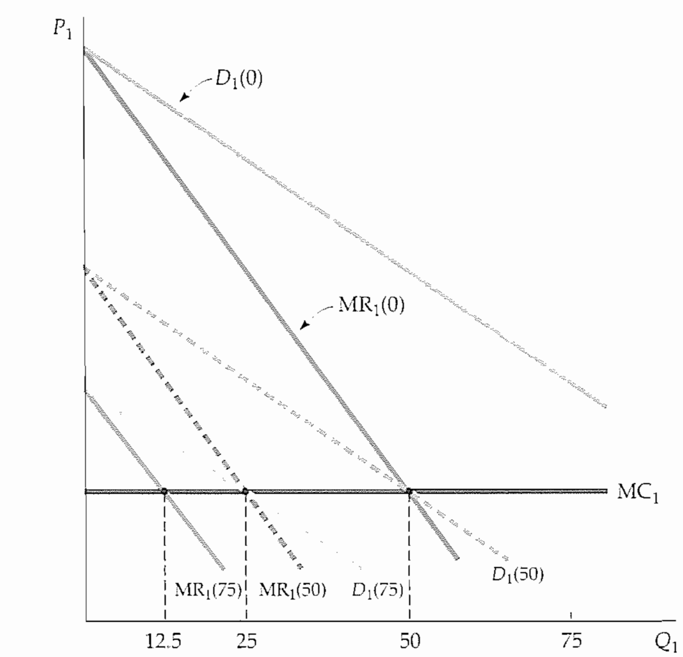

FIGURE 12.3 Firm 1's Output Decision

Firm 1's profit-maximizing output depends on how much it thinks Firm 2 will produce. If it thinks Firm 2 will produce nothing, its demand curve, labeled $D_1(0)$, is the market demand curve. The corresponding marginal revenue curve, labeled $MR_1(0)$, intersects Firm 1's marginal cost curve MC_1 at an output of 50 units. If Firm 1 thinks Firm 2 will produce 50 units, its demand curve, $D_1(50)$, is shifted to the left by this amount. Profit maximization now implies an output of 25 units. Finally, if Firm 1 thinks Firm 2 will produce 75 units, Firm 1 will produce only 12.5 units.

Suppose, instead, that Firm 1 thinks Firm 2 will produce 50 units. Then Firm 1's demand curve is the market demand curve shifted to the left by 50. In Figure 12.3, this curve is labeled $D_1(50)$, and the corresponding marginal revenue curve is labeled $MR_1(50)$. Firm 1's profit-maximizing output is now 25 units, the point where $MR_1(50) = MC_1$. Now, suppose Firm 1 thinks Firm 2 will produce 75 units. Then Firm 1's demand curve is the market demand curve shifted to the left by 75. It is labeled $D_1(75)$ in Figure 12.3, and the corresponding marginal revenue curve is labeled $MR_1(75)$. Firm 1's profit-maximizing output is now 12.5 units, the point where $MR_1(75) = MC_1$. Finally, suppose Firm 1 thinks Firm 2 will produce 100 units. Then Firm 1's demand and marginal revenue curves (which are not shown in the figure) would intersect its marginal cost curve on the vertical axis; if Firm 1 thinks that Firm 2 will produce 100 units or more, it should produce nothing.

reaction curve Relationship between a firm's profit-maximizing output and the amount it thinks its competitor will produce.

Reaction Curves To summarize: If Firm 1 thinks Firm 2 will produce nothing, it will produce 50; if it thinks Firm 2 will produce 50, it will produce 25; if it thinks Firm 2 will produce 75, it will produce 12.5; and if it thinks Firm 2 will produce 100, then it will produce nothing. *Firm 1's profit-maximizing output is thus a decreasing schedule of how much it thinks Firm 2 will produce.* We call this schedule Firm 1's **reaction curve** and denote it by $Q_1^*(Q_2)$. This curve is plotted in Figure 12.4, where each of the four output combinations we found above is shown as an *x*.

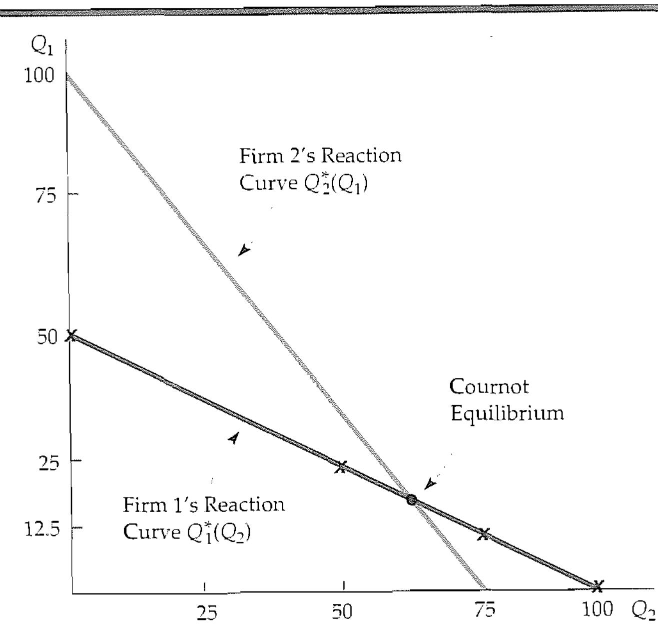

FIGURE 12.4 Reaction Curves and Cournot Equilibrium

Firm 1's reaction curve shows how much it will produce as a function of how much it thinks Firm 2 will produce. (The **x**s, at $Q_2 = 0, 50$, and 75, correspond to the examples shown in Figure 12.3.) Firm 2's reaction curve shows its output as a function of how much it thinks Firm 1 will produce. In Cournot equilibrium, each firm correctly assumes the amount that its competitor will produce and thereby maximizes its own profits. Therefore neither firm will move from this equilibrium.

We can go through the same kind of analysis for Firm 2; that is, we can determine Firm 2's profit-maximizing quantity given various assumptions about how much Firm 1 will produce. The result will be a reaction curve for Firm 2—i.e., a schedule $Q_2^*(Q_1)$ that relates its output to the output it thinks Firm 1 will produce. If Firm 2's marginal cost curve is different from that of Firm 1, its reaction curve will also differ in form. For example, Firm 2's reaction curve might look like the one drawn in Figure 12.4.

Cournot Equilibrium How much will each firm produce? Each firm's reaction curve tells it how much to produce, given the output of its competitor. In equilibrium, each firm sets output according to its own reaction curve; the equilibrium output levels are therefore found at the *intersection* of the two reaction curves. We call the resulting set of output levels a **Cournot equilibrium.** In this equilibrium, each firm correctly assumes how much its competitor will produce, and it maximizes its profit accordingly.

Cournot equilibrium Equilibrium in the Cournot model, in which each firm correctly assumes how much its competitor will produce and sets its own production level accordingly.

Note that this Cournot equilibrium is an example of a Nash equilibrium.[2] Remember that in a Nash equilibrium, each firm is doing the best it can given what its competitors are doing. As a result, no firm would individually want to change its behavior. In the Cournot equilibrium, each duopolist is producing an amount that maximizes its profit *given what its competitor is producing*, so neither would want to change its output.

Suppose the two firms are initially producing output levels that differ from the Cournot equilibrium. Will they adjust their outputs until the Cournot equilibrium is reached? Unfortunately, the Cournot model says nothing about the dynamics of the adjustment process. In fact, during any adjustment process, the model's central assumption that each firm can assume that its competitor's output is fixed will not hold. Because both firms would be adjusting their outputs, neither output would be fixed. We need different models to understand dynamic adjustment and we will examine some in Chapter 13.

When is it rational for each firm to assume that its competitor's output is fixed? It is rational if the two firms are choosing their outputs only once because then their outputs cannot change. It is also rational once they are in Cournot equilibrium because then neither firm will have any incentive to change its output. When using the Cournot model, we must therefore confine ourselves to the behavior of firms in equilibrium.

The Linear Demand Curve—An Example

Let's work through an example—two identical firms facing a linear market demand curve. This will help clarify the meaning of a Cournot equilibrium and let us compare it with the competitive equilibrium and the equilibrium that results if the firms collude and choose their output levels cooperatively.

Suppose our duopolists face the following market demand curve:

$$P = 30 - Q$$

where Q is the *total* production of both firms (i.e., $Q = Q_1 + Q_2$). Also, suppose that both firms have zero marginal cost:

$$MC_1 = MC_2 = 0$$

[2] Thus it is sometimes called a *Cournot-Nash equilibrium*.

Then we can determine the reaction curve for Firm 1 as follows. To maximize profit, it sets marginal revenue equal to marginal cost. Its total revenue R_1 is given by

$$R_1 = PQ_1 = (30 - Q)Q_1$$
$$= 30Q_1 - (Q_1 + Q_2)Q_1$$
$$= 30Q_1 - Q_1^2 - Q_2Q_1$$

Its marginal revenue MR_1 is just the incremental revenue ΔR_1 resulting from an incremental change in output ΔQ_1:

$$MR_1 = \Delta R_1/\Delta Q_1 = 30 - 2Q_1 - Q_2$$

Now, setting MR_1 equal to zero (the firm's marginal cost) and solving for Q_1, we find:

$$\textit{Firm 1's reaction curve:} \quad Q_1 = 15 - \frac{1}{2}Q_2 \qquad \textbf{(12.1)}$$

The same calculation applies to Firm 2:

$$\textit{Firm 2's reaction curve:} \quad Q_2 = 15 - \frac{1}{2}Q_1 \qquad \textbf{(12.2)}$$

The equilibrium output levels are the values for Q_1 and Q_2 that are at the intersection of the two reaction curves—i.e., the levels that solve equations (12.1) and (12.2). By replacing Q_2 in equation (12.1) with the expression on the right-hand side of (12.2), you can verify that the equilibrium output levels are

$$\textit{Cournot equilibrium:} \quad Q_1 = Q_2 = 10$$

The total quantity produced is therefore $Q = Q_1 + Q_2 = 20$, so the equilibrium market price is $P = 30 - Q = 10$.

Figure 12.5 shows the Cournot reaction curves and this Cournot equilibrium. Note that Firm 1's reaction curve shows its output Q_1 in terms of Firm 2's output Q_2. Likewise, Firm 2's reaction curve shows Q_2 in terms of Q_1. (Because the firms are identical, the two reaction curves have the same form. They look different because one gives Q_1 in terms of Q_2 and the other gives Q_2 in terms of Q_1.) The Cournot equilibrium is at the intersection of the two curves. At this point, each firm is maximizing its own profit, given its competitor's output.

We have assumed that the two firms compete with each other. Suppose, instead, that the antitrust laws were relaxed and the two firms could collude. They would set their outputs to maximize *total profit*, and presumably they would split that profit evenly. Total profit is maximized by choosing total output Q so that marginal revenue equals marginal cost, which in this example is zero. Total revenue for the two firms is

$$R = PQ = (30 - Q)Q = 30Q - Q^2$$

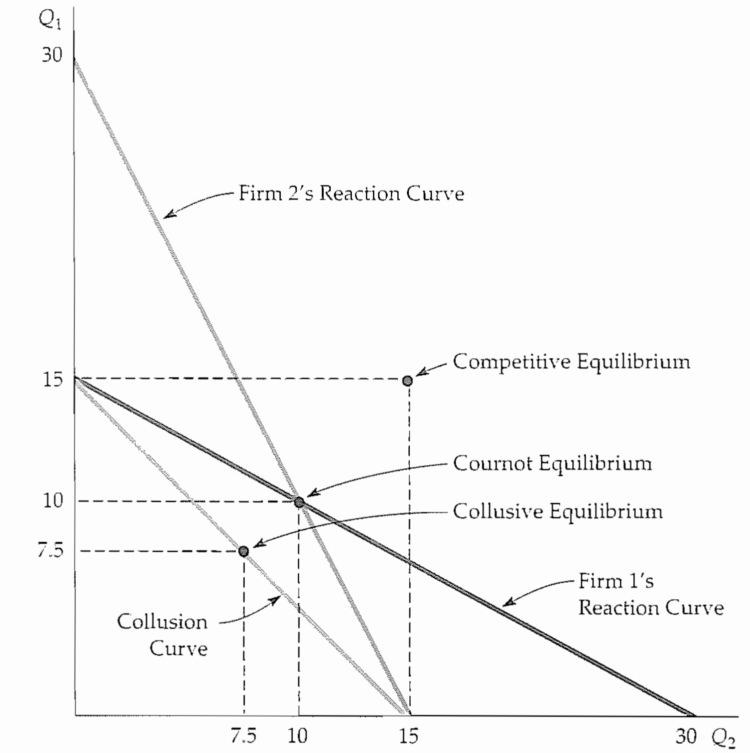

FIGURE 12.5 Duopoly Example

The demand curve is $P = 30 - Q$, and both firms have zero marginal cost. In Cournot equilibrium, each firm produces 10. The collusion curve shows combinations of Q_1 and Q_2 that maximize *total* profits. If the firms collude and share profits equally, each will produce 7.5. Also shown is the competitive equilibrium, in which price equals marginal cost and profit is zero.

Marginal revenue is therefore

$$MR = \Delta R/\Delta Q = 30 - 2Q$$

Setting MR equal to zero, we see that total profit is maximized when $Q = 15$.

Any combination of outputs Q_1 and Q_2 that add up to 15 maximizes total profit. The curve $Q_1 + Q_2 = 15$, called the *collusion curve*, therefore gives all pairs of outputs Q_1 and Q_2 that maximize total profit. This curve is also shown in Figure 12.5. If the firms agree to share profits equally, each will produce half of the total output:

$$Q_1 = Q_2 = 7.5$$

As you would expect, both firms now produce less—and earn higher profits—than in the Cournot equilibrium. Figure 12.5 shows this collusive equilibrium and the *competitive* output levels found by setting price equal to marginal cost. (You can verify that they are $Q_1 = Q_2 = 15$, which implies that each firm makes zero profit.) Note that the Cournot outcome is much better (for the firms) than perfect competition, but not as good as the outcome from collusion.

First Mover Advantage—The Stackelberg Model

We have assumed that our two duopolists make their output decisions at the same time. Now let's see what happens if one of the firms can set its output first. There are two questions of interest. First, is it advantageous to go first? Second, how much will each firm produce?

Continuing with our example, we assume both firms have zero marginal cost, and that the market demand curve is given by $P = 30 - Q$, where Q is the total output. *Suppose Firm 1 sets its output first and then Firm 2, after observing Firm 1's output, makes its output decision.* In setting output, *Firm 1 must therefore consider how Firm 2 will react.* This **Stackelberg model** of duopoly is different from the Cournot model, in which neither firm has any opportunity to react.

Stackelberg model Oligopoly model in which one firm sets its output before other firms do.

Let's begin with Firm 2. Because it makes its output decision *after* Firm 1, it takes Firm 1's output as fixed. Therefore, Firm 2's profit-maximizing output is given by its Cournot reaction curve, which we found to be

$$\textit{Firm 2's reaction curve:} \quad Q_2 = 15 - \frac{1}{2}Q_1 \tag{12.2}$$

What about Firm 1? To maximize profit, it chooses Q_1, so that its marginal revenue equals its marginal cost of zero. Recall that Firm 1's revenue is

$$R_1 = PQ_1 = 30Q_1 - Q_1^2 - Q_2Q_1 \tag{12.3}$$

Because R_1 depends on Q_2, Firm 1 must anticipate how much Firm 2 will produce. Firm 1 knows, however, that Firm 2 will choose Q_2 according to the reaction curve (12.2). Substituting equation (12.2) for Q_2 into equation (12.3), we find that Firm 1's revenue is

$$R_1 = 30Q_1 - Q_1^2 - Q_1\left(15 - \frac{1}{2}Q_1\right)$$

$$= 15Q_1 - \frac{1}{2}Q_1^2$$

Its marginal revenue is therefore

$$MR_1 = \Delta R_1/\Delta Q_1 = 15 - Q_1 \tag{12.4}$$

Setting $MR_1 = 0$ gives $Q_1 = 15$. And from Firm 2's reaction curve (12.2), we find that $Q_2 = 7.5$. Firm 1 produces twice as much as Firm 2 and makes twice as much profit. *Going first gives Firm 1 an advantage.* This may appear counterintuitive: It seems disadvantageous to announce your output first. Why, then, is going first a strategic advantage?

The reason is that announcing first creates a *fait accompli:* No matter what your competitor does, your output will be large. To maximize profit, your competitor must take your large output level as given and set a low level of output for itself. (If your competitor produced a large level of output, it would drive price down and you would both lose money. So unless your competitor views "getting even" as more important than making money, it would be irrational for it to produce a large amount.) As we will see in Chapter 13, this kind of "first mover advantage" occurs in many strategic situations.

The Cournot and Stackelberg models are alternative representations of oligopolistic behavior. Which model is the more appropriate depends on the industry. For an industry composed of roughly similar firms, none of which has a strong operating advantage or leadership position, the Cournot model is probably the more appropriate. On the other hand, some industries are dominated by a large firm that usually takes the lead in introducing new products or setting price; the mainframe computer market is an example, with IBM the leader. Then the Stackelberg model may be more realistic.

12.3 Price Competition

We have assumed that our oligopolistic firms compete by setting quantities. In many oligopolistic industries, however, competition occurs along price dimensions. For example, for GM, Ford, and Daimler-Chrysler, price is a key strategic variable, and each firm chooses its price with its competitors in mind. In this section we use the Nash equilibrium concept to study price competition, first in an industry that produces a homogeneous good and then in an industry with some degree of product differentiation.

Price Competition with Homogeneous Products—The Bertrand Model

The **Bertrand model** was developed in 1883 by another French economist, Joseph Bertrand. Like the Cournot model, it applies to firms that produce the same homogeneous good and make their decisions at the same time. In this case, however, the firms choose *prices* instead of quantities. As we will see, this change can dramatically affect the market outcome.

Bertrand model Oligopoly model in which firms produce a homogeneous good, each firm treats the price of its competitors as fixed, and all firms decide simultaneously what price to charge.

Let's return to the duopoly example of the last section, in which the market demand curve is

$$P = 30 - Q$$

where $Q = Q_1 + Q_2$ is again total production of a homogeneous good. This time, we will assume that both firms have a marginal cost of \$3:

$$MC_1 = MC_2 = 3$$

As an exercise, you can show that the Cournot equilibrium for this duopoly, which results when both firms choose output simultaneously, is $Q_1 = Q_2 = 9$. You can also check that in this Cournot equilibrium, the market price is \$12, so that each firm makes a profit of \$81.

Now suppose that these two duopolists compete by simultaneously choosing a *price* instead of a quantity. What price will each firm choose, and how much profit will each earn? To answer these questions, note that because the good is homogeneous, consumers will purchase only from the lowest-price seller. Thus, if the two firms charge different prices, the lower-priced firm will supply the entire market and the higher-priced firm will sell nothing. If both firms charge the same price, consumers will be indifferent as to which firm they buy from and each firm will supply half the market.

What is the Nash equilibrium in this case? If you think about this a little, you will see that because of the incentive to cut prices, the Nash equilibrium is the competitive outcome—i.e., both firms set price equal to marginal cost:

$P_1 = P_2 = \$3$. Then industry output is 27 units, of which each firm produces 13.5 units. And since price equals marginal cost, both firms earn zero profit. To check that this is a Nash equilibrium, ask whether either firm would have any incentive to change its price. Suppose Firm 1 raised its price. It would then lose all of its sales to Firm 2 and therefore be no better off. If instead it lowered its price, it would capture the entire market but would lose money on every unit it produced; again, it would be worse off. Therefore, Firm 1 (and likewise Firm 2) has no incentive to deviate: It is doing the best it can to maximize profit, given what its competitor is doing.

Why couldn't there be a Nash equilibrium in which the firms charged the same price, but a higher one (say, \$5), so that each made some profit? Because in this case, if either firm lowered its price just a little, it could capture the entire market and nearly double its profit. Thus each firm would want to undercut its competitor. Such undercutting would continue until the price dropped to \$3.

By changing the strategic choice variable from output to price, we get a dramatically different outcome. In the Cournot model, because each firm produces only 9 units, the market price is \$12. Now the market price is \$3. In the Cournot model, each firm made a profit; in the Bertrand model, the firms price at marginal cost and make no profit.

The Bertrand model has been criticized on several counts. First, when firms produce a homogeneous good, it is more natural to compete by setting quantities rather than prices. Second, even if firms do set prices *and* choose the same price (as the model predicts), what share of total sales will go to each one? We *assumed* that sales would be divided equally among the firms, but there is no reason why this must be the case. But despite these shortcomings, the Bertrand model is useful because it shows how the equilibrium outcome in an oligopoly can depend crucially on the firms' choice of strategic variable.[3]

Price Competition with Differentiated Products

Oligopolistic markets often have at least some degree of product differentiation.[4] Market shares are determined not just by prices, but also by differences in the design, performance, and durability of each firm's product. In such cases, it is natural for firms to compete by choosing prices rather than quantities.

To see how price competition with differentiated products can work, let's go through the following simple example. Suppose each of two duopolists has fixed costs of \$20 but zero variable costs, and that they face the same demand curves:

$$\text{Firm 1's demand:} \quad Q_1 = 12 - 2P_1 + P_2 \qquad (12.5a)$$

$$\text{Firm 2's demand:} \quad Q_2 = 12 - 2P_2 + P_1 \qquad (12.5b)$$

[3] Also, it has been shown that if firms produce a homogeneous good and compete by first setting output *capacities* and then setting price, the Cournot equilibrium in quantities again results. See David Kreps and Jose Scheinkman, "Quantity Precommitment and Bertrand Competition Yield Cournot Outcomes," *Bell Journal of Economics* 14 (1983):326–38.

[4] Product differentiation can exist even for a seemingly homogeneous product. Consider gasoline, for example. Although gasoline itself is a homogeneous good, service stations differ in terms of location and services provided. As a result, gasoline prices may differ from one service station to another.

where P_1 and P_2 are the prices that Firms 1 and 2 charge, respectively, and Q_1 and Q_2 are the resulting quantities that they sell. Note that the quantity that each firm can sell decreases when it raises its own price but increases when its competitor charges a higher price.

If both firms set their prices at the same time, we can use the Cournot model to determine the resulting equilibrium. Each firm will choose its own price, takings its competitor's price as fixed. Now consider Firm 1. Its profit π_1 is its revenue P_1Q_1 less its fixed cost of \$20. Substituting for Q_1 from the demand curve of equation (12.5a), we have

$$\pi_1 = P_1Q_1 - 20 = 12P_1 - 2P_1^2 + P_1P_2 - 20$$

At what price P_1 is this profit maximized? The answer depends on P_2, which Firm 1 assumes to be fixed. However, whatever price Firm 2 is charging, Firm 1's profit is maximized when the incremental profit from a very small increase in its own price is just zero. Taking P_2 as fixed, Firm 1's profit-maximizing price is therefore given by

$$\Delta\pi_1/\Delta P_1 = 12 - 4P_1 + P_2 = 0$$

This equation can be rewritten to give the following pricing rule, or *reaction curve*, for Firm 1:

$$\text{Firm 1's reaction curve:} \quad P_1 = 3 + \frac{1}{4}P_2$$

This tells Firm 1 what price to set, given the price P_2 that Firm 2 is setting. We can similarly find the following pricing rule for Firm 2:

$$\text{Firm 2's reaction curve:} \quad P_2 = 3 + \frac{1}{4}P_1$$

These reaction curves are drawn in Figure 12.6. The Nash equilibrium is at the point where the two reaction curves cross; you can verify that each firm is then charging a price of \$4 and earning a profit of \$12. *At this point, because each firm is doing the best it can given the price its competitor has set, neither firm has an incentive to change its price.*

Now suppose the two firms collude: Instead of choosing their prices independently, they both decide to charge the same price, which will be the price that maximizes both of their profits. You can verify that the firms would then charge \$6, and that they would be better off colluding because each would now earn a profit of \$16.[5] Figure 12.6 shows this collusive equilibrium.

[5] The firms have the same costs, so they will charge the same price P. Total profit is given by

$$\pi_T = \pi_1 + \pi_2 = 24P - 4P^2 + 2P^2 - 40 = 24P - 2P^2 - 40.$$

This is maximized when $\Delta\pi_T/\Delta P = 0$. $\Delta\pi_T/\Delta P = 24 - 4P$, so the joint profit-maximizing price is $P = 6$. Each firm's profit is therefore

$$\pi_1 = \pi_2 = 12P - P^2 - 20 = 72 - 36 - 20 = \$16$$

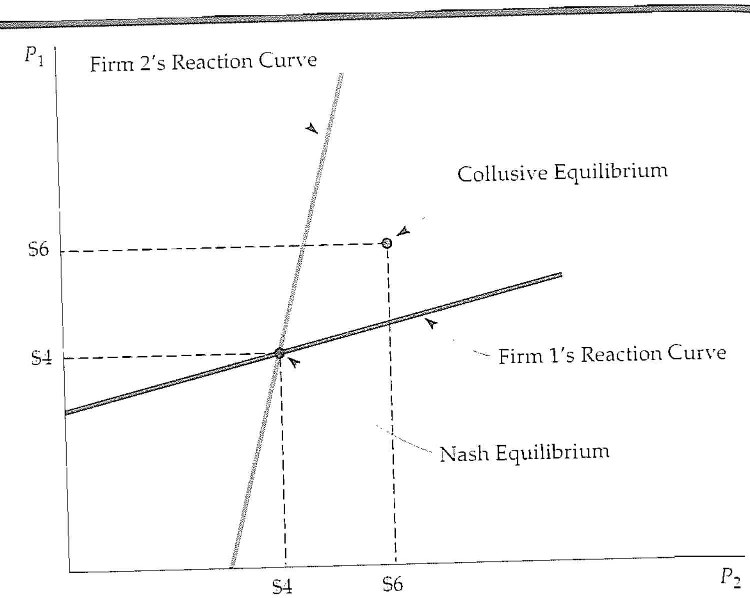

FIGURE 12.6 Nash Equilibrium in Prices

Here two firms sell a differentiated product, and each firm's demand depends both on its own price and on its competitor's price. The two firms choose their prices at the same time, each taking its competitor's price as given. Firm 1's reaction curve gives its profit-maximizing price as a function of the price that Firm 2 sets, and similarly for Firm 2. The Nash equilibrium is at the intersection of the two reaction curves; when each firm charges a price of $4, it is doing the best it can given its competitor's price and has no incentive to change price. Also shown is the collusive equilibrium: If the firms cooperatively set price, they would choose $6.

Finally, suppose Firm 1 sets its price first and, after observing Firm 1's decision, Firm 2 makes its pricing decision. Unlike the Stackelberg model in which the firms set their quantities, in this case Firm 1 would be at a distinct *disadvantage* by moving first. (To see this, calculate Firm 1's profit-maximizing price, *taking Firm 2's reaction curve into account*.) Why is moving first now a disadvantage? Because it gives the firm that moves second an opportunity to undercut slightly and thereby capture a larger market share. (See Exercise 10 at the end of the chapter.)

EXAMPLE 12.2 A Pricing Problem for Procter & Gamble

When Procter & Gamble (P&G) planned to enter the Japanese market for Gypsy Moth Tape, it knew its production costs and understood the market demand curve but found it hard to determine the right price to charge because two other firms—Kao Soap, Ltd., and Unilever, Ltd.—were also planning to enter the market. All three firms would be choosing their prices at about the same time, and P&G had to take this into account when setting its own price.[6]

[6] This example is based on classroom material developed by Professor John Hauser of MIT. To protect P&G's proprietary interests, some of the facts about the product and the market have been altered. The fundamental description of P&G's problem, however, is accurate.

Because all three firms were using the same technology for producing Gypsy Moth Tape, they had the same production costs. Each firm faced a fixed cost of $480,000 per month and a variable cost of $1 per unit. From market research, P&G ascertained that its demand curve for monthly sales was

$$Q = 3375P^{-3.5}(P_U)^{.25}(P_K)^{.25}$$

where Q is monthly sales in thousands of units, and P, P_U, and P_K are P&G's, Unilever's, and Kao's prices, respectively. Now, put yourself in P&G's position. Assuming that Unilever and Kao face the same demand conditions, *with what price should you enter the market, and how much profit should you expect to earn?*

You might begin by calculating the profit you would earn as a function of the price you charge, under alternative assumptions about the prices that Unilever and Kao will charge. Using the demand curve and cost numbers given above, we have done these calculations and tabulated the results in Table 12.2. Each entry shows your profit, in thousands of dollars per month, for a particular combination of prices (while assuming in each case that Unilever and Kao set the same price). For example, if you charge $1.30 and Unilever and Kao both charge $1.50, you will earn a profit of $15,000 per month.

Remember that in all likelihood, the managers of Unilever and Kao are making the same calculations that you are and probably have their own versions of Table 12.2. Now suppose your competitors charge $1.50 or more. As the table shows, you would want to charge only $1.40 because that price gives you the highest profit. (For example, if they charged $1.50, you would make $29,000 per month by charging $1.40 but only $20,000 by charging $1.50, and $15,000 by charging $1.30.) Consequently, you would not want to charge $1.50 (or more). Assuming that your competitors have followed the same reasoning, you should not expect them to charge $1.50 (or more) either.

What if your competitors charge $1.30? In that case, you will lose money, but you will lose the least amount of money ($6,000 per month) by charging $1.40. Your competitors would therefore not expect you to charge $1.30, and by the same reasoning, you should not expect them to charge a price this low. What price lets you do the best you can, given your competitors' prices? It is $1.40.

TABLE 12.2 P&G's Profit (in thousands of dollars per month)

P&G's Price ($)	COMPETITOR'S (EQUAL) PRICES ($) 1.10	1.20	1.30	1.40	1.50	1.60	1.70	1.80
1.10	−226	−215	−204	−194	−183	−174	−165	−155
1.20	−106	−89	−73	−58	−43	−28	−15	−2
1.30	−56	−37	−19	2	15	31	47	62
1.40	−44	−25	−6	12	29	46	62	78
1.50	−52	−32	−15	3	20	36	52	68
1.60	−70	−51	−34	−18	−1	14	30	44
1.70	−93	−76	−59	−44	−28	−13	1	15
1.80	−118	−102	−87	−72	−57	−44	−30	−17

This is also the price at which your competitors are doing the best *they* can, so it is a Nash equilibrium.[7] As the table shows, in this equilibrium you and your competitors each make a profit of $12,000 per month.

If you could *collude* with your competitors, you could make a larger profit. You would all agree to charge $1.50, and each of you would earn $20,000. But this collusive agreement might be hard to enforce: You could increase your profit further at your competitor's expense by dropping your price below theirs, and of course your competitors could do the same thing to you.

12.4 Competition versus Collusion: The Prisoners' Dilemma

A Nash equilibrium is a *noncooperative* equilibrium: Each firm makes the decisions that give it the highest possible profit, given the actions of its competitors. As we have seen, the resulting profit earned by each firm is higher than it would be under perfect competition but lower than if the firms colluded.

Collusion is, however, illegal, and most managers prefer to stay out of jail. But if cooperation can lead to higher profits, why don't firms cooperate *without* explicitly colluding? In particular, if you and your competitor can both figure out the profit-maximizing price you would agree to charge *if* you were to collude, *why not just set that price and hope your competitor will do the same*? If your competitor *does* do the same, you will both make more money.

The problem is that your competitor *probably won't* choose to set price at the collusive level. Why not? *Because your competitor would do better by choosing a lower price, even if it knew that you were going to set price at the collusive level.*

To understand this, let's go back to our example of price competition from the last section. The firms in that example each have a fixed cost of $20, have zero variable cost, and face the following demand curves:

$$\textit{Firm 1's demand:} \quad Q_1 = 12 - 2P_1 + P_2 \qquad (12.6a)$$

$$\textit{Firm 2's demand:} \quad Q_2 = 12 - 2P_2 + P_1 \qquad (12.6b)$$

We found that in the Nash equilibrium each firm will charge a price of $4 and earn a profit of $12, whereas if the firms collude, they will charge a price of $6 and earn a profit of $16. Now suppose that the firms do not collude, but that Firm 1 charges the $6 collusive price, hoping that Firm 2 will do the same. If Firm 2 *does* do the same, it will earn a profit of $16. But what if it charges the $4 price instead? In that case, Firm 2 would earn a profit of

$$\pi_2 = P_2Q_2 - 20 = (4)[12 - (2)(4) + 6] - 20 = \$20$$

Firm 1, on the other hand, will earn a profit of only

$$\pi_1 = P_1Q_1 - 20 = (6)[12 - (2)(6) + 4] - 20 = \$4$$

[7] This Nash equilibrium can also be derived algebraically from the demand curve and cost data above. We leave this to you as an exercise.

TABLE 12.3 Payoff Matrix for Pricing Game

		FIRM 2	
		Charge $4	Charge $6
FIRM 1	Charge $4	$12, $12	$20, $4
	Charge $6	$4, $20	$16, $16

So if Firm 1 charges $6 but Firm 2 charges only $4, Firm 2's profit will increase to $20. And it will do so at the expense of Firm 1's profit, which will fall to $4. Clearly, Firm 2 does best by charging only $4. And similarly, Firm 1 does best by charging only $4. If Firm 2 charges $6 and Firm 1 charges $4, Firm 1 will earn a $20 profit and Firm 2 only $4.

Payoff Matrix

Table 12.3 summarizes the results of these different possibilities. In deciding what price to set, the two firms are playing a **noncooperative game:** Each firm independently does the best it can, taking its competitor into account. Table 12.3 is called the **payoff matrix** for this game because it shows the profit (or payoff) to each firm given its decision and the decision of its competitor. For example, the upper left-hand corner of the payoff matrix tells us that if both firms charge $4, each will make a $12 profit. The upper right-hand corner tells us that if Firm 1 charges $4 and Firm 2 charges $6, Firm 1 will make $20 and Firm 2 $4.

noncooperative game Game in which negotiation and enforcement of binding contracts are not possible.

payoff matrix Table showing profit (or payoff) to each firm given its decision and the decision of its competitor.

This payoff matrix can clarify the answer to our original question: Why don't firms behave cooperatively, and thereby earn higher profits, even if they can't collude? In this case, cooperating means *both* firms charging $6 instead of $4 and thereby earning $16 instead of $12. The problem is that each firm always makes more money by charging $4, *no matter what its competitor does*. As the payoff matrix shows, if Firm 2 charges $4, Firm 1 does best by charging $4. And if Firm 2 charges $6, Firm 1 still does best by charging $4. Similarly, Firm 2 always does best by charging $4, no matter what Firm 1 does. As a result, unless the two firms can sign an enforceable agreement to charge $6, neither firm can expect its competitor to charge $6, and both will charge $4.

The Prisoners' Dilemma

A classic example in game theory, called the **prisoners' dilemma**, illustrates the problem faced by oligopolistic firms. It goes as follows: Two prisoners have been accused of collaborating in a crime. They are in separate jail cells and cannot communicate with each other. Each has been asked to confess. If both prisoners confess, each will receive a prison term of five years. If neither confesses, the prosecution's case will be difficult to make, so the prisoners can expect to plea bargain and receive terms of two years. On the other hand, if one prisoner confesses and the other does not, the one who confesses will receive a term of only one year, while the other will go to prison for ten years. If you were one of these prisoners, what would you do—confess or not confess?

prisoners' dilemma Game theory example in which two prisoners must decide separately whether to confess to a crime; if a prisoner confesses, he will receive a lighter sentence and his accomplice will receive a heavier one, but if neither confesses, sentences will be lighter than if both confess.

The payoff matrix in Table 12.4 summarizes the possible outcomes. (Note that the "payoffs" are negative; the entry in the lower right-hand corner of the payoff matrix means a two-year sentence for each prisoner.) As the table shows, our prisoners face a dilemma. If they could both agree not to confess (in a way that

TABLE 12.4 Payoff Matrix for Prisoners' Dilemma

		PRISONER B	
		Confess	Don't confess
PRISONER A	Confess	−5, −5	−1, −10
	Don't confess	−10, −1	−2, −2

would be binding), then each would go to jail for only two years. But they can't talk to each other, and even if they could, can they trust each other? If Prisoner *A* does not confess, he risks being taken advantage of by his former accomplice. After all, *no matter what Prisoner A does, Prisoner B comes out ahead by confessing.* Likewise, Prisoner *A* always comes out ahead by confessing, so Prisoner *B* must worry that by not confessing, she will be taken advantage of. Therefore, both prisoners will probably confess and go to jail for five years.

Oligopolistic firms often find themselves in a prisoners' dilemma. They must decide whether to compete aggressively, attempting to capture a larger share of the market at their competitor's expense, or to "cooperate" and compete more passively, coexisting with their competitors and settling for their current market share, and perhaps even implicitly colluding. If the firms compete passively, setting high prices and limiting output, they will make higher profits than if they compete aggressively.

Like our prisoners, however, each firm has an incentive to "fink" and undercut its competitors, and each knows that its competitors have the same incentive. As desirable as cooperation is, each firm worries—with good reason—that if it competes passively, its competitor might decide to compete aggressively and seize the lion's share of the market. In the pricing problem illustrated in Table 12.3, both firms do better by "cooperating" and charging a high price. But the firms are in a prisoners' dilemma, where neither can trust its competitor to set a high price.

EXAMPLE 12.3 Procter & Gamble in a Prisoners' Dilemma

In Example 12.2, we examined the problem that arose when P&G, Unilever, and Kao Soap all planned to enter the Japanese market for Gypsy Moth Tape at the same time. They all faced the same cost and demand conditions, and each firm had to decide on a price that took its competitors into account. In Table 12.2, we tabulated the profits to P&G corresponding to alternative prices that it and its competitors might charge. We argued that P&G should expect its competitors to charge a price of $1.40 and should do the same.[8]

P&G would be better off if it *and its competitors* all charged a price of $1.50. This is clear from the payoff matrix in Table 12.5. This payoff matrix is the portion of Table 12.2 corresponding to prices of $1.40 and $1.50, with the payoffs to P&G's competitors also tabulated.[9] If all the firms charge $1.50, each will make a profit of $20,000 per month, instead of the $12,000 per month they make by charging $1.40. Then why don't they charge $1.50?

[8] As in Example 12.2, some of the facts about the product and the market have been altered to protect P&G's proprietary interests.

TABLE 12.5 Payoff Matrix for Pricing Problem

		UNILEVER AND KAO	
		Charge $1.40	Charge $1.50
P&G	Charge $1.40	$12, $12	$29, $11
	Charge $1.50	$3, $21	$20, $20

Because these firms are in a prisoners' dilemma. No matter what Unilever and Kao do, P&G makes more money by charging $1.40. For example, if Unilever and Kao charge $1.50, P&G can make $29,000 per month by charging $1.40, versus $20,000 by charging $1.50. This is also true for Unilever and Kao. For example, if P&G charges $1.50 and Unilever and Kao both charge $1.40, P&G's competitors will each make $21,000, instead of $20,000.[10] As a result, P&G knows that if it sets a price of $1.50, its competitors will have a strong incentive to undercut and charge $1.40. P&G will then have only a small share of the market and make only $3000 per month profit. Should P&G make a leap of faith and charge $1.50? If you were faced with this dilemma, what would you do?

12.5 Implications of the Prisoners' Dilemma for Oligopolistic Pricing

Does the prisoners' dilemma doom oligopolistic firms to aggressive competition and low profits? Not necessarily. Although our imaginary prisoners have only one opportunity to confess, most firms set output and price over and over again, continually observing their competitors' behavior and adjusting their own accordingly. This allows firms to develop reputations from which trust can arise. As a result, oligopolistic coordination and cooperation can sometimes prevail.

Take, for example, an industry made up of three or four firms that have coexisted for a long time. Over the years, the managers of those firms might grow tired of losing money because of price wars, and an implicit understanding might arise by which all the firms maintain high prices and no firm tries to take market share from its competitors. Although each firm might be tempted to

[9] This payoff matrix assumes that Unilever and Kao both charge the same price. Entries represent profits in thousands of dollars per month.

[10] If P&G and Kao both charged $1.50 and *only* Unilever undercut and charged $1.40, Unilever would make $29,000 per month. It is especially profitable to be the only firm charging the low price.

undercut its competitors, the managers know that the resulting gains will be short lived: Competitors will retaliate, and the result will be renewed warfare and lower profits over the long run.

This resolution of the prisoners' dilemma occurs in some industries, but not in others. Sometimes managers are not content with the moderately high profits resulting from implicit collusion and prefer to compete aggressively in order to increase market share. Sometimes implicit understandings are difficult to reach. For example, firms with different costs and different assessments of market demand might disagree about the "correct" collusive price. Firm *A* might think the "correct" price is $10, while Firm *B* thinks it is $9. When it sets a $9 price, Firm *A* might view this as an attempt to undercut and retaliate by lowering its price to $8. The result is a price war.

In many industries, therefore, implicit collusion is short lived. There is often a fundamental layer of mistrust, so warfare erupts as soon as one firm is perceived by its competitors to be "rocking the boat" by changing its price or increasing advertising.

Price Rigidity

price rigidity Characteristic of oligopolistic markets by which firms are reluctant to change prices even if costs or demands change.

Because implicit collusion tends to be fragile, oligopolistic firms often have a strong desire for stability, particularly with respect to price. This is why **price rigidity** can be a characteristic of oligopolistic industries. Even if costs or demand change, firms are reluctant to change price. If costs fall or market demand declines, they fear that lower prices might send the wrong message to their competitors and set off a round of price warfare. And if costs or demand rises, they are reluctant to raise prices because they are afraid that their competitors may refuse to raise theirs.

kinked demand curve model Oligopoly model in which each firm faces a demand curve kinked at the currently prevailing price: at higher prices demand is very elastic, whereas at lower prices it is inelastic.

This price rigidity is the basis of the **kinked demand curve** model of oligopoly. According to this model, each firm faces a demand curve kinked at the currently prevailing price P^*. (See Figure 12.7.) At prices above P^*, the demand curve is very elastic. The reason is that the firm believes that if it raises its price above P^*, other firms will not follow suit, and it will therefore lose sales and much of its market share. On the other hand, the firm believes that if it lowers its price below P^*, other firms will follow suit because they will not want to lose *their* shares of the market. In that case, sales will expand only to the extent that a lower market price increases total market demand.

Because the firm's demand curve is kinked, its marginal revenue curve is discontinuous. (The bottom part of the marginal revenue curve corresponds to the less elastic part of the demand curve, as shown by the solid portions of each curve.) As a result, the firm's costs can change without resulting in a change in price. As shown in the figure, marginal cost could increase but still equal marginal revenue at the same output level, so that price stays the same.

Although the kinked demand curve model is attractively simple, it does not really explain oligopolistic pricing. It says nothing about how firms arrived at price P^* in the first place, and why they didn't arrive at some different price. It is useful mainly as a *description* of price rigidity rather than as an *explanation* of it.[11] The explanation for price rigidity comes from the prisoners' dilemma and from firms' desires to avoid mutually destructive price competition.

[11] In addition, the model has not stood up well to empirical tests; there is evidence that rival firms do match price increases as well as decreases.

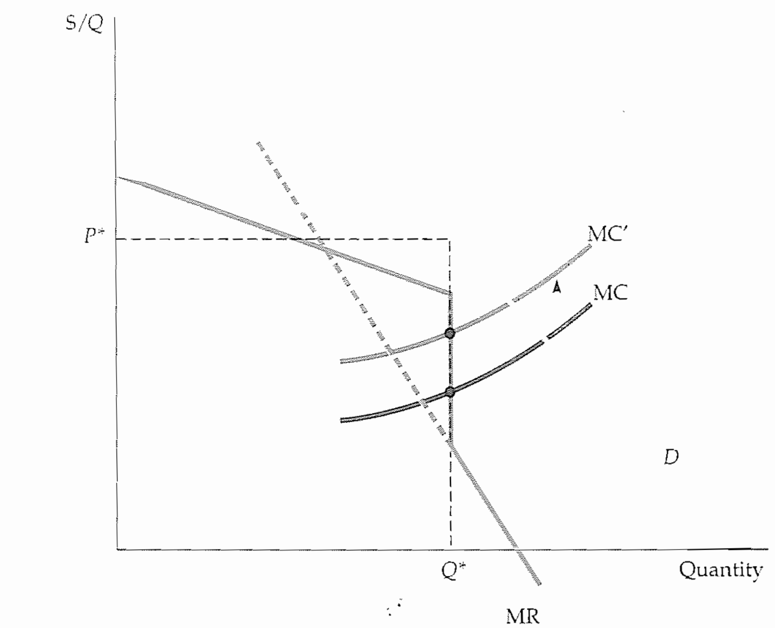

FIGURE 12.7 The Kinked Demand Curve

Each firm believes that if it raises its price above the current price P^*, none of its competitors will follow suit, so it will lose most of its sales. Each firm also believes that if it lowers price, everyone will follow suit, and its sales will increase only to the extent that market demand increases. As a result, the firm's demand curve *D* is kinked at price P^*, and its marginal revenue curve MR is discontinuous at that point. If marginal cost increases from MC to MC′, the firm will still produce the same output level Q^* and charge the same price P^*.

Price Signaling and Price Leadership

One of the main impediments to implicitly collusive pricing is the fact that it is difficult for firms to agree (without talking to each other) on what the price should be. Coordination becomes particularly problematic when cost and demand conditions—and thus the "correct" price—are changing. **Price signaling** is a form of implicit collusion that sometimes gets around this problem. For example, a firm might announce that it has raised its price (perhaps through a press release) and hope that its competitors will take this announcement as a signal that they should also raise prices. If competitors follow suit, all of the firms (in at least the short run) will earn higher profits.

price signaling Form of implicit collusion in which a firm announces a price increase in the hope that other firms will follow suit.

Sometimes a pattern is established whereby one firm regularly announces price changes and other firms in the industry follow suit. This pattern is called **price leadership:** One firm is implicitly recognized as the "leader," while the other firms, the "price followers," match its prices. This behavior solves the problem of coordinating price: Everyone simply charges what the leader is charging.

price leadership Pattern of pricing in which one firm regularly announces price changes that other firms then match.

Suppose, for example, that three oligopolistic firms are currently charging $10 for their product. (If they all know the market demand curve, this might be the Nash equilibrium price.) Suppose that by colluding, they could all set a price of $20 and greatly increase their profits. Meeting and agreeing to set a price of $20 is illegal. But suppose instead that Firm *A* raises its price to $15, and announces to the business press that it is doing so because higher prices are needed to restore economic vitality to the industry. Firms *B* and *C* might view this as a clear message—namely, that Firm *A* is seeking their cooperation in raising prices. They might then raise their own prices to $15. Firm *A* might then increase price further—say, to $18—and Firms *B* and *C* might raise their prices as well. Whether or not the profit-maximizing price of $20 is reached (or surpassed), a pattern of coordination and implicit collusion has now been established that, from the firm's point of view, may be nearly as effective as meeting and formally agreeing on a price.[12]

This example of signaling and price leadership is extreme and might lead to an antitrust lawsuit. But in some industries, a large firm might naturally emerge as a leader, with the other firms deciding that they are best off just matching the leader's prices, rather than trying to undercut the leader or each other. An example is the U.S. automobile industry, where General Motors has traditionally been the price leader.

Price leadership can also serve as a way for oligopolistic firms to deal with the reluctance to change prices, a reluctance that arises out of the fear of being undercut or "rocking the boat." As cost and demand conditions change, firms may find it increasingly necessary to change prices that have remained rigid for some time. In that case, they might look to a price leader to signal when and by how much price should change. Sometimes a large firm will naturally act as leader; sometimes different firms will act as leader from time to time. The example that follows illustrates this.

EXAMPLE 12.4 Price Leadership and Price Rigidity in Commercial Banking

Commercial banks borrow money from individuals and companies who deposit funds in checking accounts, savings accounts, and certificates of deposit. They then use this money to make loans to household and corporate borrowers. By lending at an interest rate higher than the rate they pay on their deposits, they earn a profit.

The largest commercial banks in the United States—BankAmerica, Chase Manhattan, Citicorp, and First Chicago Corp, among others—compete with each other to make loans to large corporate clients. The main form of competition is over price—in this case the interest rates that banks charge corporate clients. If competition becomes aggressive, the interest rates fall, and so do profits. The incentive to avoid aggressive competition leads to price rigidity, and a form of price leadership.

The interest rate that banks charge large corporate clients is called the *prime rate*. Because it is widely cited in newspapers, it is a convenient focal point for price leadership. Most large banks charge the same or nearly the same prime rate; they avoid making frequent changes in the rate that might be destabilizing and lead to competitive warfare. The prime rate changes only when money market conditions cause other interest rates to rise or fall substantially. When

[12] For a formal model of how such price leadership can facilitate collusion, see Julio J. Rotemberg and Garth Saloner, "Collusive Price Leadership," *Journal of Industrial Economics*, 1990.

TABLE 12.6 The Prime Rate

DATE	BANK	RATE CHANGE
March 23, 1994	Major commercial banks	$6 \rightarrow 6\frac{1}{4}$
April 18, 1994	Banc One, Citicorp, Chemical Bank, Bank of New York	$6\frac{1}{4} \rightarrow 6\frac{3}{4}$
May 17, 1994	Citicorp, First Chicago, Bank of New York	$6\frac{3}{4} \rightarrow 7\frac{1}{4}$
August 16, 1994	Citicorp, BankAmerica, Chemical Bank, Chase Manhattan, Norwest	$7\frac{1}{4} \rightarrow 7\frac{3}{4}$
November 15, 1994	First Chicago	$7\frac{3}{4} \rightarrow 8\frac{1}{2}$
February 1, 1995	Major commercial banks	$8\frac{1}{2} \rightarrow 9$
July 6, 1995	Banc One, Bank of America	$9 \rightarrow 8\frac{3}{4}$
December 20, 1995	Banc One	$8\frac{3}{4} \rightarrow 8\frac{1}{2}$
January 31, 1996	Citicorp, NationsBank, Chase Manhattan	$8\frac{1}{2} \rightarrow 8\frac{1}{4}$
March 25, 1997	Banc One, KeyCorp, Norwest	$8\frac{1}{4} \rightarrow 8\frac{1}{2}$
September 30, 1998	Norwest, U.S. Bank of Nebraska, First Chicago	$8\frac{1}{2} \rightarrow 8\frac{1}{4}$
October 15, 1998	Banc One Corp., First Chicago	$8\frac{1}{4} \rightarrow 8$
November 18, 1998	KeyCorp, TCF Bank	$8 \rightarrow 7\frac{3}{4}$
June 30, 1999	Fleet Bank, Bank of America, KeyCorp, Wells Fargo Bank	$7\frac{3}{4} \rightarrow 8$

that happens, one of the major banks announces a change in its rate and other banks quickly follow suit. Different banks act as leader from time to time, but when one bank announces a change, the others follow within two or three days.

Table 12.6 shows the evolution of the prime rate from March 1994 through June 1999. Note that when the prime rate changed, several large banks raised or lowered their rates at about the same time, and other banks quickly followed suit. In most cases, all banks changed their rates within the same day. The table also shows that changes in the prime rate were relatively infrequent. Other market interest rates were fluctuating considerably during this period, but the prime rate changed only after other rates had changed substantially. Figure 12.8 shows this pattern by comparing the prime rate with the interest rate on high-grade (AAA) long-term corporate bonds during the

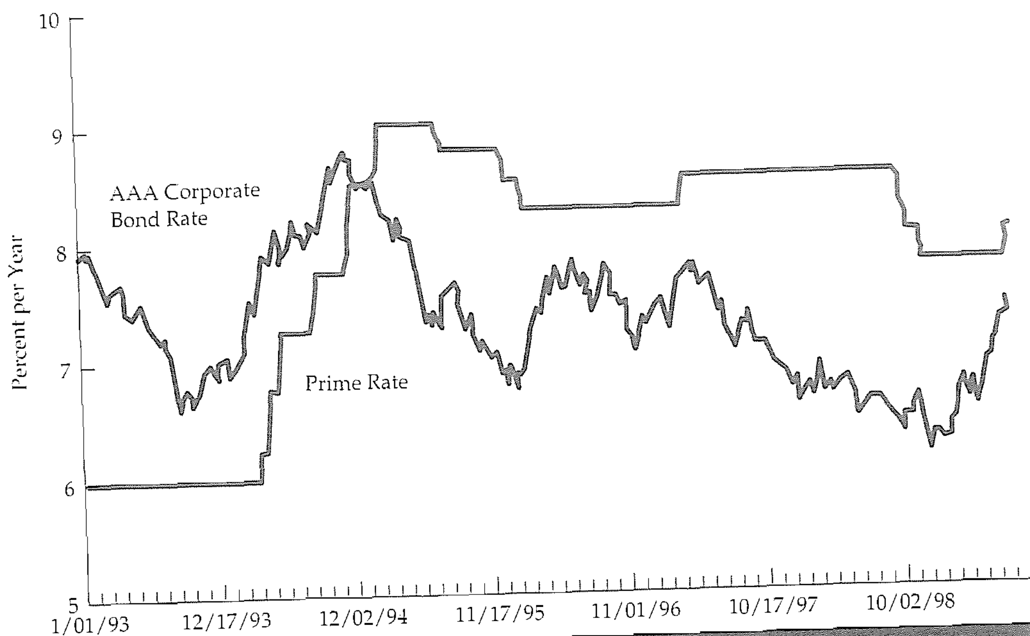

FIGURE 12.8 Prime Rate versus Corporate Bond Rate

The prime rate is the rate that major banks charge large corporate customers for short-term loans. It changes only infrequently because banks are reluctant to undercut one another. When a change does occur, it begins with one bank, and other banks quickly follow suit. The corporate bond rate is the return on long-term corporate bonds. Because these bonds are widely traded, this rate fluctuates with market conditions.

same period. Observe that although the corporate bond rate fluctuated continually, there were extended periods during which prime rate did not change.

The Dominant Firm Model

dominant firm Firm with a large share of total sales that sets price to maximize profits, taking into account the supply response of smaller firms.

In some oligopolistic markets, one large firm has a major share of total sales while a group of smaller firms supplies the remainder of the market. The large firm might then act as a **dominant firm,** setting a price that maximizes its own profits. The other firms, which individually could have little influence over price anyway, would then act as perfect competitors; they take the price set by the dominant firm as given and produce accordingly. But what price should the dominant firm set? To maximize profit, it must take into account how the output of the other firms depends on the price it sets.

Figure 12.9 shows how a dominant firm sets its price. Here, D is the market demand curve, and S_F is the supply curve (i.e., the aggregate marginal cost curve of the smaller fringe firms). The dominant firm must determine *its* demand curve D_D. As the figure shows, this curve is just the difference between market demand and the supply of fringe firms. For example, at price P_1 the supply of fringe firms is just equal to market demand; thus the dominant firm can sell nothing at this price. At a price P_2 or less, fringe firms will not supply any of the good, so the dominant firm faces the market demand curve. At prices between P_1 and P_2, the dominant firm faces the demand curve D_D.

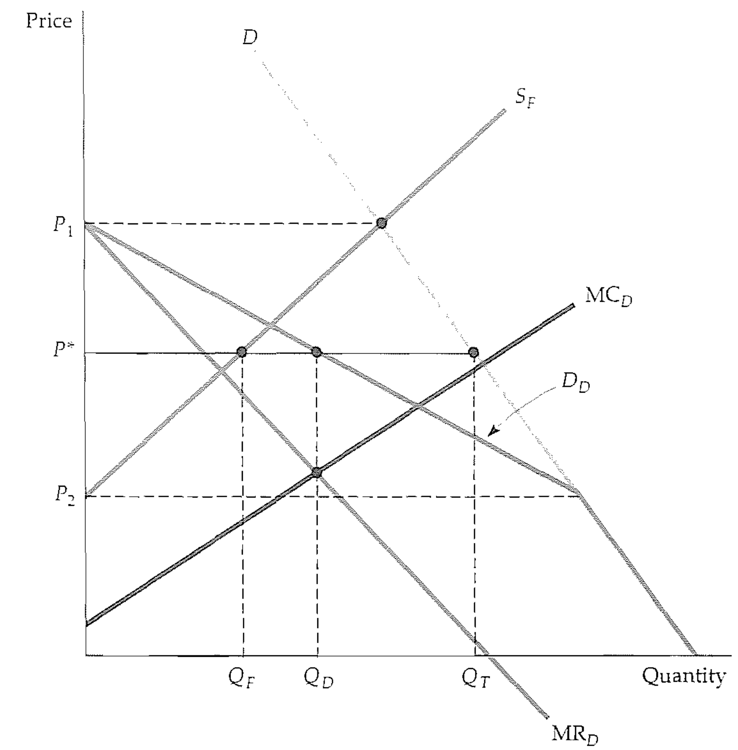

FIGURE 12.9 Price Setting by a Dominant Firm

The dominant firm sets price, and the other firms sell as much as they want at that price. The dominant firm's demand curve, D_D, is the difference between market demand D and the supply of fringe firms S_F. The dominant firm produces a quantity Q_D at the point where its marginal revenue MR_D is equal to its marginal cost MC_D. The corresponding price is P^*. At this price, fringe firms sell Q_F, so that total sales is Q_T.

Corresponding to D_D is the dominant firm's marginal revenue curve MR_D. MC_D is the dominant firm's marginal cost curve. To maximize its profit, the dominant firm produces quantity Q_D at the intersection of MR_D and MC_D. From the demand curve D_D, we find price P^*. At this price, fringe firms sell a quantity Q_F; thus the total quantity sold is $Q_T = Q_D + Q_F$.

12.6 Cartels

Producers in a *cartel* explicitly agree to cooperate in setting prices and output levels. Not all the producers in an industry need to join the cartel, and most cartels involve only a subset of producers. But if enough producers adhere to the cartel's agreements, and if market demand is sufficiently inelastic, the cartel may drive prices well above competitive levels.

Cartels are often international. While U.S. antitrust laws prohibit American companies from colluding, those of other countries are much weaker and are sometimes poorly enforced. Furthermore, nothing prevents countries, or companies owned or controlled by foreign governments, from forming cartels. For example, the OPEC cartel is an international agreement among oil-producing countries which, for over a decade, succeeded in raising world oil prices far above competitive levels.

Other international cartels have also succeeded in raising prices. During the mid-1970s, for example, the International Bauxite Association (IBA) quadrupled bauxite prices, and a secretive international uranium cartel pushed up uranium prices. Some cartels had longer successes: From 1928 through the early 1970s, a cartel called Mercurio Europeo kept the price of mercury close to monopoly levels, and an international cartel monopolized the iodine market from 1878 through 1939. However, most cartels have failed to raise prices. An international copper cartel operates to this day, but it has never had a significant impact on copper prices. Cartel attempts to drive up the prices of tin, coffee, tea, and cocoa have also failed.[13]

Conditions for Cartel Success Why do some cartels succeed while others fail? There are two conditions for cartel success. First, a stable cartel organization must be formed whose members agree on price and production levels and then adhere to that agreement. Unlike our prisoners in the prisoners' dilemma, cartel members can talk to each other to formalize an agreement. This does not mean, however, that agreeing is easy. Different members may have different costs, different assessments of market demand, and even different objectives, and they may therefore want to set price at different levels. Furthermore, each member of the cartel will be tempted to "cheat" by lowering its price slightly to capture a larger market share than it was allotted. Most often, only the threat of a long-term return to competitive prices deters cheating of this sort. But if the profits from cartelization are large enough, that threat may be sufficient.

Recall from §10.2 that monopoly power refers to market power on the part of a seller—the ability of a firm to price its product above its marginal cost of production.

The second condition is the potential for monopoly power. Even if a cartel can solve its organizational problems, there will be little room to raise price if it faces a highly elastic demand curve. Potential monopoly power may be the most important condition for success; if the potential gains from cooperation are large, cartel members will have more incentive to solve their organizational problems.

Analysis of Cartel Pricing

Only rarely do *all* the producers of a good combine to form a cartel. A cartel usually accounts for only a portion of total production and must take into account the supply response of competitive (noncartel) producers when it sets price. Cartel pricing can thus be analyzed by using the dominant firm model discussed earlier. We will apply this model to two cartels, the OPEC oil cartel and the CIPEC copper cartel.[14] This will help us understand why OPEC was successful in raising price while CIPEC was not.

Analyzing OPEC Figure 12.10 illustrates the case of OPEC. Total Demand TD is the total world demand curve for crude oil, and S_c is the competitive (non-OPEC) supply curve. The demand for OPEC oil D_{OPEC} is the difference between

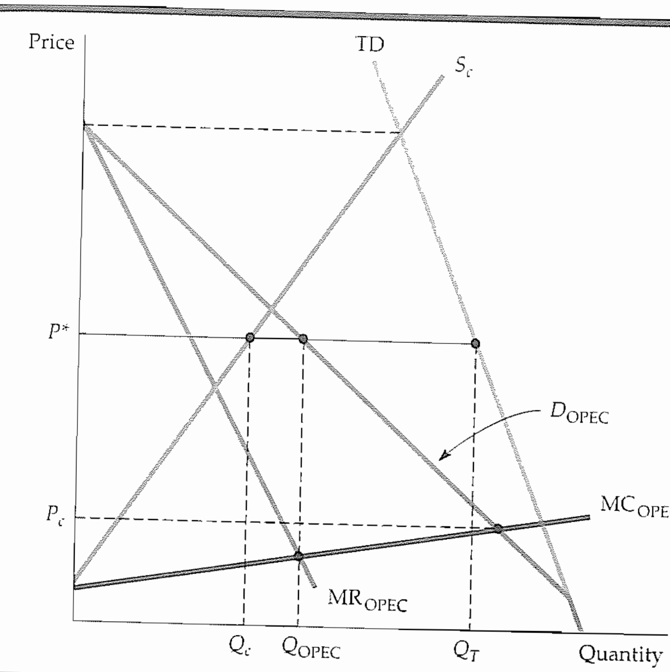

FIGURE 12.10 The OPEC Oil Cartel

TD is the total world demand curve for oil, and S_c is the competitive (non-OPEC) supply curve. OPEC's demand D_{OPEC} is the difference between the two. Because both total demand and competitive supply are inelastic, OPEC's demand is inelastic. OPEC's profit-maximizing quantity Q_{OPEC} is found at the intersection of its marginal revenue and marginal cost curves; at this quantity, OPEC charges price P^*. If OPEC producers had not cartelized, price would be P_c, where OPEC's demand and marginal cost curves intersect.

total demand and competitive supply, and MR_{OPEC} is the corresponding marginal revenue curve. MC_{OPEC} is OPEC's marginal cost curve; as you can see, OPEC has much lower production costs than do non-OPEC producers. OPEC's marginal revenue and marginal cost are equal at quantity Q_{OPEC}, which is the quantity that OPEC will produce. We see from OPEC's demand curve that the price will be P^*, at which competitive supply is Q_c.

Suppose petroleum-exporting countries had not formed a cartel but had instead produced competitively. Price would then have equaled marginal cost. We can therefore determine the competitive price from the point where OPEC's demand curve intersects its marginal cost curve. That price, labeled P_c, is much lower than the cartel price P^*. Because both total demand and non-OPEC supply are inelastic, the demand for OPEC oil is also fairly inelastic; thus the cartel has substantial monopoly power. In the 1970s, it used that power to drive prices well above competitive levels.

In Chapter 2 we stressed the importance of distinguishing between short-run and long-run supply and demand. That distinction is important here. The total demand and non-OPEC supply curves in Figure 12.10 apply to a short- or intermediate-run analysis. In the long run, both demand and supply will be much more elastic, which means that OPEC's demand curve will also be much more elastic. We would thus expect that in the long run, OPEC would be unable

[13] See Jeffrey K. MacKie-Mason and Robert S. Pindyck, "Cartel Theory and Cartel Experience in International Minerals Markets," in *Energy: Markets and Regulation* (Cambridge, MA: MIT Press, 1986).

[14] CIPEC is the French acronym for International Council of Copper Exporting Countries.

to maintain a price that is so much above the competitive level. Indeed, during 1982–1989, oil prices fell in real terms, largely because of the long-run adjustment of demand and non-OPEC supply.

Analyzing CIPEC Figure 12.11 provides a similar analysis of CIPEC. CIPEC consists of four copper-producing countries: Chile, Peru, Zambia, and Congo (formerly Zaire), that collectively account for less than half of world copper production. In these countries, production costs are lower than those of non-CIPEC producers, but except for Chile, not much lower. In Figure 12.11, CIPEC's marginal cost curve is therefore drawn only a little below the non-CIPEC supply curve. CIPEC's demand curve D_{CIPEC} is the difference between total demand TD and non-CIPEC supply S_c. CIPEC's marginal cost and marginal revenue curves intersect at quantity Q_{CIPEC}, with the corresponding price P^*. Again, the competitive price P_c is found at the point where CIPEC's demand curve intersects its marginal cost curve. Note that this price is very close to the cartel price P^*.

Why can't CIPEC increase copper prices much? As Figure 12.11 shows, the total demand for copper is more elastic than that for oil. (Other materials, such as aluminum, can easily be substituted for copper.) Also, competitive supply is much more elastic. Even in the short run, non-CIPEC producers can easily expand supply if prices should rise (in part because of the availability of supply from scrap metal). Thus CIPEC's potential monopoly power is small.

As the examples of OPEC and CIPEC illustrate, successful cartelization requires two things. First, the total demand for the good must not be very price elastic. Second, either the cartel must control nearly all the world's supply or, if it does not, the supply of noncartel producers must not be price elastic. Most international commodity cartels have failed because few world markets meet both conditions.

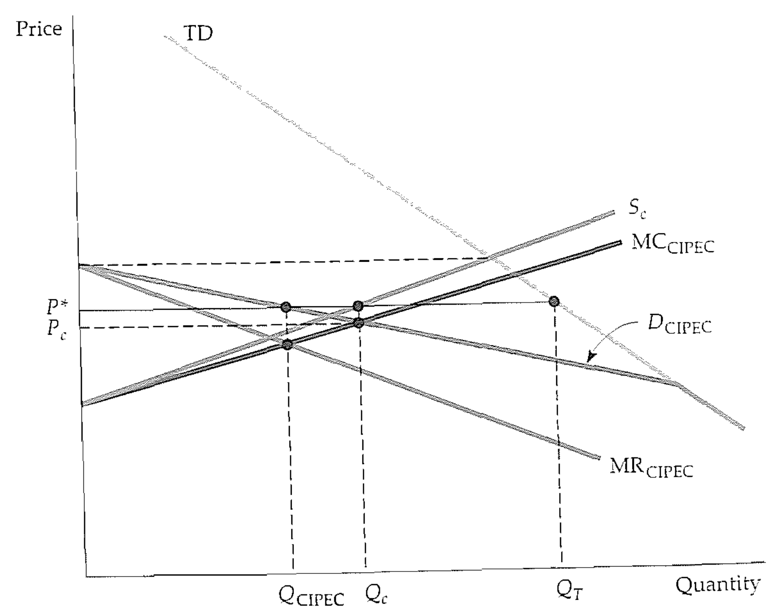

FIGURE 12.11 The CIPEC Copper Cartel

TD is the total demand for copper and S_c is the competitive (non-CIPEC) supply. CIPEC's demand D_{CIPEC} is the difference between the two. Both total demand and competitive supply are relatively elastic, so CIPEC's demand curve is elastic, and CIPEC has very little monopoly power. Note that CIPEC's optimal price P^* is close to the competitive price P_c.

EXAMPLE 12.5 The Cartelization of Intercollegiate Athletics

Many people think of intercollegiate athletics as an extracurricular activity for college students and a diversion for fans. They assume that universities support athletics because it not only gives amateur athletes a chance to develop their skills and play football or basketball before large audiences but also provides entertainment and promotes school spirit and alumni support. Although it does these things, intercollegiate athletics is also a big—and an extremely profitable—industry.

Like any industry, intercollegiate athletics has firms and consumers. The "firms" are the universities that support and finance teams. The inputs to production are the coaches, student athletes, and capital in the form of stadiums and playing fields. The consumers, many of whom are current or former college students, are the fans who buy tickets to games and the TV and radio networks that pay to broadcast them. There are many firms and consumers, which suggests that the industry is competitive. But the persistently high level of profits in this industry is inconsistent with competition—a large state university can regularly earn more than $6 million a year in profits from football games alone.[15] This profitability is the result of monopoly power, obtained via cartelization.

The cartel organization is the National Collegiate Athletic Association (NCAA). The NCAA restricts competition in a number of important activities. To reduce bargaining power by student athletes, the NCAA creates and enforces rules regarding eligibility and terms of compensation. To reduce competition by universities, it limits the number of games that can be played each season and the number of teams that can participate in each division. And to limit price competition, the NCAA has, until 1984, been the sole negotiator for all football television contracts, thereby monopolizing one of the main sources of industry revenues.[16]

Has the NCAA been a successful cartel? Like most cartels, its members have occasionally broken its rules and regulations. But until 1984, it had increased the monopoly power of this industry well above what it would have been otherwise. In 1984, however, the Supreme Court ruled that the NCAA's monopolization of football television contracts was illegal and that individual universities could negotiate their own contracts. The ensuing competition led to a drop in contract fees. As a result, more college football is shown on television but, because of the lower fees, the revenues to the schools have dropped somewhat. But although the Supreme Court's ruling reduced the NCAA's monopoly power, it did not eliminate it. Thanks to the cartel, intercollegiate athletics remains very profitable.

[15] See "In Big-Time College Athletics, the Real Score Is in Dollars," *New York Times*, March 1, 1987.

[16] See James V. Koch, "The Intercollegiate Athletics Industry," in Walter Adams, *The Structure of American Industry*, 7th ed. (New York: Macmillan, 1986). Koch provides a detailed and informative discussion of the nature of this industry and the behavior of the NCAA cartel.

EXAMPLE 12.6 The Milk Cartel

The U.S. government has supported the price of milk since the Great Depression and continues to do so today. The government, however, has been scaling back price supports during the 1990s, and as a result, wholesale prices of milk have fluctuated more widely. Not surprisingly, farmers have been complaining.

In response to these complaints, the federal government has allowed milk producers in the six New England states to cartelize. The cartel—called the Northeast Interstate Dairy Compact—sets a minimum wholesale price for milk, and is exempt from the antitrust laws.

In 1999, Congress responded to the lobbying efforts of farmers in other states by attempting to expand the milk cartel. Legislation was introduced that would allow dairy farmers in New York, New Jersey, Maryland, Delaware, and Pennsylvania to join the New England states and thereby form a cartel covering most of the northeast United States.[17]

Studies have suggested that the original cartel (covering only the New England states) has caused retail prices of milk to rise by only a few cents a gallon. Why so little? The reason is that the New England cartel is surrounded by a fringe of noncartel producers—namely, dairy farmers in New York, New Jersey, and other states. Expanding the cartel, however, will shrink the competitive fringe. This action is likely to enable the cartel to have a greater impact on milk prices.

Not wanting to be left out, dairy farmers in the South also lobbied Congress for higher milk prices. As a result, the 1999 legislation also authorizes 16 Southern states, including Texas, Florida, and Georgia, to create their own regional cartel. Unfortunately for consumers in the South, this can only result in higher milk prices.

SUMMARY

1. In a monopolistically competitive market, firms compete by selling differentiated products, which are highly substitutable. New firms can enter or exit easily. Firms have only a small amount of monopoly power. In the long run, entry will occur until profits are driven to zero. Firms then produce with excess capacity (i.e., at output levels below those that minimize average cost).
2. In an oligopolistic market, only a few firms account for most or all of production. Barriers to entry allow some firms to earn substantial profits, even over the long run. Economic decisions involve strategic considerations—each firm must consider how its actions will affect its rivals, and how they are likely to react.
3. In the Cournot model of oligopoly, firms make their output decisions at the same time, each taking the other's output as fixed. In equilibrium, each firm is maximizing its profit, given the output of its competitor, so no firm has an incentive to change its output. The firms are therefore in a Nash equilibrium. Each firm's profit is higher than it would be under perfect competition but less than what it would earn by colluding.
4. In the Stackelberg model, one firm sets its output first. That firm has a strategic advantage and earns a higher profit. It knows it can choose a large output and that its competitors will have to choose small outputs if they want to maximize profits.
5. The Nash equilibrium concept can also be applied to markets in which firms produce substitute goods and compete by setting price. In equilibrium, each firm maximizes its profit, given the prices of its competitors, and so has no incentive to change price.
6. Firms would earn higher profits by collusively agreeing to raise prices, but the antitrust laws usually prohibit this. They might all set a high price without colluding, each hoping its competitors will do the same, but they are in a prisoners' dilemma, which makes this unlikely. Each firm has an incentive to cheat by lowering its price and capturing sales from competitors.
7. The prisoners' dilemma creates price rigidity in oligopolistic markets. Firms are reluctant to change prices for fear of setting off price warfare.
8. Price leadership is a form of implicit collusion that sometimes gets around the prisoners' dilemma. One firm sets price and other firms follow suit.
9. In a cartel, producers explicitly collude in setting prices and output levels. Successful cartelization requires that the total demand not be very price elastic, and that either the cartel control most supply or else the supply of noncartel producers be inelastic.

[17] "Congress Weighs an Expanded Milk Cartel That Would Aid Farmers by Raising Prices," *New York Times*, May 2, 1999. At the time this book went to press, Congress had not yet passed this legislation. For an update, go to the following Web site: www.dairycompact.org.

QUESTIONS FOR REVIEW

1. What are the characteristics of a monopolistically competitive market? What happens to the equilibrium price and quantity in such a market if one firm introduces a new, improved product?
2. Why is the firm's demand curve flatter than the total market demand curve in monopolistic competition? Suppose a monopolistically competitive firm is making a profit in the short run. What will happen to its demand curve in the long run?
3. Some experts have argued that too many brands of breakfast cereal are on the market. Give an argument to support this view. Give an argument against it.
4. Why is the Cournot equilibrium stable (i.e., why don't firms have any incentive to change their output levels once in equilibrium)? Even if they can't collude, why don't firms set their outputs at the joint profit-maximizing levels (i.e., the levels they would have chosen had they colluded)?
5. In the Stackelberg model, the firm that sets output first has an advantage. Explain why.
6. Explain the meaning of a Nash equilibrium when firms are competing with respect to price. Why is the equilibrium stable? Why don't the firms raise prices to the level that maximizes joint profits?
7. The kinked demand curve describes price rigidity. Explain how the model works. What are its limitations? Why does price rigidity arise in oligopolistic markets?
8. Why does price leadership sometimes evolve in oligopolistic markets? Explain how the price leader determines a profit-maximizing price.
9. Why has the OPEC oil cartel succeeded in raising prices substantially while the CIPEC copper cartel has not? What conditions are necessary for successful cartelization? What organizational problems must a cartel overcome?

EXERCISES

1. Suppose all firms in a monopolistically competitive industry were merged into one large firm. Would that new firm produce as many different brands? Would it produce only a single brand? Explain.
2. Consider two firms facing the demand curve $P = 10 - Q$, where $Q = Q_1 + Q_2$. The firms' cost functions are $C_1(Q_1) = 4 + 2Q_1$ and $C_2(Q_2) = 3 + 3Q_2$.
   a. Suppose both firms have entered the industry. What is the joint profit-maximizing level of output? How much will each firm produce? How would your answer change if the firms have not yet entered the industry?
   b. What is each firm's equilibrium output and profit if they behave noncooperatively? Use the Cournot model. Draw the firms' reaction curves and show the equilibrium.
   c. How much should Firm 1 be willing to pay to purchase Firm 2 if collusion is illegal but the takeover is not?
3. A monopolist can produce at a constant average (and marginal) cost of $AC = MC = 5$. It faces a market demand curve given by $Q = 53 - P$.
   a. Calculate the profit-maximizing price and quantity for this monopolist. Also calculate its profits.

b. Suppose a second firm enters the market. Let Q_1 be the output of the first firm and Q_2 be the output of the second. Market demand is now given by

$$Q_1 + Q_2 = 53 - P$$

Assuming that this second firm has the same costs as the first, write the profits of each firm as functions of Q_1 and Q_2.

c. Suppose (as in the Cournot model) that each firm chooses its profit-maximizing level of output on the assumption that its competitor's output is fixed. Find each firm's "reaction curve" (i.e., the rule that gives its desired output in terms of its competitor's output).

d. Calculate the Cournot equilibrium (i.e., the values of Q_1 and Q_2 for which both firms are doing as well as they can given their competitor's output). What are the resulting market price and profits of each firm?

*e. Suppose there are N firms in the industry, all with the same constant marginal cost, $MC = 5$. Find the Cournot equilibrium. How much will each firm produce, what will be the market price, and how much profit will each firm earn? Also, show that as N becomes large, the market price approaches the price that would prevail under perfect competition.

4. This exercise is a continuation of Exercise 3. We return to two firms with the same constant average and marginal cost, $AC = MC = 5$, facing the market demand curve $Q_1 + Q_2 = 53 - P$. Now we will use the Stackelberg model to analyze what will happen if one of the firms makes its output decision before the other.
   a. Suppose Firm 1 is the Stackelberg leader (i.e., makes its output decisions before Firm 2). Find the reaction curves that tell each firm how much to produce in terms of the output of its competitor.
   b. How much will each firm produce, and what will its profit be?

5. Two firms compete in selling identical widgets. They choose their output levels Q_1 and Q_2 simultaneously and face the demand curve

$$P = 30 - Q$$

where $Q = Q_1 + Q_2$. Until recently, both firms had *zero marginal costs*. Recent environmental regulations have increased Firm 2's marginal cost to \$15. Firm 1's marginal cost remains constant at zero. True or false: As a result, the market price will rise to the *monopoly* level.

6. Suppose that two identical firms produce widgets and that they are the only firms in the market. Their costs are given by $C_1 = 30Q_1$ and $C_2 = 30Q_2$, where Q_1 is the output of Firm 1 and Q_2 the output of Firm 2. Price is determined by the following demand curve:

$$P = 150 - Q$$

where $Q = Q_1 + Q_2$.

   a. Find the Cournot-Nash equilibrium. Calculate the profit of each firm at this equilibrium.
   b. Suppose the two firms form a cartel to maximize joint profits. How many widgets will be produced? Calculate each firm's profit.
   c. Suppose Firm 1 were the only firm in the industry. How would market output and Firm 1's profit differ from that found in part (b) above?
   d. Returning to the duopoly of part (b), suppose Firm 1 abides by the agreement but Firm 2 cheats by increasing production. How many widgets will Firm 2 produce? What will be each firm's profits?

7. Suppose that two competing firms, A and B, produce a homogeneous good. Both firms have a marginal cost of $MC = \$50$. Describe what would happen to output and price in each of the following situations if the firms are at (i) Cournot equilibrium, (ii) collusive equilibrium, and (iii) Bertrand equilibrium.
   a. Firm A must increase wages and its MC increases to \$80.
   b. The marginal cost of both firms increases.
   c. The demand curve shifts to the right.

8. Suppose the airline industry consisted of only two firms: American and Texas Air Corp. Let the two firms have identical cost functions, $C(q) = 40q$. Assume the demand curve for the industry is given by $P = 100 - Q$ and that each firm expects the other to behave as a Cournot competitor.
   a. Calculate the Cournot-Nash equilibrium for each firm, assuming that each chooses the output level that maximizes its profits when taking its rival's output as given. What are the profits of each firm?
   b. What would be equilibrium quantity if Texas Air had constant marginal and average costs of 25 and American had constant marginal and average costs of 40?
   c. Assuming that both firms have the original cost function, $C(q) = 40q$, how much should Texas Air be willing to invest to lower its marginal cost from 40 to 25, assuming that American will not follow suit? How much should American be willing to spend to reduce its marginal cost to 25, assuming that Texas Air will have marginal costs of 25 regardless of American's actions?

*9. Demand for light bulbs can be characterized by $Q = 100 - P$, where Q is in millions of boxes of lights sold and P is the price per box. There are two producers of lights, Everglow and Dimlit. They have identical cost functions:

$$C_i = 10Q_i + \tfrac{1}{2}Q_i^2 \; (i = E, D)$$

$$Q = Q_E + Q_D$$

   a. Unable to recognize the potential for collusion, the two firms act as short-run perfect competitors. What are the equilibrium values of Q_E, Q_D, and P? What are each firm's profits?

   b. Top management in both firms is replaced. Each new manager independently recognizes the oligopolistic nature of the light bulb industry and plays Cournot. What are the equilibrium values of Q_E, Q_D, and P? What are each firm's profits?
   c. Suppose the Everglow manager guesses correctly that Dimlit has a Cournot conjectural variation, so Everglow plays Stackelberg. What are the equilibrium values of Q_E, Q_D, and P? What are each firm's profits?
   d. If the managers of the two companies collude, what are the equilibrium values of Q_E, Q_D, and P? What are each firm's profits?

10. Two firms produce luxury sheepskin auto seat covers: Western Where (WW) and B.B.B. Sheep (BBBS). Each firm has a cost function given by

$$C(q) = 20q + q^2$$

The market demand for these seat covers is represented by the inverse demand equation

$$P = 200 - 2Q$$

where $Q = q_1 + q_2$, total output.

   a. If each firm acts to maximize its profits, taking its rival's output as given (i.e., the firms behave as Cournot oligopolists), what will be the equilibrium quantities selected by each firm? What is total output, and what is the market price? What are the profits for each firm?
   b. It occurs to the managers of WW and BBBS that they could do a lot better by colluding. If the two firms collude, what will be the profit-maximizing choice of output? The industry price? The output and the profit for each firm in this case?
   c. The managers of these firms realize that explicit agreements to collude are illegal. Each firm must decide on its own whether to produce the Cournot quantity or the cartel quantity. To aid in making the decision, the manager of WW constructs a payoff matrix like the one below. Fill in each box with the profit of WW and the profit of BBBS. Given this payoff matrix, what output strategy is each firm likely to pursue?

PROFIT PAYOFF MATRIX (WW PROFIT, BBBS PROFIT)		**BBBS** Produce Cournot q	**BBBS** Produce Cartel q
WW	Produce Cournot q		
WW	Produce Cartel q		

   d. Suppose WW can set its output level *before* BBBS does. How much will WW choose to produce in this case? How much will BBBS produce? What is the market price, and what is the profit for each firm? Is WW better off by choosing its output first? Explain why or why not.

*11. Two firms compete by choosing price. Their demand functions are

$$Q_1 = 20 - P_1 + P_2$$

and

$$Q_2 = 20 + P_1 - P_2$$

where P_1 and P_2 are the prices charged by each firm, respectively, and Q_1 and Q_2 are the resulting demands. Note that the demand for each good depends only on the difference in prices; if the two firms colluded and set the same price, they could make that price as high as they wanted, and earn infinite profits. Marginal costs are zero.

   a. Suppose the two firms set their prices at the *same time*. Find the resulting Nash equilibrium. What price will each firm charge, how much will it sell, and what will its profit be? (*Hint:* Maximize the profit of each firm with respect to its price.)
   b. Suppose Firm 1 sets its price *first* and then Firm 2 sets its price. What price will each firm charge, how much will it sell, and what will its profit be?
   c. Suppose you are one of these firms, and there are three ways you could play the game: (i) Both firms set price at the same time. (ii) You set price first. (iii) Your competitor sets price first. If you could choose among these options, which would you prefer? Explain why.

*12. The dominant firm model can help us understand the behavior of some cartels. Let's apply this model to the OPEC oil cartel. We will use isoelastic curves to describe world demand W and noncartel (competitive) supply S. Reasonable numbers for the price elasticities of world demand and noncartel supply are $-1/2$ and $1/2$, respectively. Then, expressing W and S in millions of barrels per day (mb/d), we could write

$$W = 160P^{-1/2}$$

and

$$S = (3\tfrac{1}{3})P^{1/2}$$

Note that OPEC's net demand is $D = W - S$.

   a. Draw the world demand curve W, the non-OPEC supply curve S, OPEC's net demand curve D, and OPEC's marginal revenue curve. For purposes of approximation, assume OPEC's production cost is zero. Indicate OPEC's optimal price, OPEC's optimal production, and non-OPEC production on the diagram. Now, show on the diagram how the various curves will shift and how OPEC's optimal price will change if non-OPEC supply becomes more expensive because reserves of oil start running out.

b. Calculate OPEC's optimal (profit-maximizing) price. (*Hint:* Because OPEC's cost is zero, just write the expression for OPEC revenue and find the price that maximizes it.)

c. Suppose the oil-consuming countries were to unite and form a "buyers' cartel" to gain monopsony power. What can we say, and what can't we say, about the impact this action would have on price?

*13. A lemon-growing cartel consists of four orchards. Their total cost functions are

$$TC_1 = 20 + 5Q_1^2$$

$$TC_2 = 25 + 3Q_2^2$$

$$TC_3 = 15 + 4Q_3^2$$

$$TC_4 = 20 + 6Q_4^2$$

TC is in hundreds of dollars, and Q is in cartons per month picked and shipped.

a. Tabulate total, average, and marginal costs for each firm for output levels between 1 and 5 cartons per month (i.e., for 1, 2, 3, 4, and 5 cartons).

b. If the cartel decided to ship 10 cartons per month and set a price of $25 per carton, how should output be allocated among the firms?

c. At this shipping level, which firm has the most incentive to cheat? Does any firm *not* have an incentive to cheat?

CHAPTER 13

Game Theory and Competitive Strategy

Chapter Outline

List of Examples

In Chapter 12, we began to explore some of the strategic output and pricing decisions that firms must often make. We saw how a firm can take into account the likely responses of its competitors when it makes these decisions. However, there are many questions about market structure and firm behavior that we have not yet addressed. For example, why do firms tend to collude in some markets and to compete aggressively in others? How do some firms manage to deter entry by potential competitors? And how should firms make pricing decisions when demand or cost conditions are changing or new competitors are entering the market?

To answer these questions, we will use game theory to extend our analysis of strategic decision making. The application of game theory has been an important development in microeconomics. This chapter explains some key aspects of this theory and shows how it can be used to understand how markets evolve and operate, and how managers should think about the strategic decisions they continually face. We will see, for example, what happens when oligopolistic firms must set and adjust prices strategically over time, so that the prisoners' dilemma, which we discussed in Chapter 12, is repeated over and over. We will show how firms can make strategic moves that give them advantages over competitors or an edge in bargaining situations. And we will see how firms can use threats, promises, or more concrete actions to deter entry by potential competitors.

13.1 Gaming and Strategic Decisions

First, we should clarify what gaming and strategic decision making are all about. A **game** is any situation in which *players* (the participants) make *strategic decisions*—i.e., decisions that take into account each other's actions and responses. Examples of games include firms competing with each other by setting prices, or a group of consumers bidding against each other at an auction for a work of art. Strategic decisions result in **payoffs** to the players: outcomes that generate

game Situation in which players (participants) make strategic decisions that take into account each other's actions and responses.

payoff Outcome of a game that generates rewards or benefits for the player.

strategy Rule or plan of action for playing a game.

optimal strategy Strategy that maximizes player's expected payoff.

rewards or benefits. For the price-setting firms, the payoffs are profits; for the bidders at the auction, the winner's payoff is her consumer surplus—i.e., the value she places on the artwork less the amount she must pay.

A key objective of game theory is to determine the optimal strategy for each player. A **strategy** is a rule or plan of action for playing the game. For our price-setting firms, a strategy might be: "I'll keep my price high as long as my competitors do the same, but once a competitor lowers his price, I'll lower mine even more." For a bidder at an auction, a strategy might be: "I'll make a first bid of $2000 to convince the other bidders that I'm serious about winning, but I'll drop out if other bidders push the price above $5000." The **optimal strategy** for a player is the one that maximizes her expected payoff.

We will focus on games involving players who are *rational*, in the sense that they think through the consequences of their actions. In essence, we are concerned with the following question: *If I believe that my competitors are rational and act to maximize their own payoffs, how should I take their behavior into account when making my decisions?* Of course in real life you may encounter competitors who are irrational, or are less capable than you of thinking through the consequences of their actions. Nonetheless, a good place to start is by assuming that your competitors are just as rational and just as smart as you are.[1] As we will see, taking competitors' behavior into account is not as simple as it might seem. Determining optimal strategies can be difficult, even under conditions of complete symmetry and perfect information (i.e., my competitors and I have the same cost structure and are fully informed about each others' costs, about demand, etc.). Moreover, we will be concerned with more complex situations in which firms face different costs, different types of information, and various degrees and forms of competitive "advantage" and "disadvantage."

Noncooperative versus Cooperative Games

cooperative game Game in which participants can negotiate binding contracts that allow them to plan joint strategies.

noncooperative game Game in which negotiation and enforcement of binding contracts between players is not possible.

The economic games that firms play can be either *cooperative* or *noncooperative*. In a **cooperative game,** players can negotiate binding contracts that allow them to plan joint strategies. In a **noncooperative game,** negotiation and enforcement of binding contracts are not possible.

An example of a cooperative game is the bargaining between a buyer and a seller over the price of a rug. If the rug costs $100 to produce and the buyer values the rug at $200, a cooperative solution to the game is possible: An agreement to sell the rug at any price between $101 and $199 will maximize the sum of the buyer's consumer surplus and the seller's profit, while making both parties better off. Another cooperative game would involve two firms negotiating a joint investment to develop a new technology (assuming that neither firm would have enough know-how to succeed on its own). If the firms can sign a binding contract to divide the profits from their joint investment, a cooperative outcome that makes both parties better off is possible.[2]

An example of a noncooperative game is a situation in which two competing firms take each other's likely behavior into account when independently setting their prices. Each firm knows that by undercutting its competitor it can capture more market share, but doing so risks setting off a price war. Another noncooperative game is the auction mentioned above; each bidder must take the likely behavior of the other bidders into account when determining an optimal bidding strategy.

[1] When we asked, 80 percent of our students told us that they were smarter and more capable than most of their classmates. We hope that you don't find it too much of a strain to imagine competing against people who are as smart and capable as you are.

[2] Bargaining over a rug is called a *constant sum* game because no matter what the selling price, the sum of consumer surplus and profit will be the same. Negotiating over a joint venture is a *nonconstant sum* game: the total profit that results from the venture will depend on the outcome of the negotiations (for instance, the resources that each firm devotes to the venture).

Note that the fundamental difference between cooperative and noncooperative games lies in the contracting possibilities. In cooperative games, binding contracts are possible; in noncooperative games, they are not.

We will be concerned mostly with noncooperative games. In any game, however, the most important aspect of strategic decision making is *understanding your opponent's point of view, and (assuming your opponent is rational) deducing his or her likely responses to your actions.* This may seem obvious—of course, one must understand an opponent's point of view. Yet even in simple gaming situations, people often ignore or misjudge opponents' positions and the rational responses those positions imply.

How to Buy a Dollar Bill As an example, consider the following game devised by Martin Shubik.[3] A dollar bill is auctioned, but in an unusual way. The highest bidder receives the dollar in return for the amount bid. However, the second-highest bidder must also hand over the amount he or she bid—and get nothing in return. *If you were playing this game, how much would you bid for the dollar bill?*

Classroom experience shows that students often end up bidding more than a dollar for the dollar. In a typical scenario, one player bids 20 cents, and another 30 cents. The lower bidder now stands to lose 20 cents but figures he can earn a dollar by raising his bid, and so bids 40 cents. The escalation continues until two players carry the bidding to a dollar against 90 cents. Now the 90-cent bidder has to choose between bidding $1.10 for the dollar or paying 90 cents to get nothing. Most often, he raises his bid, and the bidding escalates further. In some experiments, the "winning" bidder has ended up paying more than $3 for the dollar!

How could intelligent students put themselves in this position? By failing to think through the likely response of the other players and the sequence of events it implies. How much would you bid for the dollar? We hope nothing.

In the rest of this chapter, we will examine simple games that involve pricing, advertising, and investment decisions. The games are simple in that, *given some behavioral assumptions,* we can determine the best strategy for each firm. But even for these simple games, we will find that the correct behavioral assumptions are not always easy to make, and will depend on how the game is played (e.g., how long the firms stay in business, their reputations, etc.). Therefore, when reading this chapter, you should try to understand the basic issues involved in making strategic decisions. You should also keep in mind the importance of carefully assessing your opponent's position and rational response to your actions, as Example 13.1 illustrates.

EXAMPLE 13.1 Acquiring a Company

You represent Company *A* (the acquirer), which is considering acquiring Company *T* (the target).[4] You plan to offer cash for all of Company *T*'s shares, but you are unsure what price to offer. The complication is this: The

[3] Martin Shubik, *Game Theory in the Social Sciences* (Cambridge, MA: MIT Press, 1982).

[4] This is a revised version of an example designed by Max Bazerman for a course at MIT.

value of Company *T*—indeed, its viability—depends on the outcome of a major oil exploration project. If the project fails, Company *T* under current management will be worth nothing. But if it succeeds, Company *T*'s value under current management could be as high as $100/share. All share values between $0 and $100 are considered equally likely.

It is well known, however, that Company *T* will be worth much more under the progressive management of Company *A* than under current management. In fact, whatever the ultimate value under current management, *Company T will be worth 50 percent more under the management of Company A*. If the project fails, Company *T* is worth $0/share under either management. If the exploration project generates a $50/share value under current management, the value under Company *A* will be $75/share. Similarly, a $100/share value under Company *T* implies a $150/share value under Company *A*, and so on.

You must determine what price Company *A* should offer for Company *T*'s shares. This offer must be made *now—before* the outcome of the exploration project is known. From all indications, Company *T* would be happy to be acquired by Company *A—for the right price*. You expect Company *T* to delay a decision on your bid until the exploration results are in and then accept or reject your offer before news of the drilling results reaches the press.

Thus, *you (Company A) will not know the results of the exploration project when submitting your price offer, but Company T will know the results when deciding whether to accept your offer. Also, Company T will accept any offer by Company A that is greater than the (per share) value of the company under current management.* As the representative of Company *A*, you are considering price offers in the range $0/share (i.e., making no offer at all) to $150/share. *What price per share should you offer for Company T's stock?*

Note: The typical response—to offer between $50 and $75 per share—is wrong. The correct answer to this problem appears at the end of this chapter, but we urge you to try to answer it on your own.

13.2 Dominant Strategies

How can we decide on the best strategy for playing a game? How can we determine a game's likely outcome? We need something to help us determine how the rational behavior of each player will lead to an equilibrium solution. Some strategies may be successful if competitors make certain choices but fail if they make other choices. Other strategies, however, may be successful regardless of what competitors do. We begin with the concept of a **dominant strategy**—*one that is optimal no matter what an opponent does.*

dominant strategy Strategy that is optimal no matter what an opponent does.

The following example illustrates this in a duopoly setting. Suppose Firms *A* and *B* sell competing products and are deciding whether to undertake advertising campaigns. Each firm will be affected by its competitor's decision. The possible outcomes of the game are illustrated by the payoff matrix in Table 13.1. (Recall that the payoff matrix summarizes the possible outcomes of the game; the first number in each cell is the payoff to *A* and the second is the payoff to *B*.) Observe that if both firms decide to advertise, Firm *A* will make a profit of 10 and Firm *B* a profit of 5. If Firm *A* advertises and Firm *B* does not, Firm *A* will earn 15 and Firm *B* zero. The table also shows the outcomes for the other two possibilities.

In §12.5, we explain that a payoff matrix is a table showing the payoffs to each player given her decision and the decision of her competitor.

TABLE 13.1 Payoff Matrix for Advertising Game

		FIRM B	
		Advertise	Don't advertise
FIRM A	Advertise	10, 5	15, 0
	Don't advertise	6, 8	10, 2

What strategy should each firm choose? First consider Firm *A*. It should clearly advertise because no matter what firm *B* does, Firm *A* does best by advertising. If Firm *B* advertises, *A* earns a profit of 10 if it advertises but only 6 if it doesn't. If *B* does not advertise, *A* earns 15 if it advertises but only 10 if it doesn't. Thus advertising is a dominant strategy for Firm *A*. The same is true for Firm *B*; no matter what firm *A* does, Firm *B* does best by advertising. Therefore, assuming that both firms are rational, we know that the outcome for this game is that both *firms will advertise*. This outcome is easy to determine because both firms have dominant strategies.

When every player has a dominant strategy, we call the outcome of the game an **equilibrium in dominant strategies.** Such games are straightforward to analyze because each player's optimal strategy can be determined without worrying about the actions of the other players.

equilibrium in dominant strategies Outcome of a game in which each firm is doing the best it can regardless of what its competitors are doing.

Unfortunately, not every game has a dominant strategy for each player. To see this, let's change our advertising example slightly. The payoff matrix in Table 13.2 is the same as in Table 13.1 except for the bottom right-hand corner—if neither firm advertises, Firm *B* will again earn a profit of 2, but Firm *A* will earn a profit of 20. Perhaps Firm *A*'s ads are largely defensive, designed to refute Firm *B*'s claims, and expensive; by not advertising, Firm *A* can thus reduce its expenses considerably.

Now Firm *A* has no dominant strategy. *Its optimal decision depends on what Firm B does.* If Firm *B* advertises, Firm *A* does best by advertising; but if Firm *B* does not advertise, Firm *A* also does best by not advertising. Now suppose both firms must make their decisions at the same time. What should Firm *A* do?

To answer this, Firm *A* must put itself in Firm *B*'s shoes. What decision is best from Firm *B*'s point of view, and what is Firm *B* likely to do? The answer is clear: Firm *B* has a dominant strategy—advertise, no matter what Firm *A* does. (If Firm *A* advertises, *B* earns 5 by advertising and 0 by not advertising; if *A* doesn't advertise, *B* earns 8 if it advertises and 2 if it doesn't.) Therefore, Firm *A* can conclude that Firm *B* will advertise. This means that Firm *A* should advertise (and thereby earn 10 instead of 6). The equilibrium is that both firms will advertise. It is the logical outcome of the game because Firm *A* is doing the best it can, given Firm *B*'s decision; and Firm *B* is doing the best it can, given Firm *A*'s decision.

TABLE 13.2 Modified Advertising Game

		FIRM B	
		Advertise	Don't advertise
FIRM A	Advertise	10, 5	15, 0
	Don't advertise	6, 8	20, 2

13.3 The Nash Equilibrium Revisited

To determine the likely outcome of a game, we have been seeking "self-enforcing," or "stable" strategies. Dominant strategies are stable, but in many games, one or more players do not have a dominant strategy. We therefore need a more general equilibrium concept. In Chapter 12 we introduced the concept of a *Nash equilibrium* and saw that it is widely applicable and intuitively appealing.[5]

Recall that a Nash equilibrium is a set of strategies (or actions) such that *each player is doing the best it can given the actions of its opponents.* Because each player has no incentive to deviate from its Nash strategy, the strategies are stable. In the example shown in Table 13.2, the Nash equilibrium is that both firms advertise: Given the decision of its competitor, each firm is satisfied that it has made the best decision possible, and so has no incentive to change its decision.

In Chapter 12, we used the Nash equilibrium to study output and pricing by oligopolistic firms. In the Cournot model, for example, each firm sets its own output while taking the outputs of its competitors as fixed. We saw that in a Cournot equilibrium, no firm has an incentive to change its output unilaterally because each firm is doing the best it can given the decisions of its competitors. Thus a Cournot equilibrium is a Nash equilibrium.[6] We also examined models in which firms choose price, taking the prices of their competitors as fixed. Again, in the Nash equilibrium, each firm is earning the largest profit it can given the prices of its competitors, and thus has no incentive to change its price.

In §12.2, we explain that the Cournot equilibrium is a Nash equilibrium in which each firm correctly assumes how much its competitor will produce.

It is helpful to compare the concept of a Nash equilibrium with that of an equilibrium in dominant strategies:

Dominant Strategies:	I'm doing the best I can *no matter what you do.* You're doing the best you can *no matter what I do.*
Nash Equilibrium:	I'm doing the best I can *given what you are doing.* You're doing the best you can *given what I am doing.*

Note that a dominant strategy equilibrium is a special case of a Nash equilibrium.

In the advertising game of Table 13.2, there is a single Nash equilibrium—both firms advertise. In general, a game need not have a single Nash equilibrium. Sometimes there is no Nash equilibrium, and sometimes there are several (i.e., several sets of strategies are stable and self-enforcing). A few more examples will help to clarify this.

The Product Choice Problem Consider the following "product choice" problem. Two breakfast cereal companies face a market in which two new variations of cereal can be successfully introduced—provided that each variation is introduced by only one firm. There is a market for a new "crispy" cereal and for a new "sweet" cereal, but each firm has the resources to introduce only one new product. The payoff matrix for the two firms might look like the one in Table 13.3.

[5] Our discussion of the Nash equilibrium, and of game theory in general, is at an introductory level. For a more in-depth discussion of game theory and its applications, see James W. Friedman, *Game Theory with Applications to Economics* (New York: Oxford University Press, 1990); Drew Fudenberg and Jean Tirole, *Game Theory* (Cambridge, MA: MIT Press, 1991); and Avinash Dixit and Susan Skeath, *Games of Strategy* (New York: Norton, 1999).

[6] A *Stackelberg equilibrium* is also a Nash equilibrium. In the Stackelberg model, however, the rules of the game are different: One firm makes its output decision before its competitor does. Under these rules, each firm is doing the best it can given the decision of its competitor.

TABLE 13.3 Product Choice Problem

		FIRM 2	
		Crispy	Sweet
FIRM 1	Crispy	−5, −5	10, 10
	Sweet	10, 10	−5, −5

In this game, each firm is indifferent about which product it produces—so long as it does not introduce the same product as its competitor. If coordination were possible, the firms would probably agree to divide the market. But what if the firms must behave *noncooperatively?* Suppose that somehow—perhaps through a news release—Firm 1 indicates it is about to introduce the sweet cereal, and Firm 2 (after hearing this) indicates it will introduce the crispy one. Given the action it believes its opponent is taking, neither firm has an incentive to deviate from its proposed action. If it takes the proposed action, its payoff is 10, but if it deviates—and its opponent's action remains unchanged—its payoff will be −5. Therefore, the strategy set given by the bottom left-hand corner of the payoff matrix is stable and constitutes a Nash equilibrium: Given the strategy of its opponent, each firm is doing the best it can and has no incentive to deviate.

Note that the upper right-hand corner of the payoff matrix is also a Nash equilibrium, which might occur if Firm 1 indicated it was about to produce the crispy cereal. Each Nash equilibrium is stable because *once the strategies are chosen,* no player will unilaterally deviate from them. However, without more information, we have no way of knowing *which* equilibrium (crispy/sweet vs. sweet/crispy) is likely to result—or if *either* will result. Of course, both firms have a strong incentive to reach *one* of the two Nash equilibria—if they both introduce the same type of cereal, they will both lose money. The fact that the two firms are not allowed to collude does not mean that they will not reach a Nash equilibrium. As an industry develops, understandings often evolve as firms "signal" each other about the paths the industry is to take.

The Beach Location Game Suppose that you (Y) and a competitor (C) are planning to sell soft drinks on a beach this summer. The beach is 200 yards long, and sunbathers are spread evenly across its length. You and your competitor sell the same soft drinks at the same prices, so customers will walk to the closest vendor. Where on the beach will you locate, and where do you think your competitor will locate?

If you think about this for a minute, you will see that the only Nash equilibrium calls for both you and your competitor to locate at the same spot in the center of the beach (see Figure 13.1). To see why, suppose your competitor located at some other point A, which is three quarters of the way to the end of the beach. In that case you would no longer want to locate in the center; you would locate near your competitor, just to her left. You would thus capture nearly three-fourths of all sales, while your competitor got only the remaining fourth. This outcome is not an equilibrium because your competitor would then want to move to the center of the beach, and you would do the same.

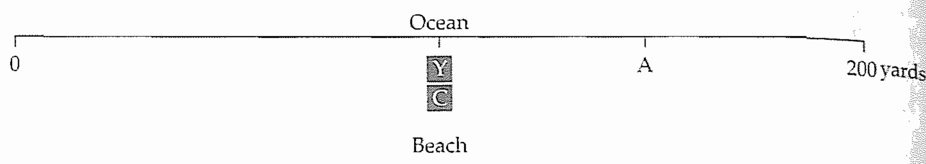

FIGURE 13.1 Beach Location Game

You (Y) and a competitor (C) plan to sell soft drinks on a beach. If sunbathers are spread evenly across the beach and will walk to the closest vendor, the two of you will locate next to each other at the center of the beach. This is the only Nash equilibrium. If your competitor located at point *A*, you would want to move until you were just to her left, where you could capture three-fourths of all sales. But your competitor would then want to move back to the center, and you would do the same.

The "beach location game" can help us understand a variety of phenomena. Have you ever noticed how, along a two- or three-mile stretch of road, two or three gas stations, or several car dealerships, will be located close to each other? Likewise, as a U.S. presidential election approaches, the Democratic and Republican candidates typically move close to the center as they define their political positions.

Maximin Strategies

The concept of a Nash equilibrium relies heavily on individual rationality. Each player's choice of strategy depends not only on its own rationality, but also on that of its opponent. This can be a limitation, as the example in Table 13.4 shows.

In this game, two firms compete in selling file-encryption software. Because both firms use the same encryption standard, files encrypted by one firm's software can be read by the other's—an advantage for consumers. Nonetheless, Firm 1 has a much larger market share (it entered the market earlier and its software has a better user interface). Both firms are now considering an investment in a new encryption standard.

Note that investing is a dominant strategy for Firm 2 because by doing so, it will do better (earning $10 million rather than 0) regardless of what Firm 1 does. Thus Firm 1 should expect Firm 2 to invest. In this case, Firm 1 would do better by also investing (and earning $20 million) than by not investing (and losing $10 million). Clearly the outcome (invest, invest) is a Nash equilibrium for this game, and you can verify that it is the only Nash equilibrium. But note that Firm 1's managers had better be sure that Firm 2's managers understand the game and are rational. If Firm 2 should happen to make a mistake and fail to invest, it would be extremely costly to Firm 1. (Consumer confusion over incompatible standards would arise, and Firm 1, with its dominant market share, would lose $100 million.)

TABLE 13.4 Maximin Strategy

		FIRM 2	
		Don't invest	Invest
FIRM 1	Don't invest	0, 0	− 10, 10
	Invest	− 100, 0	20, 10

If you were Firm 1, what would you do? If you tend to be cautious, and if you are concerned that the managers of Firm 2 might not be fully informed or rational, you might choose to play "don't invest." In that case, the worst that can happen is that you will lose $10 million; you no longer have a chance of losing $100 million. Such a strategy is called a **maximin strategy** because it *maximizes the minimum gain that can be earned.* If both firms used maximin strategies, the outcome would be that Firm 1 does not invest and Firm 2 does. A maximin strategy is conservative, but it is not profit-maximizing (Firm 1, for example, loses $10 million rather than earning $20 million). Note that if Firm 1 *knew for certain* that Firm 2 was using a maximin strategy, it would prefer to invest (and earn $20 million) instead of following its own maximin strategy of not investing.

maximin strategy Strategy that maximizes the minimum gain that can be earned.

Maximizing the Expected Payoff The maximin strategy is conservative. If Firm 1 is unsure about what Firm 2 will do but can assign probabilities to each possible action for Firm 2, it could instead use a strategy that *maximizes its expected payoff.* Suppose, for example, that Firm 1 thinks that there is only a 10-percent chance that Firm 2 will not invest. In that case, Firm 1's expected payoff from investing is (.1)(− 100) + (.9)(20) = $8 million. Its expected payoff if it doesn't invest is (.1)(0) + (.9)(− 10) = − $9 million. In this case, Firm 1 should invest.

On the other hand, suppose Firm 1 thinks that the probability that Firm 2 will not invest is 30 percent. In this case, Firm 1's expected payoff from investing is (.3)(− 100) + (.7)(20) = − $16 million, while its expected payoff from not investing is (.3)(0) + (.7)(− 10) = − $7 million. Thus Firm 1 will choose not to invest.

You can see that Firm 1's strategy depends critically on its assessment of the probabilities of different actions by Firm 2. Determining these probabilities may seem like a tall order. However, firms often face uncertainty (over market conditions, future costs, and the behavior of competitors), and must make the best decisions they can based on probability assessments and expected values.

For a review of expected value, see §5.1, where it is defined as a weighted average of the payoffs associated with all possible outcomes, with the probabilities of each outcome used as weights.

The Prisoners' Dilemma What is the Nash equilibrium for the prisoners' dilemma discussed in Chapter 12? Table 13.5 shows the payoff matrix for the prisoners' dilemma. Recall that the ideal outcome is one in which neither prisoner confesses, so that they both get two years in prison. Confessing, however, is a *dominant strategy* for each prisoner—it yields a higher payoff regardless of the strategy of the other prisoner. Dominant strategies are also maximin strategies. Therefore, the outcome in which both prisoners confess is both a Nash equilibrium and a maximin solution. Thus, in a very strong sense, it is rational for each prisoner to confess.

TABLE 13.5 Prisoners' Dilemma

		PRISONER B	
		Confess	Don't confess
PRISONER A	Confess	− 5, − 5	− 1, − 10
	Don't confess	− 10, − 1	− 2, − 2

*Mixed Strategies

In all of the games that we examined so far, we have considered strategies in which players make a specific choice or take a specific action: advertise or don't advertise, set a price of $4 or a price of $6, and so on. Strategies of this kind are called **pure strategies.** There are games, however, in which pure strategies are not the best way to play.

pure strategy Strategy in which a player makes a specific choice or takes a specific action.

Matching Pennies An example is the game of "Matching Pennies." In this game, each player chooses heads or tails and the two players reveal their coins at the same time. If the coins match (i.e., both are heads or both are tails), Player *A* wins and receives a dollar from Player *B*. If the coins do not match, Player *B* wins and receives a dollar from Player *A*. The payoff matrix is shown in Table 13.6.

Note that there is no Nash equilibrium in pure strategies for this game. Suppose, for example, that Player *A* chose the strategy of playing heads. Then Player *B* would want to play tails. But if Player *B* plays tails, Player *A* would also want to play tails. No combination of heads or tails leaves both players satisfied—one player or the other will always want to change strategies.

Although there is no Nash equilibrium in pure strategies, there is a Nash equilibrium in **mixed strategies:** *strategies in which players make random choices among two or more possible actions, based on sets of chosen probabilities.* In this game, for example, Player *A* might simply flip the coin, thereby playing heads with probability 1/2 and playing tails with probability 1/2. In fact, if Player *A* follows this strategy and Player *B* does the same, we will have a Nash equilibrium; both players will be doing the best they can given what the opponent is doing. Note that the outcome of the game is random, but the *expected payoff* is 0 for each player.

mixed strategy Strategy in which a player makes a random choice among two or more possible actions, based on a set of chosen probabilities.

It may seem strange to play a game by choosing actions randomly. But put yourself in the position of Player *A* and think what would happen if you followed a strategy *other* than just flipping the coin. Suppose, for example, you decided to play heads. If Player *B* knows this, she would play tails and you would lose. Even if Player *B* didn't know your strategy, if the game were played over and over again, she could eventually discern your pattern of play and choose a strategy that countered it. Of course, you would then want to change your strategy—which is why this would not be a Nash equilibrium. Only if you and your opponent both choose heads or tails randomly with probability 1/2 would neither of you have any incentive to change strategies. (You can check that the use of different probabilities, say 3/4 for heads and 1/4 for tails, does not generate a Nash equilibrium.)

TABLE 13.6 Matching Pennies

		PLAYER B	
		Heads	Tails
PLAYER A	Heads	1, −1	−1, 1
	Tails	−1, 1	1, −1

One reason to consider mixed strategies is that some games (such as "Matching Pennies") do not have any Nash equilibria in pure strategies. It can be shown, however, that once we allow for mixed strategies, *every* game has at least one Nash equilibrium.[7] Hence, mixed strategies provide solutions to games when pure strategies fail. Of course, whether solutions involving mixed strategies are reasonable will depend on the particular game and players. Mixed strategies are likely to be very reasonable for "Matching Pennies," poker, and other such games. A firm, on the other hand, might not find it reasonable to believe that its competitor will set its price randomly.

The Battle of the Sexes Some games have Nash equilibria both in pure strategies and in mixed strategies. An example is "The Battle of the Sexes," a game that you might find familiar. It goes like this. Jim and Joan would like to spend Saturday night together but have different tastes in entertainment. Joan would like to go to the opera, but Jim prefers mud wrestling. (Feel free to reverse these preferences.) As the payoff matrix in Table 13.7 shows, Joan would most prefer to go to the opera with Jim, but prefers watching mud wrestling with Jim to going to the opera alone, and similarly for Jim.

First, note that there are two Nash equilibria in pure strategies for this game—the one in which Jim and Joan both watch mud wrestling, and the one in which they both go to the opera. Jim, of course, would prefer the first of these outcomes and Joan the second, but both outcomes are equilibria—neither Jim nor Joan would want to change his or her decision, given the decision of the other.

This game also has an equilibrium in mixed strategies: Jim chooses wrestling with probability 2/3 and opera with probability 1/3, and Joan chooses wrestling with probability 1/3 and opera with probability 2/3. You can check that if Joan uses this strategy, Jim cannot do better with any other strategy, and vice versa.[8] The outcome is random, and Jim and Joan will each have an expected payoff of 2/3.

Should we expect Jim and Joan to use these mixed strategies? Unless they're very risk loving or in some other way a strange couple, probably not. By agreeing to either form of entertainment, each will have a payoff of at least 1, which

TABLE 13.7 The Battle of the Sexes

		JOAN	
		Wrestling	Opera
JIM	Wrestling	2, 1	0, 0
	Opera	0, 0	1, 2

[7] More precisely, every game with a finite number of players and a finite number of actions has at least one Nash equilibrium. For a proof, see David M. Kreps, *A Course in Microeconomic Theory* (Princeton, NJ: Princeton University Press, 1990), p. 409.

[8] Suppose Jim randomizes, letting p be the probability of wrestling, and $(1 - p)$ the probability of opera. Since Joan is using probabilities of 1/3 for wrestling and 2/3 for opera, the probability that both will choose wrestling is $(1/3)p$, and the probability that both will choose opera is $(2/3)(1 - p)$. Hence Jim's expected payoff is $2(1/3)p + 1(2/3)(1 - p) = (2/3)p + 2/3 - (2/3)p = 2/3$. This is independent of p, so Jim cannot do better in terms of expected payoff no matter what he chooses.

exceeds the expected payoff of 2/3 from randomizing. In this game as in many others, mixed strategies provide another solution, but not a very realistic one. Hence, for the remainder of this chapter we will focus on pure strategies.

13.4 Repeated Games

We saw in Chapter 12 that in oligopolistic markets, firms often find themselves in a prisoners' dilemma when making output or pricing decisions. Can firms find a way out of this dilemma, so that oligopolistic coordination and cooperation (whether explicit or implicit) could prevail?

To answer this question, we must recognize that the prisoners' dilemma, as we have described it so far, is limited: Although some prisoners may have only one opportunity in life to confess or not, most firms set output and price over and over again. In real life, firms play **repeated games:** Actions are taken and payoffs received over and over again. In repeated games, strategies can become more complex. For example, with each repetition of the prisoners' dilemma, each firm can develop a reputation about its own behavior and can study the behavior of its competitors.

repeated game Game in which actions are taken and payoffs received over and over again.

How does repetition change the likely outcome of the game? Suppose you are Firm 1 in the prisoners' dilemma illustrated by the payoff matrix in Table 13.8. If you and your competitor both charge a high price, you will both make a higher profit than if you both charged a low price. However, you are afraid to charge a high price because if your competitor charges a low price, you will lose money and, to add insult to injury, your competitor will get rich. But suppose this game is repeated over and over again—for example, you and your competitor simultaneously announce your prices on the first day of every month. Should you then play the game differently, perhaps changing your price over time in response to your competitor's behavior?

In an interesting study, Robert Axelrod asked game theorists to come up with the best strategy they could think of to play this game in a repeated manner.[9] (A possible strategy might be: "I'll start off with a high price, then lower my price. But then if my competitor lowers his price, I'll raise mine for a while before lowering it again, etc.") Then, in a computer simulation, Axelrod played these strategies off against one another to see which worked best.

Tit-for-Tat Strategy

As you would expect, any given strategy would work better against some strategies than it would against others. The objective, however, was to find the strategy that was most robust, i.e., that would work best on average against *all*, or almost all, other strategies. The result was surprising. The strategy that worked best was an extremely simple **"tit-for-tat" strategy:** I start out with a high price, which I maintain so long as you continue to "cooperate" and also charge a high price. As soon as you lower your price, however, I follow suit and lower mine. If you later decide to cooperate and raise your price again, I'll immediately raise my price as well.

TABLE 13.8 Pricing Problem

		FIRM 2	
		Low price	High price
FIRM 1	Low price	10, 10	100, – 50
	High price	– 50, 100	50, 50

[9] See Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984).

tit-for-tat strategy Repeated-game strategy in which a player responds in kind to an opponent's previous play, cooperating with cooperative opponents and retaliating against uncooperative ones.

Why does this tit-for-tat strategy work best? In particular, can I expect that using the tit-for-tat strategy will induce my competitor to behave cooperatively (and charge a high price)?

Suppose the game is *infinitely repeated*. In other words, my competitor and I repeatedly set price month after month, *forever*. Cooperative behavior (i.e., charging a high price) is then the rational response to a tit-for-tat strategy. (This assumes that my competitor knows, or can figure out, that I am using a tit-for-tat strategy.) To see why, suppose that in one month my competitor sets a low price and undercuts me. In that month he will make a large profit. But my competitor knows that the following month I will set a low price, so that his profit will fall and will remain low as long as we both continue to charge a low price. Since the game is infinitely repeated, the cumulative loss of profits that results must outweigh any short-term gain that accrued during the first month of undercutting. Thus, it is not rational to undercut.

In fact, with an infinitely repeated game, my competitor need not even be sure that I am playing tit-for-tat to make cooperation its own rational strategy. Even if my competitor believes there is only *some* chance that I am playing tit-for-tat, he will still find it rational to start by charging a high price and maintain it as long as I do. Why? With infinite repetition of the game, the *expected* gains from cooperation will outweigh those from undercutting. This will be true even if the probability that I am playing tit-for-tat (and so will continue cooperating) is small.

Now suppose the game is repeated a *finite* number of times—say, N months. (N can be large as long as it is finite.) If my competitor (Firm 2) is rational *and believes that I am rational*, he will reason as follows: "Because Firm 1 is playing tit-for-tat, I (Firm 2) cannot undercut—that is, *until the last month*. I *should* undercut in the last month because then I can make a large profit that month, and afterward the game is over, so Firm 1 cannot retaliate. Therefore, I will charge a high price until the last month, and then I will charge a low price."

However, since I (Firm 1) have also figured this out, I also plan to charge a low price in the last month. Of course, Firm 2 can figure this out as well, and therefore *knows* that I will charge a low price in the last month. But then what about the next-to-last month? Because there will be no cooperation in the last month, anyway, Firm 2 figures that it should undercut and charge a low price in the next-to-last month. But, of course, I have figured this out too, so I *also* plan to charge a low price in the next-to-last month. And because the same reasoning applies to each preceding month, the only rational outcome is for both of us to charge a low price every month.

Since most of us do not expect to live forever, the tit-for-tat strategy seems of little value; once again we are stuck in the prisoners' dilemma. However, there *is* a way out if my competitor *has even a slight doubt about my "rationality."*

Suppose my competitor *thinks* (and he need not be certain) that I am playing tit-for-tat. He also thinks that *perhaps* I am playing tit-for-tat "blindly," or with limited rationality, in the sense that I have failed to work out the logical implications of a finite time horizon as discussed above. My competitor thinks, for example, that perhaps I have not figured out that he will undercut me in the last month, so that I should also charge a low price in the last month, and so on.

"*Perhaps*," thinks my competitor, "Firm 1 will play tit-for-tat blindly, charging a high price as long as I charge a high price." Then (if the time horizon is long enough), it is rational for my competitor to maintain a high price until the last month (when he will undercut me).

Note that we have stressed the word "perhaps." My competitor need not be sure that I am playing tit-for-tat "blindly," or even that I am playing tit-for-tat at all. Just the *possibility* can make cooperative behavior a good strategy (until near the end) if the time horizon is long enough. Although my competitor's conjecture about how I am playing the game might be wrong, cooperative behavior is profitable *in expected value terms*. With a long time horizon, the sum of current and future profits, weighted by the probability that the conjecture is correct, can exceed the sum of profits from warfare, even if the competitor is the first to undercut. After all, if I am wrong and my competitor charges a low price, I can shift my strategy at the cost of only one period's profit—a minor cost in light of the substantial profit that I can make if we both choose to set a high price.

Most managers don't know how long they will be competing with their rivals, and this also serves to make cooperative behavior a good strategy. If the end point of the repeated game is unknown, the unravelling argument that begins with a clear expectation of undercutting in the last month no longer applies. As with an infinitely repeated game, it will be rational to play tit-for-tat.

Thus, in a repeated game, the prisoners' dilemma can have a cooperative outcome. In most markets, the game is in fact repeated over a long and uncertain length of time, and managers have doubts about how "perfectly rationally" they and their competitors operate. As a result, in some industries, particularly those in which only a few firms compete over a long period under stable demand and cost conditions, cooperation prevails, even though no contractual arrangements are made. (The water meter industry, discussed below, is an example.) In many other industries, however, there is little or no cooperative behavior.

Sometimes cooperation breaks down or never begins because there are too many firms. More often, failure to cooperate is the result of rapidly shifting demand or cost conditions. Uncertainties about demand or costs make it difficult for the firms to reach an implicit understanding of what cooperation should entail. (Remember that an *explicit* understanding, arrived at through meetings and discussions, could lead to an antitrust violation.) Suppose, for example, that cost differences or different beliefs about demand lead one firm to conclude that cooperation means charging $50 while a second firm thinks it means $40. If the second firm charges $40, the first firm might view that as a grab for market share and respond in tit-for-tat fashion with a $35 price. A price war could then develop.

EXAMPLE 13.2 Oligopolistic Cooperation in the Water Meter Industry

For more than 30 years, almost all the water meters sold in the United States have been produced by four American companies: Rockwell International, Badger Meter, Neptune Water Meter Company, and Hersey Products. Rockwell has had about a 35 percent share of the market, and the other three firms have together had about a 50 to 55 percent share.[10]

Most buyers of water meters are municipal water utilities, who install the meters in residential and commercial establishments in order to measure water consumption and bill consumers accordingly. Since the cost of meters is a small part of the total cost of providing water, utilities are concerned mainly that the meters be accurate and reliable. Price is not a primary issue, and demand is very inelastic. Demand is also very stable; because every residence or commercial establishment must have a water meter, demand grows slowly along with the population.

In addition, utilities tend to have long-standing relationships with suppliers and are reluctant to shift from one to another. Because any new entrant will find it difficult to lure customers from existing firms, this creates a barrier to entry. Substantial economies of scale create a second barrier to entry: To capture a significant share of the market, a new entrant must invest in a large factory. This requirement virtually precludes entry by new firms.

With inelastic and stable demand and little threat of entry by new firms, the existing four firms could earn substantial monopoly profits if they set prices cooperatively. If, on the other hand, they compete aggressively, with each firm cutting price to increase its own share of the market, profits would fall to nearly competitive levels. The firms thus face a prisoners' dilemma. Can cooperation prevail?

It can and *has* prevailed. Remember that the same four firms have been playing a *repeated game* for decades. Demand has been stable and predictable, and over the years, the firms have been able to assess their own and each other's costs. In this situation, tit-for-tat strategies work well; it pays each firm to cooperate, as long as its competitors are cooperating.

As a result, the four firms operate as though they were members of a country club. There is rarely an attempt to undercut price, and each firm appears satisfied with its share of the market. While the business may appear dull, it is certainly profitable. All four firms have been earning returns on their investments that far exceed those in more competitive industries.

[10] This example is based in part on Nancy Taubenslag, "Rockwell International," Harvard Business School Case No. 9-383-019, July 1983. In 1979, Neptune Water Meter Company was acquired by Wheelabrator-Frye. Hersey Products is a small privately held company.

EXAMPLE 13.3 Competition and Collusion in the Airline Industry

In March 1983, American Airlines, whose president, Robert Crandall, had become notable for his use of the telephone (see Example 10.5), proposed that all airlines adopt a uniform fare schedule based on mileage. The rate per mile would depend on the length of the trip, with the lowest rate of 15 cents per mile for trips over 2500 miles, higher rates for shorter trips, and the highest rate, 53 cents per mile, for trips under 250 miles. For example, a one-way coach ticket from Boston to Chicago, a distance of 932 miles, would cost $233 (based on a rate of 25 cents per mile for trips between 751 and 1000 miles).

This proposal would have done away with the many different fares (some heavily discounted) then available. The cost of a ticket from one city to another would depend only on the number of miles between those cities. As a senior vice-president of American Airlines said, "The new streamlined fare structure will help reduce fare confusion." Most other major airlines reacted favorably to the plan and began to adopt it. A vice-president of TWA said, "It's a good move. It's very businesslike." United Airlines quickly announced that it would adopt the plan on routes where it competes with American, which includes most of its system, and TWA and Continental said that they would adopt it for all their routes.[11]

Why did American Airlines propose this fare structure, and what made it so attractive to the other airlines? Was it really to "help reduce fare confusion"? No, the aim was to reduce price competition and achieve a collusive pricing arrangement. Prices had been driven down by competitive undercutting, as airlines competed for market share. And as Robert Crandall had learned less than a year earlier, fixing prices over the telephone is illegal. Instead, the companies would implicitly fix prices by agreeing to use the same fare-setting formula.

The plan failed, a victim of the prisoners' dilemma. Only two weeks after the plan was announced and adopted by most airlines, Pan Am, which was dissatisfied with its small share of the U.S. market, dropped its fares. American, United, and TWA, afraid of losing their own shares of the market, quickly dropped their fares to match Pan Am. The price-cutting continued, and fortunately for consumers, the plan was soon dead.

This episode exemplifies the problem of oligopolistic pricing. One economist summarized it accurately: "You can't blame American Airlines for trying. After all, it is the American Way to try to cartelize prices with a simple formula. But it is also in the great tradition of open competition in this country to frustrate any such establishment of cartel prices by competitive chiseling."[12]

American Airlines introduced another simplified, four-tier fare structure in April 1992, which was quickly adopted by most major carriers. But it, too, soon fell victim to competitive discounts. In May 1992, Northwest Airlines announced a "kids fly free" program, and American responded with a summer half-price sale, which other carriers matched. As a result, the airline industry lost billions of dollars in 1992.

Why is airline pricing so intensively competitive? Airlines plan route capacities two or more years into the future, but they make pricing decisions over short horizons—month by month or even week by week. In the short run, the marginal cost of adding passengers to a flight is very low—essentially the cost of a soft drink and a bag of peanuts. Each airline, therefore, has an incentive to lower fares in order to capture passengers from its competitors. In addition, the demand for air travel often fluctuates unpredictably. Such factors as these stand in the way of implicit price cooperation.

13.5 Sequential Games

sequential game Game in which players move in turn, responding to each other's actions and reactions.

In most of the games we have discussed so far, both players move at the same time. In the Cournot model of duopoly, for example, both firms set output at the same time. In **sequential games,** players move in turn. The Stackelberg model discussed in Chapter 12 is an example of a sequential game; one firm sets output before the other does. There are many other examples: an advertising decision by one firm and the response by its competitor; entry-deterring investment by an incumbent firm and the decision whether to enter the market by a potential competitor; or a new government regulatory policy and the investment and output response of the regulated firms.

[11] "American to Base Fares on Mileage," *New York Times*, March 15, 1983; "Most Big Airlines Back American's Fare Plan," *New York Times*, March 17, 1983.

[12] Paul W. MacAvoy, "A Plan That Won't Endure Competition," *New York Times*, April 3, 1983.

TABLE 13.9 Modified Product Choice Problem

		FIRM 2	
		Crispy	Sweet
FIRM 1	Crispy	−5, −5	10, 20
	Sweet	20, 10	−5, −5

We will look at a variety of sequential games in the remainder of this chapter. As we will see, they are often easier to analyze than games in which the players move at the same time. In a sequential game, the key is to think through the possible actions and rational reactions of each player.

As a simple example, let's return to the product choice problem first discussed in Section 13.3. This problem involves two companies facing a market in which two new variations of breakfast cereal can be successfully introduced as long as each firm introduces only one variation. This time, let's change the payoff matrix slightly. As Table 13.9 shows, the new sweet cereal will inevitably be a better seller than the new crispy cereal, earning a profit of 20 rather than 10 (perhaps because consumers prefer sweet things to crispy things). Both new cereals will still be profitable, however, as long as each is introduced by only one firm. (Compare Table 13.9 with Table 13.3).

Suppose that both firms, in ignorance of each other's intentions, must announce their decisions independently and simultaneously. In that case, both will probably introduce the sweet cereal—and both will lose money.

Now suppose that Firm 1 can gear up its production faster and introduce its new cereal first. We now have a sequential game: Firm 1 introduces a new cereal, and then Firm 2 introduces one. What will be the outcome of this game? When making its decision, Firm 1 must consider the rational response of its competitor. It knows that whichever cereal it introduces, Firm 2 will introduce the other kind. Thus it will introduce the sweet cereal, knowing that Firm 2 will respond by introducing the crispy one.

The Extensive Form of a Game

extensive form of a game Representation of possible moves in a game in the form of a decision tree.

Although this outcome can be deduced from the payoff matrix in Table 13.9, sequential games are sometimes easier to visualize if we represent the possible moves in the form of a decision tree. This representation is called the **extensive form of a game,** and is shown in Figure 13.2. The figure shows the possible choices of Firm 1 (introduce a crispy or a sweet cereal) and the possible responses of Firm 2 to each of those choices. The resulting payoffs are given at the end of each branch. For example, if Firm 1 produces a crispy cereal and Firm 2 responds by also producing a crispy cereal, each firm will have a payoff of −5.

To find the solution to the extensive form game, work backward from the end. For Firm 1, the best sequence of moves is the one in which it earns 20 and Firm 2 earns 10. Thus it can deduce that it should produce the sweet cereal, because Firm 2's best response is then to produce the crispy cereal.

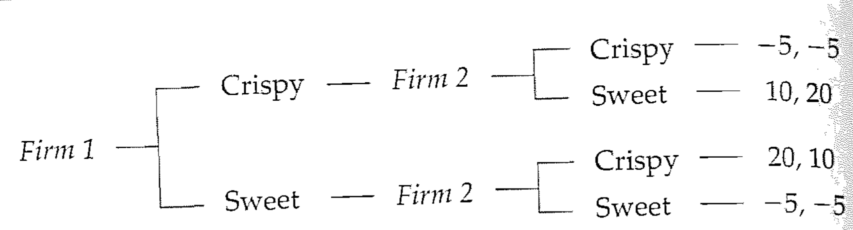

FIGURE 13.2 Product Choice Game in Extensive Form

The Advantage of Moving First

In this product-choice game, there is a clear advantage to moving first: By introducing the sweet cereal, Firm 1 creates a fait accompli that leaves Firm 2 little choice but to introduce the crispy one. This is much like the first-mover advantage that we saw in the Stackelberg model in Chapter 12. In that model, the firm that moves first can choose a large level of output, thereby giving its competitor little choice but to choose a small level.

In §12.3, we explain that the Stackelberg model is an oligopoly model in which one firm sets its output before other firms do.

To clarify the nature of this first-mover advantage, it will be useful to review the Stackelberg model and compare it to the Cournot model in which both firms choose their outputs simultaneously. As in Chapter 12, we will use the example in which two duopolists face the market demand curve

Recall that in §12.2, we explain that in the Cournot model, each firm treats the output of its competitors as fixed and that all firms simultaneously decide how much to produce.

$$P = 30 - Q$$

where Q is the total production, i.e., $Q = Q_1 + Q_2$. As before, we will also assume that both firms have zero marginal cost. Recall that the Cournot equilibrium is then $Q_1 = Q_2 = 10$, so that $P = 10$ and each firm earns a profit of 100. Recall also that if the two firms colluded, they would set $Q_1 = Q_2 = 7.5$, so that $P = 15$ and each firm earns a profit of 112.50. Finally, recall from Section 12.3 that in the Stackelberg model, in which Firm 1 moves first, the outcome is $Q_1 = 15$ and $Q_2 = 7.5$, so that $P = 7.50$ and the firms' profits are 112.50 and 56.25, respectively.

These and a few other possible outcomes are summarized in the payoff matrix of Table 13.10. If both firms move simultaneously, the only solution to the game is that both firms produce 10 and earn 100. In this Cournot equilibrium, each firm is doing the best it can given what its competitor is doing. If Firm 1 moves first, however, it knows that its decision will constrain Firm 2's choice.

TABLE 13.10 Choosing Output

		FIRM 2		
		7.5	10	15
FIRM 1	7.5	112.50, 112.50	93.75, 125	56.25, 112.50
	10	125, 93.75	100, 100	50, 75
	15	112.50, 56.25	75, 50	0, 0

Observe from the payoff matrix that if Firm 1 sets $Q_1 = 7.5$, Firm 2's best response will be to set $Q_2 = 10$. This will give Firm 1 a profit of 93.75 and Firm 2 a profit of 125. If Firm 1 sets $Q_1 = 10$, Firm 2 will set $Q_2 = 10$, and both firms will earn 100. But if Firm 1 sets $Q_1 = 15$, Firm 2 will set $Q_2 = 7.5$, so that Firm 1 earns 112.50, and Firm 2 earns 56.25. Therefore the most that Firm 1 can earn is 112.50, and it does so by setting $Q_1 = 15$. Compared to the Cournot outcome, when Firm 1 moves first, it does better—and Firm 2 does much worse.

13.6 Threats, Commitments, and Credibility

The product choice problem and the Stackelberg model are two examples of how a firm that moves first can create a fait accompli that gives it an advantage over its competitor. In this section we'll take a broader look at the advantage a firm can have by moving first, and also consider what determines *which* firm goes first. We will focus on the following question: *What actions can a firm take to gain advantage in the marketplace?* For example, how might a firm deter entry by potential competitors, or induce existing competitors to raise prices, reduce output, or leave the market altogether? Or how might a firm reach an implicit agreement with its competitors that is heavily weighted in its own favor?

An action that gives a firm this kind of advantage is called a **strategic move**. A good definition of a strategic move was given by Thomas Schelling, who first explained the concept and its implications: "A strategic move is one that influences the other person's choice in a manner favorable to one's self, by affecting the other person's expectations of how one's self will behave. One constrains the partner's choice by constraining one's own behavior."[13]

strategic move Action that gives a player an advantage by constraining his behavior.

The idea of constraining your own behavior to gain an advantage may seem paradoxical, but we'll soon see that it is not. Let's consider a few examples.

First, let's return once more to the product-choice problem shown in Table 13.9. The firm that introduces its new breakfast cereal first will do best. *But which firm will introduce its cereal first?* Even if both firms require the same amount of time to gear up production, each has an incentive to *commit itself first to the sweet cereal*. The key word is "commit." If Firm 1 simply announces it will produce the sweet cereal, Firm 2 will have little reason to believe it. After all, Firm 2, knowing the incentives, can make the same announcement louder and more vociferously. Firm 1 must constrain its own behavior in some way that convinces Firm 2 that Firm 1 has *no choice* but to produce the sweet cereal. Firm 1 might launch an expensive advertising campaign describing the new sweet cereal well before its introduction, thereby putting its reputation on the line. Firm 1 might also sign a contract for the forward delivery of a large quantity of sugar (and make the contract public, or at least send a copy to Firm 2). The idea is for Firm 1 to *commit itself* to produce the sweet cereal. Commitment is a strategic move that will induce Firm 2 to make the decision that Firm 1 wants it to make—namely, to produce the crispy cereal.

[13] Thomas C. Schelling, *The Strategy of Conflict* (New York: Oxford University Press, 1960), p. 160 (1980 edition published by Harvard University Press). For a general discussion of strategic moves in business planning, see Michael E. Porter, *Competitive Strategy* (New York: Free Press, 1980).

TABLE 13.11 Pricing of Computers and Word Processors

		FIRM 2	
		High price	Low price
FIRM 1	High price	100, 80	80, 100
	Low price	20, 0	10, 20

Why can't Firm 1 simply *threaten* Firm 2, vowing to produce the sweet cereal even if Firm 2 does the same? Because Firm 2 has little reason to believe the threat—and can make the same threat itself. A threat is useful only if it is credible. The following example should help make this clear.

Empty Threats

Suppose Firm 1 produces personal computers that can be used both as word processors and to do other tasks. Firm 2 produces only dedicated word processors. As the payoff matrix in Table 13.11 shows, as long as Firm 1 charges a high price for its computers, both firms can make a good deal of money. Even if Firm 2 charges a low price for its word processors, many people will still buy Firm 1's computers (because they can do so many other things), although some buyers will be induced by the price differential to buy the dedicated word processor instead. However, if Firm 1 charges a low price, Firm 2 will also have to charge a low price (or else make zero profit), and the profit of both firms will be significantly reduced.

Firm 1 would prefer the outcome in the upper left-hand corner of the matrix. For Firm 2, however, charging a low price is clearly a dominant strategy. Thus the outcome in the upper right-hand corner will prevail (no matter which firm sets its price first).

Firm 1 would probably be viewed as the "dominant" firm in this industry because its pricing actions will have the greatest impact on overall industry profits. Can Firm 1 induce Firm 2 to charge a high price by *threatening* to charge a low price if Firm 2 charges a low price? No, as the payoff matrix in Table 13.11 makes clear: *Whatever* Firm 2 does, Firm 1 will be much worse off if it charges a low price. As a result, its threat is not credible.

Commitment and Credibility

Sometimes firms can make threats credible. To see how, consider the following example. Race Car Motors, Inc., produces cars, and Far Out Engines, Ltd., produces specialty car engines. Far Out Engines sells most of its engines to Race Car Motors, and a few to a limited outside market. Nonetheless, it depends heavily on Race Car Motors, and makes its production decisions in response to Race Car's production plans.

We thus have a sequential game in which Race Car is the "leader." It will decide what kind of cars to build, and Far Out Engines will then decide what kind of engines to produce. The payoff matrix in Table 13.12(a) shows the possible outcomes of this game. (Profits are in millions of dollars.) Observe that Race Car will do best by deciding to produce small cars. It knows that in response to this decision, Far Out will produce small engines, most of which Race Car will then buy. As a result, Far Out will make $3 million and Race Car $6 million.

TABLE 13.12(a) Production Choice Problem

		RACE CAR MOTORS	
		Small cars	Big cars
FAR OUT ENGINES	Small engines	3, 6	3, 0
	Big engines	1, 1	8, 3

Far Out, however, would much prefer the outcome in the lower right-hand corner of the payoff matrix. If it could produce big engines, *and* if Race Car produced big cars and therefore bought the big engines, it would make $8 million. (Race Car, however, would make only $3 million). Can Far Out induce Race Car to produce big cars instead of small ones?

Suppose Far Out *threatens* to produce big engines no matter what Race Car does; suppose, too, that no other engine producer can easily satisfy the needs of Race Car. If Race Car believed Far Out's threat, it would produce big cars: Otherwise, it would have trouble finding engines for its small cars and would earn only $1 million instead of $3 million. But the threat is not credible: Once Race Car responded by announcing its intentions to produce small cars, Far Out would have no incentive to carry out its threat.

Far Out can make its threat credible by visibly and irreversibly reducing some of its own payoffs in the matrix, so that its choices become constrained. In particular, Far Out must reduce its profits from small engines (the payoffs in the top row of the matrix). It might do this by *shutting down or destroying some of its small engine production capacity*. This would result in the payoff matrix shown in Table 13.12(b). Now Race Car *knows* that whatever kind of car it produces, Far Out will produce big engines. If Race Car produces the small cars, Far Out will sell the big engines as best it can to other car producers and settle for making only $1 million. But this is better than making no profits by producing small engines. Because Race Car will have to look elsewhere for engines, its profit will also be lower ($1 million). Now it is clearly in Race Car's interest to produce large cars. By making a strategic move that *seemingly puts itself at a disadvantage*, Far Out has improved the outcome of the game.

Although strategic commitments of this kind can be effective, they are risky and depend heavily on having accurate knowledge of the payoff matrix and the industry. Suppose, for example, that Far Out commits itself to producing big engines but is surprised to find that another firm can produce small engines at a low cost. The commitment may then lead Far Out to bankruptcy rather than to continued high profits.

TABLE 13.12(b) Modified Production Choice Problem

		RACE CAR MOTORS	
		Small cars	Big cars
FAR OUT ENGINES	Small engines	0, 6	0, 0
	Big engines	1, 1	8, 3

The Role of Reputation Developing the right kind of *reputation* can also give one a strategic advantage. Again, consider Far Out Engines' desire to produce big engines for Race Car Motors' big cars. Suppose that the managers of Far Out Engines develop a reputation for being irrational—perhaps downright crazy. They threaten to produce big engines no matter what Race Car Motors does. (Refer to Table 13.12a.) Now the threat might be credible without any further action; after all, you can't be sure that an irrational manager will always make a profit-maximizing decision. In gaming situations, the party that is known (or thought) to be a little crazy can have a significant advantage.

Developing a reputation can be an especially important strategy in a repeated game. A firm might find it advantageous to behave irrationally for several plays of the game. This might give it a reputation that will allow it to increase its long-run profits substantially.

EXAMPLE 13.4 Wal-Mart Stores' Preemptive Investment Strategy

Wal-Mart Stores, Inc., is an enormously successful chain of discount retail stores started by Sam Walton in 1969.[14] Its success was unusual in the industry. During the 1960s and 1970s, rapid expansion by existing firms and the entry and expansion of new firms made discount retailing increasingly competitive. During the 1970s and 1980s, industrywide profits fell, and large discount chains—including such giants as King's, Korvette's, Mammoth Mart, W. T. Grant, and Woolco—went bankrupt. Wal-Mart Stores, however, kept on growing (from 153 stores in 1976 to 1009 in 1986) and became even more profitable. By the end of 1985, Sam Walton was one of the richest people in the United States.

How did Wal-Mart Stores succeed where others failed? The key is in Wal-Mart's expansion strategy. To charge less than ordinary department stores and small retail stores, discount stores rely on size, no frills, and high inventory turnover. Through the 1960s, the conventional wisdom held that a discount store could succeed only in a city with a population of 100,000 or more. Sam Walton disagreed and decided to open his stores in small Southwestern towns; by 1970, there were 30 Wal-Mart stores in small towns in Arkansas, Missouri, and Oklahoma. The stores succeeded because Wal-Mart had created 30 "local monopolies." Discount stores that had opened in larger towns and cities were competing with other discount stores, which drove down prices and profit margins. These small towns, however, had room for only one discount operation. Wal-Mart could undercut the nondiscount retailers but never had to worry that another discount store would open and compete with it.

By the mid-1970s, other discount chains realized that Wal-Mart had a profitable strategy: Open a store in a small town that could support only one discount store and enjoy a local monopoly. There are a lot of small towns in the United States, so the issue became who would get to each town first. Wal-Mart now found itself in a *preemption game* of the sort illustrated by the payoff matrix in Table 13.13. As the matrix shows, if Wal-Mart enters a town but Company X does not, Wal-Mart will make 20 and Company X will make 0.

[14] This example is based in part on information in Pankaj Ghemawat, "Wal-Mart Stores' Discount Operations," Harvard Business School, 1986.

TABLE 13.13 The Discount Store Preemption Game

		COMPANY X	
		Enter	Don't enter
WAL-MART	Enter	− 10, − 10	20, 0
	Don't enter	0, 20	0, 0

Similarly, if Wal-Mart doesn't enter but Company X does, Wal-Mart makes 0 and Company X makes 20. But if Wal-Mart and Company X *both* enter, *they both lose 10.*

This game has two Nash equilibria—the lower left-hand corner and the upper right-hand corner. Which equilibrium results depends on *who moves first.* If Wal-Mart moves first, it can enter, knowing that the rational response of Company X will be not to enter, so that Wal-Mart will be assured of earning 20. *The trick, therefore, is to preempt*—to set up stores in other small towns quickly, before Company X (or Company Y or Z) can do so. That is exactly what Wal-Mart did. By 1986, it had 1009 stores in operation and was earning an annual profit of $450 million. And while other discount chains were going under, Wal-Mart continued to grow. By 1993, it had over 1800 stores and was earning an annual profit of over $1.5 billion. By 1999, Wal-Mart had 2454 stores in the United States and another 729 stores in the rest of the world, and had annual sales of $138 billion.

13.7 Entry Deterrence

Barriers to entry, which are an important source of monopoly power and profits, sometimes arise naturally. For example, economies of scale, patents and licenses, or access to critical inputs can create entry barriers. However, firms themselves can sometimes deter entry by potential competitors.

To deter entry, *the incumbent firm must convince any potential competitor that entry will be unprofitable.* To see how this might be done, put yourself in the position of an incumbent monopolist facing a prospective entrant, Firm X. Suppose that to enter the industry, Firm X will have to pay a (sunk) cost of $80 million to build a plant. You, of course, would like to induce Firm X to stay out of the industry. If X stays out, you can continue to charge a high price and enjoy monopoly profits. As shown in the upper right-hand corner of the payoff matrix in Table 13.14(a), you would earn $200 million in profits.

In §7.1, we explain that a sunk cost is an expenditure that has been made and cannot be recovered.

TABLE 13.14(a) Entry Possibilities

		POTENTIAL ENTRANT	
		Enter	Stay out
INCUMBENT	High price (accommodation)	100, 20	200, 0
	Low price (warfare)	70, − 10	130, 0

If Firm X does enter the market, you must make a decision. You can be "accommodating," maintaining a high price in the hope that X will do the same. In that case, you will earn only $100 million in profit because you will have to share the market. New entrant X will earn a *net* profit of $20 million: $100 million minus the $80 million cost of constructing a plant. (This outcome is shown in the upper left-hand corner of the payoff matrix.) Alternatively, you can increase your production capacity, produce more, and lower your price. The lower price will give you a greater market share and a $20 million increase in revenues. Increasing production capacity, however, will cost $50 million, reducing your net profit to $70 million. Because warfare will also reduce the entrant's revenue by $30 million, it will have a net loss of $10 million. (This outcome is shown in the lower left-hand corner of the payoff matrix.) Finally, if Firm X stays out but you expand capacity and lower price nonetheless, your net profit will fall by $70 million (from $200 million to $130 million): the $50 million cost of the extra capacity and a $20 million reduction in revenue from the lower price with no gain in market share. (Clearly this choice, shown in the lower right-hand corner of the matrix, would make no sense.)

If Firm X thinks you will be accommodating and maintain a high price after it has entered, it will find it profitable to enter and will do so. Suppose you threaten to expand output and wage a price war in order to keep X out. If X takes the threat seriously, it will not enter the market because it can expect to lose $10 million. The threat, however, is not credible. As Table 13.14(a) shows (and as the potential competitor knows), *once entry has occurred, it will be in your best interest to accommodate and maintain a high price.* Firm X's rational move is to enter the market; the outcome will be the upper left-hand corner of the matrix.

But what if you can make an irrevocable commitment that will alter your incentives once entry occurred—a commitment that will give you little choice but to charge a low price if entry occurs? In particular, suppose you invest the $50 million *now,* rather than later, in the extra capacity needed to increase output and engage in competitive warfare should entry occur. Of course, if you later maintain a high price (whether or not X enters), this added cost will reduce your payoffs.

We now have a new payoff matrix, as shown in Table 13.14(b). As a result of your decision to invest in additional capacity, your threat to engage in competitive warfare is *completely credible.* Because you already have the additional capacity with which to wage war, you will do better in competitive warfare than you would by maintaining a high price. Because the potential competitor now knows that entry will result in warfare, it is rational for it to stay out of the market. Meanwhile, having deterred entry, you can maintain a high price and earn a profit of $150 million.

Can an incumbent monopolist deter entry without making the costly move of installing additional production capacity? Earlier we saw that a reputation for irrationality can bestow a strategic advantage. Suppose the incumbent firm has

TABLE 13.14(b) Entry Deterrence

		POTENTIAL ENTRANT	
		Enter	Stay out
INCUMBENT	High price (accommodation)	50, 20	150, 0
	Low price (warfare)	70, −10	130, 0

such a reputation. Suppose also that with vicious price-cutting this firm has eventually driven out every entrant in the past, even though it incurred (rationally unwarranted) losses in doing so. Its threat might then indeed be credible. In this case, the incumbent's irrationality suggests to the potential competitor that it might be better off staying away.

Of course, if the game described above were to be *indefinitely repeated,* then the incumbent might have a *rational* incentive to carry out the threat of warfare whenever entry actually occurs. Why? Because short-term losses from warfare might be outweighed by longer-term gains from preventing entry. Understanding this, the potential competitor might find the incumbent's threat of warfare credible and decide to stay out. Now the incumbent relies on its reputation for being rational—and in particular for being far-sighted—to provide the credibility needed to deter entry. The success of this strategy depends on the time horizon and the relative gains and losses associated with accommodation and warfare.

We have seen that the attractiveness of entry depends largely on the way incumbents can be expected to react. In general, incumbents cannot be expected to maintain output at the preentry level once entry has occurred. Eventually, they may back off and reduce output, raising price to a new joint profit-maximizing level. Because potential entrants know this, incumbent firms must create a credible threat of warfare to deter entry. A reputation for irrationality can help do this. Indeed, this seems to be the basis for much of the entry-preventing behavior that goes on in actual markets. The potential entrant must consider that *rational* industry discipline can break down after entry occurs. By fostering an image of irrationality and belligerence, an incumbent firm might convince potential entrants that the risk of warfare is too high.[15]

Strategic Trade Policy and International Competition

We have seen how a preemptive investment can give a firm an advantage by creating a credible threat to potential competitors. In some situations, a preemptive investment—subsidized or otherwise encouraged by the government—can give a *country* an advantage in international markets and so be an important instrument of trade policy.

Does this conflict with what you have learned about the benefits of free trade? In Chapter 9, for example, we saw how trade restrictions such as tariffs or quotas lead to deadweight losses. In Chapter 16 we go further and show how, in a general way, free trade between people (or between countries) is mutually beneficial. Given the virtues of free trade, how can government intervention in an international market ever be warranted? An emerging literature in international trade theory suggests that in certain situations, a country can benefit by adopting policies that give its domestic industries a competitive advantage.

[15]There is an analogy here to *nuclear deterrence.* Consider the use of a nuclear threat to deter the former Soviet Union from invading Western Europe during the Cold War. If it invaded, would the United States actually react with nuclear weapons, knowing that the Soviets would then respond in kind? It is not rational for the United States to react this way, so a nuclear threat might not seem credible. But this assumes that everyone is rational; there is a reason to fear an *irrational* response by the United States. Even if an irrational response is viewed as very improbable, it can be a deterrent, given the costliness of an error. The United States can thus gain by promoting the idea that it might act irrationally, or that events might get out of control once an invasion occurs. This is the "rationality of irrationality." See Schelling, *The Strategy of Conflict.*

To see how this might occur, consider an industry with substantial economies of scale—one in which a few large firms can produce much more efficiently than many small ones. Suppose that by granting subsidies or tax breaks, the government can encourage domestic firms to expand faster than they would otherwise. This might prevent firms in other countries from entering the world market, so that the domestic industry can enjoy higher prices and greater sales. Such a policy works by creating a credible threat to potential entrants. Large domestic firms, taking advantage of scale economies, would be able to satisfy world demand at a low price; if other firms entered, price would be driven below the point at which they could make a profit.

The Commercial Aircraft Market As an example, consider the international market for commercial aircraft. The development and production of a new line of aircraft are subject to substantial economies of scale; it would not pay to develop a new aircraft unless a firm expected to sell many of them. Suppose that Boeing and Airbus (a European consortium that includes France, Germany, Britain, and Spain) are each considering developing a new aircraft (as indeed they were in the late 1970s and early 1980s). The ultimate payoff to each firm depends in part on what the other firm does. Suppose it is only economical for one firm to produce the new aircraft. Then the payoffs might look like those in Table 13.15(a).[16]

If Boeing has a head start in the development process, the outcome of the game is the upper right-hand corner of the payoff matrix. Boeing will produce a new aircraft, and Airbus, realizing that it will lose money if it does the same, will not. Boeing will then earn a profit of 100.

European governments, of course, would prefer that Airbus produce the new aircraft. Can they change the outcome of this game? Suppose they commit to subsidizing Airbus and make this commitment before Boeing has committed itself to produce. If the European governments commit to a subsidy of 20 to Airbus if Airbus produces the plane *regardless of what Boeing does*, the payoff matrix would change to the one in Table 13.15(b).

TABLE 13.15(a) Development of a New Aircraft

		AIRBUS: Produce	AIRBUS: Don't produce
BOEING	Produce	−10, −10	100, 0
	Don't produce	0, 100	0, 0

TABLE 13.15(b) Development of Aircraft after European Subsidy

		AIRBUS: Produce	AIRBUS: Don't produce
BOEING	Produce	−10, 10	100, 0
	Don't produce	0, 120	0, 0

[16] This example is drawn from Paul R. Krugman, "Is Free Trade Passé?" *Journal of Economic Perspectives* 1 (Fall 1987): 131–44.

Now Airbus will make money from a new aircraft whether or not Boeing produces one. Boeing knows that even if it commits to producing, Airbus will produce as well, and Boeing will lose money. Thus Boeing will decide not to produce, and the outcome will be the one in the lower left-hand corner of Table 13.15(b). A subsidy of 20, then, changes the outcome from one in which Airbus does not produce and earns 0 to one in which it does produce and earns 120. Of this, 100 is a transfer of profit from the United States to Europe. From the European point of view, subsidizing Airbus yields a high return.

European governments *did* commit to subsidizing Airbus, and during the 1980s, Airbus successfully introduced several new airplanes. The result, however, was not quite the one reflected in our stylized example. Boeing also introduced new airplanes (the 757 and 767 models) that were extremely profitable. As commercial air travel grew, it became clear that both companies could profitably develop and sell a new generation of airplanes. Nonetheless, Boeing's market share would have been much larger without the European subsidies to Airbus. One study estimated that those subsidies totalled $25.9 billion during the 1980s and found that Airbus would not have entered the market without them.[17]

This example shows how strategic trade policy can transfer profits from one country to another. Bear in mind, however, that a country which uses such a policy may provoke retaliation from its trading partners. If a trade war results, all countries can end up much worse off. The possibility of such an outcome must be considered before a nation adopts a strategic trade policy.

EXAMPLE 13.5 DuPont Deters Entry in the Titanium Dioxide Industry

Titanium dioxide is a whitener used in paints, paper, and other products. In the early 1970s, DuPont and National Lead each accounted for about a third of U.S. titanium dioxide sales; another seven firms produced the remainder. In 1972, DuPont was considering whether to expand capacity. The industry was changing, and with the right strategy, those changes might enable DuPont to capture more of the market and dominate the industry.[18]

Three factors had to be considered. First, although future demand for titanium dioxide was uncertain, it was expected to grow substantially. Second, the government had announced that new environmental regulations would be imposed. Third, the prices of raw materials used to make titanium dioxide were rising. The new regulations and the higher input prices would have a major effect on production cost and give DuPont a cost advantage, both because its production technology was less sensitive to the change in input prices and because its plants were in areas that made disposal of corrosive wastes much less difficult than for other producers. Because of these cost changes, DuPont anticipated that National Lead and some other producers would have to shut down part of their capacity. DuPont's competitors would in effect have to "reenter" the market by building new plants. Could DuPont deter them from taking this step?

[17] "Aid to Airbus Called Unfair in U.S. Study," *New York Times*, September 8, 1990.

[18] This example is based on Pankaj Ghemawat, "Capacity Expansion in the Titanium Dioxide Industry," *Journal of Industrial Economics* 33 (December 1984): 145–63; and P. Ghemawat, "DuPont in Titanium Dioxide," Harvard Business School, Case No. 9-385-140, June 1986.

In 1972, DuPont considered the following strategy: invest nearly $400 million in increased production capacity to try to capture 64 percent of the market by 1985. The production capacity that would be put on line would be much more than what was actually needed. The idea was to *deter competitors from investing*. Scale economies and movement down the learning curve would give DuPont a cost advantage. This would not only make it hard for other firms to compete, but would make credible the implicit threat that in the future, DuPont would fight rather than accommodate.

The strategy was sensible and seemed to work for a few years. By 1975, however, things began to go awry. First, because demand grew by much less than expected, there was excess capacity industrywide. Second, because the environmental regulations were only weakly enforced, competitors did not have to shut down capacity as expected. Finally, DuPont's strategy led to antitrust action by the Federal Trade Commission in 1978. The FTC claimed that DuPont was attempting to monopolize the market. DuPont won the case, but the decline in demand made its victory moot.

EXAMPLE 13.6 Diaper Wars

For more than a decade, the disposable diaper industry in the United States has been dominated by two firms: Procter & Gamble, with an approximately 50–60 percent market share, and Kimberly-Clark, with another 30 percent.[19] How do these firms compete? And why haven't other firms been able to enter and take a significant share of this $4 billion per year market?

Even though there are only two major firms, competition is intense. The competition occurs mostly in the form of *cost-reducing innovation*. The key to success is to perfect the manufacturing process so that a plant can manufacture diapers in high volume and at low cost. This is not as simple as it might seem. Packing cellulose fluff for absorbency, adding an elastic gatherer, and binding, folding, and packaging the diapers—at a rate of about 3000 diapers per minute and at a cost of about 8 to 10 cents per diaper—requires an innovative, carefully designed, and finely tuned process. Furthermore, small technological improvements in the manufacturing process can result in a significant competitive advantage. If a firm can shave its production cost even slightly, it can reduce price and capture market share. As a result, both firms are forced to spend heavily on research and development (R&D) in a race to reduce cost.

The payoff matrix in Table 13.16 illustrates this. If both firms spend aggressively on R&D, they can expect to maintain their current market shares. P&G will earn a profit of 40, and Kimberly-Clark (with a smaller market share) will earn 20. If neither firm spends money on R&D, their costs and prices will remain constant and the money saved will become part of profits. P&G's profit will increase to 60 and Kimberly-Clark's to 40. However, if one firm continues to do R&D and the other doesn't, the innovating firm will eventually capture most of its competitor's market share. For example, if Kimberly-Clark does R&D and P&G doesn't, P&G can expect to lose 20 while Kimberly-Clark's profit increases to 60. The two firms are therefore in a prisoners' dilemma: Spending money on R&D is a dominant strategy for each firm.

[19] Procter & Gamble makes Pampers, Ultra Pampers, and Luvs. Kimberly-Clark has only one major brand, Huggies.

TABLE 13.16 Competing through R&D

		KIMBERLY-CLARK	
		R&D	No R&D
P&G	R&D	40, 20	80, – 20
	No R&D	– 20, 60	60, 40

Why hasn't cooperative behavior evolved? After all, the two firms have been competing in this market for years, and the demand for diapers is fairly stable. For several reasons, a prisoners' dilemma involving R&D is particularly hard to resolve. First, it is difficult for a firm to monitor its competitor's R&D activities the way it can monitor price. Second, it can take several years to complete an R&D program that leads to a major product improvement. As a result, tit-for-tat strategies, in which both firms cooperate until one of them "cheats," are less likely to work. A firm may not find out that its competitor has been secretly doing R&D until the competitor announces a new and improved product. By then it may be too late to gear up an R&D program of its own.

The ongoing R&D expenditures by P&G and Kimberly-Clark also serve to deter entry. In addition to brand name recognition, these two firms have accumulated so much technological know-how and manufacturing proficiency that they would have a substantial cost advantage over any firm just entering the market. Besides building new factories, an entrant would have to spend a considerable amount on R&D to capture even a small share of the market. After it began producing, a new firm would have to continue to spend heavily on R&D to reduce its costs over time. Entry would be profitable only if P&G and Kimberly-Clark stop doing R&D, so that the entrant could catch up and eventually gain a cost advantage. But as we have seen, no rational firm would expect this to happen.[20]

13.8 Bargaining Strategy

In analyzing the prisoners' dilemma and related problems, we have assumed that collusion was limited by an inability to make enforceable agreements. Clearly, alternative outcomes are possible (and likely) if firms or individuals can make promises that can be enforced. The prisoners' dilemma illustrated by the pricing problem shown in Table 13.8 (p. 472) is a good example of this. If there were no antitrust laws and both firms could make an enforceable agreement about pricing, they would both charge a high price and make profits of 50. Here, the bargaining problem is simple.

[20] Example 15.3 in Chapter 15 examines in more detail the profitability of capital investment by a new entrant in the diaper market.

TABLE 13.17 Production Decision

		FIRM 2	
		Produce *A*	Produce *B*
FIRM 1	Produce *A*	40, 5	50, 50
	Produce *B*	60, 40	5, 45

Other bargaining situations are more complicated, however, and the outcome can depend on the ability of either side to make a strategic move that alters its relative bargaining position. For example, consider two firms that are each planning to introduce one of two products, which happen to be complementary goods. As the payoff matrix in Table 13.17 shows, Firm 1 has a cost advantage over Firm 2 in producing *A*. Therefore, if both firms produce *A*, Firm 1 can maintain a lower price and earn a much higher profit. Similarly, Firm 2 has a cost advantage over Firm 1 in producing product *B*. As should be clear from the payoff matrix, if the two firms could agree about who will produce what, the rational outcome would be the one in the upper right-hand corner: Firm 1 produces *A*, Firm 2 produces *B*, and both firms make profits of 50. Indeed, even *without cooperation*, this outcome will result, whether Firm 1 or Firm 2 moves first or both firms move simultaneously. Why? Because producing *B* is a dominant strategy for Firm 2, so (*A*, *B*) is the only Nash equilibrium.

Firm 1, of course, would prefer the outcome in the lower left-hand corner of the payoff matrix. But in the context of this limited set of decisions, it cannot achieve that outcome. Suppose, however, that Firms 1 and 2 are also bargaining over a second issue—whether to join a research consortium that a third firm is trying to form. Table 13.18 shows the payoff matrix for this decision problem. Clearly, the dominant strategy is for both firms to enter the consortium, thereby obtaining increased profits of 40.

Now suppose that Firm 1 *links the two bargaining problems* by announcing that it will join the consortium *only* if Firm 2 agrees to produce product *A*. In this case, it is indeed in Firm 2's interest to produce *A* (with Firm 1 producing *B*) in return for Firm 1's participation in the consortium. This example illustrates how a strategic move can be used in bargaining, and why combining issues in a bargaining agenda can sometimes benefit one side at the other's expense.

As another example, consider two people bargaining over the price of a house. Suppose I, as a potential buyer, do not want to pay more than $200,000 for a house that is actually worth $250,000 to me. The seller is willing to part with the house at any price above $180,000 but would like to receive the highest price she can. If I am the only bidder for the house, how can I make the seller think that I will walk away rather than pay more than $200,000?

TABLE 13.18 Decision to Join Consortium

		FIRM 2	
		Work alone	Enter consortium
FIRM 1	Work alone	10, 10	10, 20
	Enter consortium	20, 10	40, 40

I might declare that I will never, ever pay more than $200,000 for the house. But is such a promise credible? It is if the seller knows that I have a *strong reputation* for toughness and steadfastness and that I have never broken my word on a promise of this sort. But suppose I have no such reputation. Then the seller knows that I have every incentive to make the promise (making it costs nothing) but little incentive to keep it (this will probably be our only business transaction together). As a result, this promise by itself is not likely to improve my bargaining position.

The promise can work, however, if it is combined with a strategic move that gives it credibility. Such a strategic move must reduce my flexibility—limit my options—so that I have no choice but to keep the promise. A possible move would be to make an enforceable bet with a third party—for example, "If I pay more than $200,000 for that house, I'll pay you $60,000." Alternatively, if I am buying the house on behalf of my company, the company might insist on authorization by the Board of Directors for a price above $200,000, and announce that the board will not meet again for several months. In both cases, my promise becomes credible because I have destroyed my ability to break it. The result is less flexibility—and more bargaining power.

*13.9 Auctions

In this section we examine **auction markets**—markets in which products are bought and sold through formal bidding processes.[21] Auctions come in all sizes and shapes. They are often used for differentiated products, especially unique items such as art, antiques, and the rights to produce oil from a piece of land. In recent years, for example, the U.S. Treasury has relied on auctions to sell Treasury bills, the Federal Communications Commission has used auctions for the sale of portions of the electromagnetic spectrum for cellular telephone services, and the Department of Defense has used auctions to procure military equipment. Auctions like these have important advantages: They are likely to be less time consuming than one-on-one bargaining, and they encourage competition among buyers in a way that increases the seller's revenue.

auction markets Markets in which products are bought and sold through formal bidding processes.

The design of an auction, which involves choosing the rules under which it operates, greatly affects its outcome. A seller will usually want an auction format that maximizes the revenue from the sale of the product. A buyer collecting bids from a group of potential sellers, on the other hand, will want an auction that minimizes the expected cost of the product to be purchased.

Auction Formats

We will see that the choice of auction format can affect the seller's auction revenue. Several different kinds of auction formats are widely used:

1. **Traditional English (or oral) auction:** The seller actively solicits progressively higher bids from a group of potential buyers. At each point, all participants are aware of the current high bid. The auction stops when no bidder is willing to surpass the current high bid; the item is then sold to the highest bidder at a price equal to the amount of the high bid.

English (or oral) auction Auction in which a seller actively solicits progressively higher bids from a group of potential buyers.

[21] There is a vast literature on auctions; for example, see Paul Milgrom, "Auctions and Bidding: A Primer," *Journal of Economic Perspectives* (Summer 1989): 3–22; John McMillan, *Games, Strategies and Managers* (New York: Oxford University Press, 1992); and Avinash Dixit and Susan Skeath, *Games of Strategy* (New York: Norton, 1999).

Dutch auction Auction in which a seller begins by offering an item at a relatively high price, then reduces it by fixed amounts until the item is sold.

sealed-bid auction Auction in which all bids are made simultaneously in sealed envelopes, the winning bidder being the individual who has submitted the highest bid.

first-price auction Auction in which the sales price is equal to the highest bid.

second-price auction Auction in which the sales price is equal to the second-highest bid.

2. **Dutch auction:** The seller begins by offering the item at a relatively high price. If no potential buyer agrees to that price, the seller reduces the price by fixed amounts. The first buyer that accepts an offered price can buy the item at that price.
3. **Sealed-bid auction:** All bids are made simultaneously in sealed envelopes, and the winning bidder is the individual who has submitted the highest bid. The price paid by the winning bidder will vary, however, depending on the rules of the auction. In a **first-price auction,** the sales price is equal to the highest bid. In a **second-price auction,** the sales price is equal to the second-highest bid.

Valuation and Information

Suppose you want to sell a distinctive and valuable product such as a painting or a rare coin. Which type of auction is best for you? The answer depends on the preferences of the bidders and the information available to them. We consider two cases:

Recall from §11.5 that the reservation price is the maximum amount of money that an individual will pay for a product.

private-value auction Auction in which each bidder knows his individual valuation of the object up for bid, with valuations differing from bidder to bidder.

common-value auction Auction in which the item has the same value to all bidders, but bidders do not know that value precisely and their estimates of it vary.

1. In **private-value auctions,** each bidder knows his or her individual valuation or *reservation price,* and valuations differ from bidder to bidder. In addition, each bidder is uncertain about the value that other bidders place on the product. For example, I might value a signed Mark McGwire home run baseball very highly but not know that you value it less highly.
2. In **common-value auctions,** the item to be auctioned has approximately the same value to all bidders. Bidders, however, do not know precisely what the value is—they can only estimate it, and bidders' estimates will vary. For example, in an auction of an offshore oil reserve, the value of the reserve is the price of oil minus the extraction cost, times the amount of oil in the reserve. As a result, the value should be about the same for all bidders. However, bidders will not know the amount of oil or the extraction cost—they can only estimate these numbers. Because their estimates will differ, they might bid very different amounts to get the reserve.

In principle, auctions can have both private-value and common-value elements. However, to simplify matters we will separate the two. We begin our discussion with private-value auctions and then move on to common-value auctions.

Private-Value Auctions

In private-value auctions, bidders have different reservation prices for the offered item. We might suppose, for example, that in an auction for the signed Mark McGwire baseball, individuals' reservation prices range from $1 (someone who doesn't like baseball but is bidding just for fun) to $600 (a St. Louis Cardinal fan). Of course, if you are bidding for the baseball, you don't know how many people will bid against you or what their bids will be.

Whatever the auction format, each bidder must choose his or her bidding strategy. For an open English auction, this strategy is a choice of a price at which to stop bidding. For a Dutch auction, the strategy is the price at which the individual expects to make his or her only bid. For a sealed-bid auction, the strategy is the choice of bid to place in a sealed envelope.

What are the payoffs in this bidding game? The payoff for winning is the difference between the winner's reservation price and the price paid; the payoff for losing is zero. Given these payoffs, let's examine bidding strategies and outcomes for different auction formats.

We will begin by showing that English oral auctions and second-price sealed-bid auctions generate nearly identical outcomes. Let's begin with the second-price sealed-bid auction. In this auction, bidding truthfully is a *dominant strategy*—there is no advantage to bidding below your reservation price. Why? Because the price you pay is based on the valuation of the *second highest bidder,* not on your own valuation. Suppose that your reservation price is $100. If you bid below your reservation price—say, $80—you risk losing to the second-highest bidder, who bids $85, when winning (at, say, $87) would have given you a positive payoff. If you bid above your reservation price—say $105—you risk winning but receiving a negative payoff.

Similarly, in an English auction the dominant strategy is to continue bidding a small amount—say $1—above the highest bidder *until the bidding reaches your reservation price.* Why? Because if you stop bidding at a point below your reservation price, you risk losing a positive payoff; if you continue beyond your reservation price, you will be guaranteed a negative payoff. How high will the bidding go? It will continue until the winning bid is $1 above the reservation price of the second highest bidder. Likewise, in the sealed-bid auction the winning bid will equal the reservation price of the second highest bidder. Thus, both auction formats generate nearly identical outcomes (the outcomes should differ in theory only by a dollar or two).

You know that as a seller, you should be indifferent between an oral English auction and a second-price sealed-bid auction, because bidders in each case have private values. Suppose that you plan to sell an item using a sealed-bid auction. Which should you choose, a first-price or a second-price auction? You might think that the first-price auction is better because the payment is given by the highest rather than the second-highest bid. Bidders, however, are aware of this reasoning and will alter their bidding strategies accordingly: They will bid less in anticipation of paying the winning bid if they are successful.

The second-price sealed-bid auction generates revenue equal to the second-highest reservation price. However, the revenue implications of a first-price sealed-bid auction for the seller are more complicated because the optimal strategy of bidders is more complex. The best strategy is to choose a bid that you believe will be equal to or slightly above the reservation price of the individual with the second-highest reservation price.[22] Why? Because the winner must pay

[22] This discussion assumes that the reservation prices of the bidders are statistically independent of one another. Suppose you and n other bidders have reservation prices that range uniformly from 0 to H—i.e., $H/(n+1), 2H/(n+1), \ldots (n-1)H/(n+1), nH/(n+1)$, and H. Statistical theory (which is beyond the scope of this book) tells us that the expected value of the second-highest reservation price is equal to $[(n)/(n+1)]H$. If your own reservation price is equal to V, your bid should be $[(n)/(n+1)]V$. Here's why: Each bidder wishes to maximize his expected profit, which is the product of (1) the probability of winning, p; and (2) $V - b$, the difference between the bidder's reservation price (V) and the bid (b). To evaluate p, consider the perspective of one bidder. Then, the probability that b will be a winning bid is the probability that b will be greater than the second-highest bid, which is equal to $b/[nH/(n+1)]$. The probability of winning falls as the number of bids increases. Specifically, the probability that bid b will be a winning bid if there are n other bidders is $p = \{b/[nH/(n+1)]\}^n$. Differentiating the expected profit, $(V - b)(p)$ with respect to b, equating to 0, and solving for b yields the rule of thumb, $b = [n/(n+1)]V$. More generally, if the bidding range varies between a low of L and a high of H, the equilibrium bidding strategy is to bid $b = \{L/(n+1) + [nH/(n+1)]\}$.

his or her bid, and it is never worth paying more than the second-highest reservation price. Thus, we see that the first-price and second-price sealed-bid auctions generate the same expected revenue.

Remember that while the *expected* revenue is the same under both auction formats, the revenue obtained can differ considerably in practice. We will see why when we study the common-value auction.

Common-Value Auctions

Suppose that you and four other people participate in an oral auction to purchase a large jar of pennies, which will go to the winning bidder at a price equal to the highest bid. Each bidder can examine the jar but cannot open it and count the pennies. Once you have estimated the number of pennies in the jar, what is your optimal bidding strategy? This is a classic common-value auction, because the jar of pennies has the same value for all bidders. The problem for you and other bidders is the fact that the value is unknown.

You might be tempted to do what many novices would do in this situation—bid up to your own estimate of the number of pennies in the jar, and no higher. This, however, is not the best way to bid. Remember that neither you nor the other bidders know the number of pennies for certain. All of you have independently made estimates of the number, and those estimates are subject to error—some will be too high and some too low. Who, then, will be the winning bidder? If each bidder bids up to his or her estimate, *the winning bidder is likely to be the person with the largest positive error*—i.e., the person with the largest overestimate of the number of pennies.

The Winner's Curse

To appreciate this possibility, suppose that there are actually 620 pennies in the jar. Let's say the bidders' estimates are 540, 590, 615, 650, and 690. Finally, suppose that you are the bidder whose estimate is 690, and that you win the auction with a bid of $6.80. Should you be happy about winning? No—you will have paid $6.80 for $6.20 worth of pennies. You will have fallen prey to the **winner's curse:** The winner of a common-value auction is often worse off than those who did not win because the winner was overly optimistic and, as a consequence, bid more for the item than it was actually worth.

winner's curse Situation in which the winner of a common-value auction is worse off as a consequence of overestimating the value of the item and thereby overbidding.

The winner's curse can arise in any common-value auction, and bidders often fail to take account of it. Suppose, for example, that your house needs to be painted, so you ask five painting companies to give you cost estimates for the job, telling each that you will accept the lowest estimate. Who will win the job? It will probably be the painter who has most seriously underestimated the amount of work involved. At first, that painter might be happy to have won the job, only later to realize that much more work is required than was anticipated. The same problem can arise for oil companies bidding for offshore oil reserves when the size of the reserve and cost of extraction are uncertain (so that the value of the reserve is uncertain). Unless the companies take the winner's curse into account, the winning bidder is likely to have won by overestimating the value of the reserve and will thus have paid more than the reserve is worth.

How should you take the winner's curse into account when bidding for an item in a common-value auction? You must not only estimate the value of the item you are bidding for, but also account for the fact that your estimate—and the estimates of the other bidders—are subject to error. To avoid the winner's curse, you must reduce your maximum bid below your value estimate by an amount equal to the expected error of the winning bidder. The more precise your estimate, the less you need to reduce your bid. If you can't accurately

assess the precision of your estimate directly, you can estimate the variation in the estimates of the other bidders. If there is a lot of disagreement among these bidders, it is likely that your estimate will be similarly imprecise. To measure the variation in bids, you can use the standard deviation of the estimates, which can be calculated using statistical methods.

Oil companies have been bidding for oil reserves for years, and thus are able to estimate this standard deviation quite well. They can thereby take the winner's curse into account by reducing their maximum bids below their value estimates by an amount equal to the expected error of the winning bidder. As a result, oil companies rarely feel they have made a mistake after winning an auction. Painters, on the other hand, are often less sophisticated in their bidding decisions and suffer from the winner's curse.

Maximizing Auction Revenue

Now let's return to the question of auction design from the point of view of the seller. Here are some useful tips for choosing the best auction format.

1. In a private-value auction, you should encourage as many bidders as possible: Additional bidders increase the expected bid of the winner and the expected valuation of the second-highest bidder as well.
2. In a common value auction, you should (a) use an open rather than a sealed-bid auction because as a general rule, an English (open) common-value auction will generate greater expected revenue than a sealed-bid auction; and (b) reveal information about the true value of the object being auctioned to reduce concern about the winner's curse, and consequently to encourage more bidding.

Why then use an open auction? Recall that in order to avoid the winner's curse, each bidder in a common value auction will bid below his individual valuation. The greater the uncertainty about the true value of the object, the greater the likelihood of an overbid, and therefore the greater the incentive for the bidder to reduce his bid. (If the bidder is risk-averse, this effect will be magnified.) However, the bidder faces less uncertainty in an English auction than in a sealed-bid auction because he can observe the prices at which other bidders drop out of the competition—an advantage that provides information about their valuations. In short, when you provide more information to bidders, risk-averse bidders will be encouraged to bid more because they will be more confident that they can account for the possibility of a winner's curse.

EXAMPLE 13.7 Internet Auctions

The popularity of auctions has skyrocketed in recent years with the growth of the Internet. Many Internet sites are now devoted to auctions at which participants can buy and sell a wide variety of items. Let's see how these Internet auctions work.

One of the most popular Internet auction sites is www.ebay.com. It conducts many auctions each day for items ranging from antiques and automobiles to Beanie Babies and Pokemon cards. In 1998, over 34 million individual auctions took place, and the average winning bid was $40. Any person over the age of 18 can participate as a buyer at no cost; sellers typically pay a percentage of the sale price to eBay. Two auction formats are used on eBay: (1) an increasing-bid auction for a single item in which the highest bidder at the close of the auction wins and pays a price equal to the second-highest bid to the seller; and (2) an increasing-bid auction for several identical items in which the highest n bidders win the n items sold. In both auctions, ties are broken by awarding the item to the buyer who bid first. Notice that neither of these auctions corresponds precisely to any of the four auction formats discussed above. The first approximates the standard English auction, but the existence of a fixed and known stopping time can cause bidders to place bids strategically at the end of the auction. The second is called a "Dutch" auction by eBay but differs from a conventional Dutch auction in two respects: bids are increasing rather than decreasing and the auction has a fixed and known stopping time. In both auction formats, sellers can impose a minimum acceptable bid—called a *reserve price*—and although buyers know a reserve price exists, they generally are not told what it is.

Many Internet auctions are dominated by private-value items. (However, because anyone can put an item up for sale, there is a common-value issue—how reliable is the seller?) The private-value emphasis of these auctions is especially true of unique antiques that may have considerable value to particular bidders. With private-value auctions you needn't worry so much about the prior history of bidding: The bids of others tell you about their preferences, but the value that you place on the object is personal to you. Although you want to win the bidding at a price as far below your valuation as possible, the winner's curse needn't be a concern: You can't be disappointed if your value for the object is more than what you paid for it.

Finally, a few caveats are in order when buying items via Internet auctions. Unlike traditional auction houses, low-end auction sites like eBay only provide a forum for buyers and sellers to interact; they provide no quality-control functions. While many sites, including eBay, make available feedback from buyers for each seller, this is usually the only evidence of a seller's reliability that buyers receive. Furthermore, there is obviously no feedback available for first-time sellers (or for sellers who have recently changed their eBay user names). In addition, the possibility of bid manipulation looms large in Internet auctions. At eBay, for example, a valid e-mail address is all that is required for a buyer to bid on an item. Given the relative ease of obtaining e-mail addresses (hundreds of Internet sites now offer free e-mail for the price of registering with the site), sellers may file spurious bids in order to manipulate the bidding process. For example, a seller of a common-value item could actually exacerbate the winner's curse problem by filing fictitious bids that cause buyers to increase their valuations for an item. Thus, caveat emptor is a sound philosophy when buying items on the Internet.

SUMMARY

1. A game is cooperative if the players can communicate and arrange binding contracts; otherwise it is noncooperative. In either kind of game, the most important aspect of strategy design is understanding your opponent's position, and (if your opponent is rational) correctly deducing the likely response to your actions. Misjudging an opponent's position is a common mistake, as Example 13.1, "Acquiring a Company," illustrates.[23]
2. A Nash equilibrium is a set of strategies such that all players are doing their best given the strategies of the other players. An equilibrium in dominant strategies is a special case of a Nash equilibrium; a dominant strategy is optimal no matter what the other players do. A Nash equilibrium relies on the rationality of each player. A maximin strategy is more conservative because it maximizes the minimum possible outcome.
3. Some games have no Nash equilibria in pure strategies but have one or more equilibria in mixed strategies. A mixed strategy is one in which the player makes a random choice among two or more possible actions, based on a set of chosen probabilities.
4. Strategies that are not optimal for a one-shot game may be optimal for a repeated game. Depending on the number of repetitions, a "tit-for-tat" strategy, in which you play cooperatively as long as your competitor does the same, may be optimal for the repeated prisoners' dilemma.
5. In a sequential game, the players move in turn. In some cases, the player who moves first has an advantage. Players may then have an incentive to try to precommit themselves to particular actions before their competitors can do the same.
6. An empty threat is a threat that one has no incentive to carry out. If one's competitors are rational, empty threats are of no value. To make a threat credible, it is sometimes necessary to make a strategic move to constrain one's later behavior, thereby creating an incentive to carry out the threat.
7. To deter entry, an incumbent firm must convince any potential competitor that entry will be unprofitable. This may be done by investing, and thereby giving credibility to the threat that entry will be met by price warfare. Strategic trade policies by governments sometimes have this objective.
8. Bargaining situations are examples of cooperative games. As in noncooperative games, in bargaining players can sometimes gain a strategic advantage by limiting their own flexibility.
9. Auctions can be conducted in a number of formats, including English (oral with increasing bids), Dutch (oral with decreasing bids), and sealed bid. The opportunity for a seller to raise revenue and for a buyer to obtain an object at a reasonable price depends on the auction format, and on whether the items being auctioned have the same value to all bidders (as in a common-value auction) or different values to different bidders (as in a private-value auction).

QUESTIONS FOR REVIEW

1. What is the difference between a cooperative and a noncooperative game? Give an example of each.
2. What is a dominant strategy? Why is an equilibrium stable in dominant strategies?
3. Explain the meaning of a Nash equilibrium. How does it differ from an equilibrium in dominant strategies?
4. How does a Nash equilibrium differ from a game's maximin solution? In what situations is a maximin solution a more likely outcome than a Nash equilibrium?
5. What is a "tit-for-tat" strategy? Why is it a rational strategy for the infinitely repeated prisoners' dilemma?
6. Consider a game in which the prisoners' dilemma is repeated 10 times and both players are rational and fully informed. Is a tit-for-tat strategy optimal in this case? Under what conditions would such a strategy be optimal?
7. Suppose you and your competitor are playing the pricing game shown in Table 13.8. Both of you must announce your prices at the same time. Can you improve your outcome by promising your competitor that you will announce a high price?
8. What is meant by "first-mover advantage"? Give an example of a gaming situation with a first-mover advantage.

[23] Here is the solution to Company A's problem: *It should offer nothing for Company T's stock.* Remember that Company T will accept an offer only if it is greater than the per-share value under current management. Suppose you offer \$50. Thus Company T will accept this offer only if the outcome of the exploration project results in a per-share value under current management of \$50 or less. Any values between \$0 and \$100 are equally likely. Therefore the *expected value* of Company T's stock, *given that it accepts the offer*—i.e., given that the outcome of the exploration project leads to a value less than \$50—is \$25, so that under the management of Company A the value would be (1.5)(\$25) = \$37.5, which is less than \$50. In fact, for any price P, if the offer is accepted, Company A can expect a value of only $(3/4)P$.

9. What is a "strategic move"? How can the development of a certain kind of reputation be a strategic move?
10. Can the threat of a price war deter entry by potential competitors? What actions might a firm take to make this threat credible?
11. A strategic move limits one's flexibility and yet gives one an advantage. Why? How might a strategic move give one an advantage in bargaining?
12. Why is the winner's curse potentially a problem for a bidder in a common-value auction but not in a private-value auction?

EXERCISES

1. In many oligopolistic industries, the same firms compete over a long period of time, setting prices and observing each other's behavior repeatedly. Given that the number of repetitions is large, why don't collusive outcomes typically result?
2. Many industries are often plagued by overcapacity—firms simultaneously make major investments in capacity expansion, so that total capacity far exceeds demand. This happens not only in industries in which demand is highly volatile and unpredictable, but also in industries in which demand is fairly stable. What factors lead to overcapacity? Explain each briefly.
3. Two computer firms, A and B, are planning to market network systems for office information management. Each firm can develop either a fast, high-quality system (High), or a slower, low-quality system (Low). Market research indicates that the resulting profits to each firm for the alternative strategies are given by the following payoff matrix:

		FIRM B: High	FIRM B: Low
FIRM A	High	30, 30	50, 35
	Low	40, 60	20, 20

a. If both firms make their decisions at the same time and follow *maximin* (low-risk) strategies, what will the outcome be?
b. Suppose both firms try to maximize profits, but Firm A has a head start in planning and can commit first. Now what will the outcome be? What will the outcome be if Firm B has the head start in planning and can commit first?
c. Getting a head start costs money (you have to gear up a large engineering team). Now consider the *two-stage* game in which *first*, each firm decides how much money to spend to speed up its planning, and *second*, it announces which product (H or L) it will produce. Which firm will spend more to speed up its planning? How much will it spend? Should the other firm spend *anything* to speed up its planning? Explain.

4. Two firms are in the chocolate market. Each can choose to go for the high end of the market (high quality) or the low end (low quality). Resulting profits are given by the following payoff matrix:

		FIRM 2: Low	FIRM 2: High
FIRM 1	Low	−20, −30	900, 600
	High	100, 800	50, 50

a. What outcomes, if any, are Nash equilibria?
b. If the managers of both firms are conservative and each follows a maximin strategy, what will be the outcome?
c. What is the cooperative outcome?
d. Which firm benefits most from the cooperative outcome? How much would that firm need to offer the other to persuade it to collude?

5. Two major networks are competing for viewer ratings in the 8:00–9:00 P.M. and 9:00–10:00 P.M. slots on a given weeknight. Each has two shows to fill this time period and is juggling its lineup. Each can choose to put its "bigger" show first or to place it second in the 9:00–10:00 P.M. slot. The combination of decisions leads to the following "ratings points" results:

		NETWORK 2: First	NETWORK 2: Second
NETWORK 1	First	18, 18	23, 20
	Second	4, 23	16, 16

a. Find the Nash equilibria for this game, assuming that both networks make their decisions at the same time.

b. If each network is risk-averse and uses a maximin strategy, what will be the resulting equilibrium?
c. What will be the equilibrium if Network 1 makes its selection first? If Network 2 goes first?
d. Suppose the network managers meet to coordinate schedules and Network 1 promises to schedule its big show first. Is this promise credible? What would be the likely outcome?

6. Two competing firms are each planning to introduce a new product. Each will decide whether to produce Product A, Product B, or Product C. They will make their choices at the same time. The resulting payoffs are shown below.

		FIRM 2: A	FIRM 2: B	FIRM 2: C
FIRM 1	A	−10, −10	0, 10	10, 20
	B	10, 0	−20, −20	−5, 15
	C	20, 10	15, −5	−30, −30

a. Are there any Nash equilibria in pure strategies? If so, what are they?
b. If both firms use maximin strategies, what outcome will result?
c. If Firm 1 uses a maximin strategy and Firm 2 knows this, what will Firm 2 do?

7. We can think of U.S. and Japanese trade policies as a prisoners' dilemma. The two countries are considering policies to open or close their import markets. The payoff matrix is shown below.

		JAPAN: Open	JAPAN: Close
U.S.	Open	10, 10	5, 5
	Close	−100, 5	1, 1

a. Assume that each country knows the payoff matrix and believes that the other country will act in its own interest. Does either country have a dominant strategy? What will be the equilibrium policies if each country acts rationally to maximize its welfare?
b. Now assume that Japan is not certain that the United States will behave rationally. In particular, Japan is concerned that U.S. politicians may want to penalize Japan even if that does not maximize U.S. welfare. How might this affect Japan's choice of strategy? How might this change the equilibrium?

8. You are a duopolist producer of a homogeneous good. Both you and your competitor have zero marginal costs. The market demand curve is

$$P = 30 - Q$$

where $Q = Q_1 + Q_2$. Q_1 is your output and Q_2 your competitor's output. Your competitor has also read this book.

a. Suppose you will play this game only once. If you and your competitor must announce your outputs at the same time, how much will you choose to produce? What do you expect your profit to be? Explain.
b. Suppose you are told that you must announce your output *before* your competitor does. How much will you produce in this case, and how much do you think your competitor will produce? What do you expect your profit to be? Is announcing first an advantage or a disadvantage? Explain briefly. How much would you pay to be given the option of announcing either first or second?
c. Suppose instead that you are to play the first round of a series of 10 rounds (with the same competitor). In each round, you and your competitor announce your outputs at the same time. You want to maximize the sum of your profits over the 10 rounds. How much will you produce in the first round? How much do you expect to produce in the tenth round? In the ninth round? Explain briefly.
d. Once again you will play a series of 10 rounds. This time, however, in each round your competitor will announce its output before you announce yours. How will your answers to (c) change in this case?

9. You play the following bargaining game. Player A moves first and makes Player B an offer for the division of \$100. (For example, Player A could suggest that she take \$60 and Player B take \$40). Player B can accept or reject the offer. If he rejects it, the amount of money available drops to \$90, and he then makes an offer for the division of this amount. If Player A rejects this offer, the amount of money drops to \$80 and Player A makes an offer for its division. If Player B rejects this offer, the amount of money drops to 0. Both players are rational, fully informed, and want to maximize their payoffs. Which player will do best in this game?

*10. Defendo has decided to introduce a revolutionary video game. As the first firm in the market, it will have a monopoly position for at least some time. In deciding what type of manufacturing plant to build, it has the choice of two technologies. Technology A is publicly available and will result in annual costs of

$$C^A(q) = 10 + 8q$$

Technology B is a proprietary technology developed in Defendo's research labs. It involves higher fixed cost of production but lower marginal costs:

$$C^B(q) = 60 + 2q$$

Defendo must decide which technology to adopt. Market demand for the new product is $P = 20 - Q$, where Q is total industry output.

a. Suppose Defendo were certain that it would maintain its monopoly position in the market for the entire product lifespan (about five years) without threat of entry. Which technology would you advise Defendo to adopt? What would be Defendo's profit given this choice?

b. Suppose Defendo expects its archrival, Offendo, to consider entering the market shortly after Defendo introduces its new product. Offendo will have access only to Technology A. If Offendo does enter the market, the two firms will play a Cournot game (in quantities) and arrive at the Cournot-Nash equilibrium.

   i. If Defendo adopts Technology A and Offendo enters the market, what will be the profit of each firm? Would Offendo choose to enter the market given these profits?

   ii. If Defendo adopts Technology B and Offendo enters the market, what will be the profit of each firm? Would Offendo choose to enter the market given these profits?

   iii. Which technology would you advise Defendo to adopt given the threat of possible entry? What will be Defendo's profit given this choice? What will be consumer surplus given this choice?

c. What happens to social welfare (the sum of consumer surplus and producer profit) as a result of the threat of entry in this market? What happens to equilibrium price? What might this imply about the role of *potential* competition in limiting market power?

11. Three contestants, A, B, and C, each have a balloon and a pistol. From fixed positions, they fire at each other's balloons. When a balloon is hit, its owner is out. When only one balloon remains, its owner is the winner of a \$1000 prize. At the outset, the players decide by lot the order in which they will fire, and each player can choose any remaining balloon as his target. Everyone knows that A is the best shot and always hits the target, that B hits the target with probability .9, and that C hits the target with probability .8. Which contestant has the highest probability of winning the \$1000? Explain why.

12. An antique dealer regularly buys objects at hometown auctions whose bidders are limited to other dealers. Most of her successful bids turn out to be financially worthwhile, because she is able to resell the antiques for a profit. On occasion, however, she travels to a nearby town to bid in an auction that is open to the public. She often finds that on the rare occasions in which she does bid successfully, she is disappointed—the antique cannot be sold at a profit. Can you explain the difference in her success between the two sets of circumstances?

13. You are in the market for a new house and have decided to bid for a house at auction. You believe that the value of the house is between \$125,000 and \$150,000, but you are uncertain as to where in the range it might be. You do know, however, that the seller has reserved the right to withdraw the house from the market if the winning bid is not satisfactory.

   a. Should you bid in this auction? Why or why not?

   b. Suppose you are a building contractor. You plan to improve the house and then to resell it at a profit. How does this situation affect your answer to (a)? Does it depend on the extent to which your skills are uniquely suitable to improving this particular house?

CHAPTER 14

Markets for Factor Inputs

Chapter Outline

List of Examples

So far we have concentrated on *output markets*: markets for goods and services that firms sell and consumers purchase. In this chapter, we discuss *factor markets*: markets for labor, raw materials, and other inputs to production. Much of our material will be familiar because the same forces that shape supply and demand in output markets also affect factor markets.

We have seen that some output markets are perfectly or almost perfectly competitive, while producers in others have market power. The same is true for factor markets. We will examine three different factor market structures:

1. Perfectly competitive factor markets;
2. Markets in which buyers of factors have monopsony power;
3. Markets in which sellers of factors have monopoly power.

We will also point out instances in which equilibrium in the factor market depends on the extent of market power in *output* markets.

14.1 Competitive Factor Markets

A competitive *factor market* is one in which there are a large number of sellers and buyers of a factor of production, such as labor or raw materials. Because no single seller or buyer can affect the price of a given factor, each is a price taker. For example, if individual firms that buy lumber to construct homes purchase a small share of the total volume of lumber available, their purchasing decision will have no effect on price. Likewise, if each supplier of lumber controls a small share of the market, no individual supplier's decision will affect the price of the lumber that he sells.

We begin by analyzing the demands for a factor by individual firms. These demands are added to get market demand. We then shift to the supply side of the market and show how market price and input levels are determined.

Demand for a Factor Input When Only One Input Is Variable

derived demand Demand for an input that depends on, and is derived from, both the firm's level of output and the cost of inputs.

Like demand curves for the final goods that result from the production process, demand curves for factors of production are downward sloping. Unlike consumers' demands for goods and services, however, factor demands are **derived demands:** they depend on, and are derived from, the firm's level of output and the costs of inputs. For example, the demand of the Microsoft Corporation for computer programmers is a derived demand that depends not only on the current salaries of programmers, but also on how much software Microsoft expects to sell.

To analyze factor demands, we will use the material from Chapter 7 that shows how a firm chooses its production inputs. We will assume that the firm produces its output using two inputs, capital K and labor L, that can be hired at the prices r (the rental cost of capital) and w (the wage rate), respectively.[1] We will also assume that the firm has its plant and equipment in place (as in a short-run analysis) and must only decide how much labor to hire.

Suppose that the firm has hired a certain number of workers and wants to know whether it is profitable to hire one additional worker. This will be profitable if the additional revenue from the output of the worker's labor is greater than its cost. The additional revenue from an incremental unit of labor, the **marginal revenue product of labor,** is denoted MRP_L. We will show that the firm should hire more labor if the MRP_L is at least as large as the wage rate w.

marginal revenue product Additional revenue resulting from the sale of output created by the use of one additional unit of an input.

How do we measure the MRP_L? It's *the additional output obtained from the additional unit of this labor, multiplied by the additional revenue from an extra unit of output.* The additional output is given by the marginal product of labor MP_L and the additional revenue by the marginal revenue MR.

Recall that in §8.2, marginal revenue is defined to be the increase in revenue resulting from a one-unit increase in output.

How do we measure the MRP_L? MRP_L is measured by the *additional output obtained from the additional unit of this labor, multiplied by the additional revenue from an extra unit of output*. Formally, the marginal revenue product is $\Delta R/\Delta L$, where L is the number of units of labor input and R is revenue. The additional output per unit of labor, the MP_L, is given by $\Delta Q/\Delta L$, and marginal revenue, MR, is equal to $\Delta R/\Delta Q$. Because $\Delta R/\Delta L = (\Delta R)/(\Delta Q)(\Delta Q/\Delta L)$, it follows that

$$MRP_L = (MP_L)(MR) \tag{14.1}$$

This important result holds for any competitive factor market, whether the output market is competitive or not. However, to examine the characteristics of the MRP_L, let's begin with the case of a perfectly competitive output (and input) market. In a competitive output market, a firm will sell all its output at the market price P. The marginal revenue from the sale of an additional unit of output is then equal to P. In this case, the marginal revenue product of labor is equal to the marginal product of labor times the price of the product:

In §8.2, we explain that because the demand facing each firm in a competitive market is perfectly elastic, each firm will sell its output at a price equal to its average revenue and to its marginal revenue.

$$MRP_L = (MP_L)(P) \tag{14.2}$$

The higher of the two curves in Figure 14.1 represents the MRP_L curve for a firm in a competitive output market. Note that because there are diminishing marginal returns to labor, the marginal product of labor falls as the amount of labor increases. The marginal revenue product curve thus slopes downward, even though the price of the output is constant.

[1] We implicitly assume that all inputs to production are identical in quality. Differences in workers' skills and abilities are discussed in Chapter 17

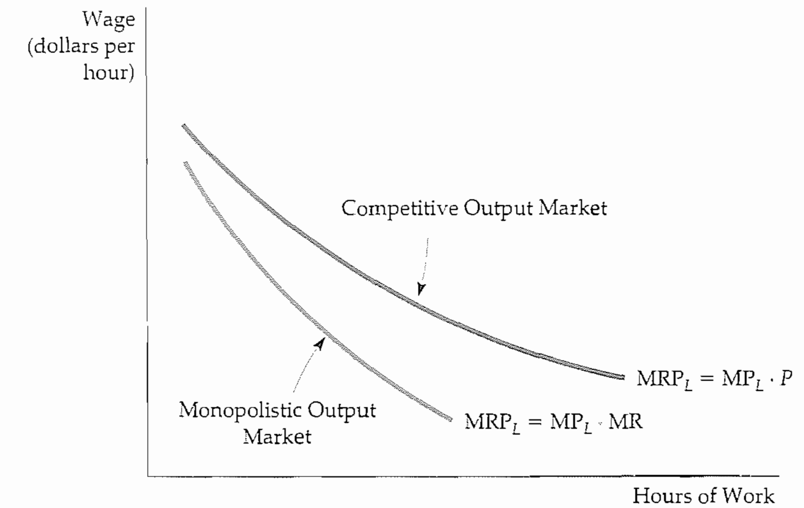

FIGURE 14.1 Marginal Revenue Product

In a competitive factor market in which the producer is a price taker, the buyer's demand for an input is given by the marginal revenue product curve. The MRP curve falls because the marginal product of labor falls as hours of work increase. When the producer of the product has monopoly power, the demand for the input is also given by the MRP curve. In this case, however, the MRP curve falls because both the marginal product of labor and marginal revenue fall.

The lower curve in Figure 14.1 is the MRP_L curve when the firm has monopoly power in the output market. When firms have monopoly power, they face a downward-sloping demand curve, and therefore must lower the price of all units of the product in order to sell more of it. As a result, marginal revenue is always less than price ($MR < P$). This explains why the monopolistic curve lies below the competitive curve, and why marginal revenue falls as output increases. Thus the marginal revenue product curve slopes downward in this case because the marginal revenue curve *and* the marginal product curve slope downward.

Note that the marginal revenue product tells us how much the firm should be willing to pay to hire an additional unit of labor. As long as the MRP_L is greater than the wage rate, the firm should hire an additional unit of labor. If the marginal revenue product is less than the wage rate, the firm should lay off workers. Only when the marginal revenue product is equal to the wage rate will the firm have hired the profit-maximizing amount of labor. The profit-maximizing condition is therefore

$$MRP_L = w \tag{14.3}$$

Figure 14.2 illustrates this condition. The demand for labor curve D_L is the MRP_L. Note that the quantity of labor demanded increases as the wage rate falls. Because the labor market is perfectly competitive, the firm can hire as many

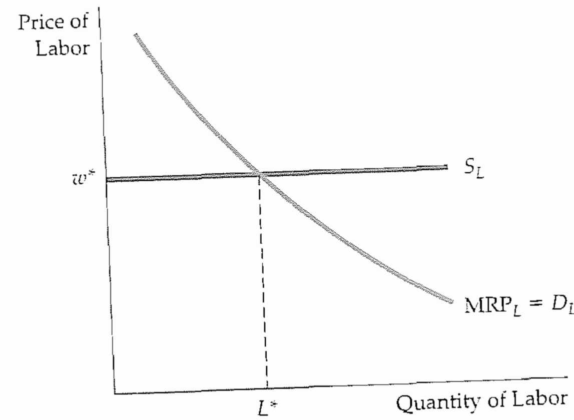

FIGURE 14.2 Hiring by a Firm in the Labor Market (with Fixed Capital)

In a competitive labor market, a firm faces a perfectly elastic supply of labor S_L and can hire as many workers as it wants at a wage rate w^*. The firm's demand for labor D_L is given by its marginal revenue product of labor MRP_L. The profit-maximizing firm will hire L^* units of labor at the point where the marginal revenue product of labor is equal to the wage rate.

workers as it wants at the market wage w^*. The supply of labor curve facing the firm, S_L, is thus a horizontal line. The profit-maximizing amount of labor that the firm hires, L^*, is at the intersection of the supply and demand curves.

Figure 14.3 shows how the quantity of labor demanded changes in response to a drop in the market wage rate from w_1 to w_2. The wage rate might decrease if more people entering the labor force are looking for jobs for the first time (as happened, for example, when all baby boomers came of age). The quantity of labor demanded by the firm is initially L_1, at the intersection of MRP_L and S_1. However, when the supply of labor curve shifts from S_1 to S_2, the wage falls from w_1 to w_2 and the quantity of labor demanded increases from L_1 to L_2.

Factor markets are similar to output markets in many ways. For example, the factor market profit-maximizing condition that the marginal revenue product of labor be equal to the wage rate is analogous to the output market condition that marginal revenue be equal to marginal cost. To see why this is true, recall that $MRP_L = (MP_L)(MR)$ and divide both sides of equation (14.3) by the marginal product of labor. Then,

$$MR = w/MP_L \tag{14.4}$$

Because MP_L measures additional output per unit of input, the right-hand side of equation (14.4) measures the marginal cost of an additional unit of output (the wage rate multiplied by the labor needed to produce one unit of output). Equation (14.4) shows that *both the hiring and output choices of the firm follow the same rule: inputs or outputs are chosen so that marginal revenue (from the sale of output) is equal to marginal cost (from the purchase of inputs)*. This principle holds in both competitive and noncompetitive markets.

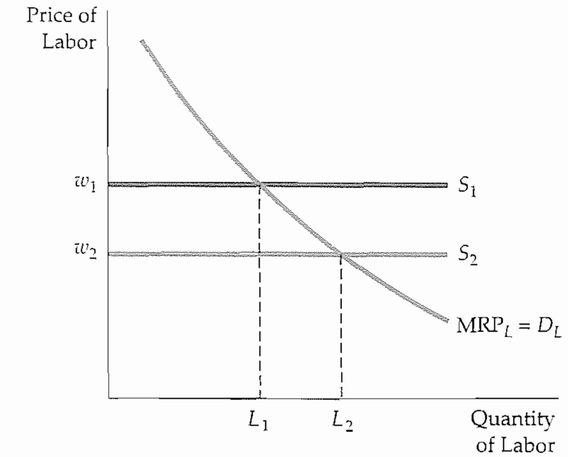

FIGURE 14.3 A Shift in the Supply of Labor

When the supply of labor facing the firm is S_1, the firm hires L_1 units of labor at wage w_1. But when the market wage rate decreases and the supply of labor shifts to S_2, the firm maximizes its profit by moving along the demand for labor curve until the new wage rate w_2 is equal to the marginal revenue product of labor. As a result, L_2 units of labor are hired.

Demand for a Factor Input When Several Inputs Are Variable

When the firm simultaneously chooses quantities of two or more variable inputs, the hiring problem becomes more difficult because a change in the price of one input will change the demand for others. Suppose, for example, that both labor and assembly-line machinery are variable inputs to producing farm equipment. Let's say that we wish to determine the firm's demand for labor curve. As the wage rate falls, more labor will be demanded, even if the firm's investment in machinery is unchanged. But as labor becomes less expensive, the marginal cost of producing the farm equipment falls. Consequently, it is profitable for the firm to increase its output. In that case, the firm is likely to invest in additional machinery to expand production capacity. Expanding the use of machinery causes the marginal revenue product of labor curve to shift to the right; in turn, the quantity of labor demanded increases.

Figure 14.4 illustrates this. Suppose that when the wage rate is $20 per hour, the firm hires 100 worker-hours, as shown by point *A* on the MRP_{L1} curve. Now consider what happens when the wage rate falls to $15 per hour. Because the marginal revenue product of labor is now greater than the wage rate, the firm will demand more labor. But the MRP_{L1} curve describes the demand for labor when the use of machinery is fixed. In fact, a greater amount of labor causes the marginal product of *capital* to rise, which encourages the firm to rent more machinery as well as hire more labor. Because there is more machinery, the marginal product of labor will increase (with more machinery, workers can be more productive). The marginal revenue product curve will therefore shift to the right (to MRP_{L2}). Thus, when the wage rate falls, the firm will use 140 hours of labor. This is shown by a new point on the demand curve, *C*, rather than 120 hours as

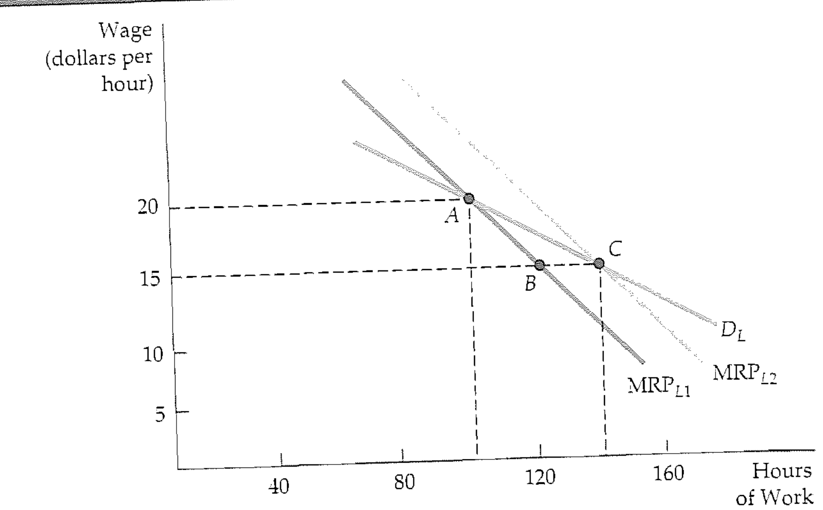

FIGURE 14.4 Firm's Demand Curve for Labor (with Variable Capital)

When two or more inputs are variable, a firm's demand for one input depends on the marginal revenue product of both inputs. When the wage rate is \$20, *A* represents one point on the firm's demand for labor curve. When the wage rate falls to \$15, the marginal product of capital rises, encouraging the firm to rent more machinery and hire more labor. As a result, the MRP curve shifts from MRP_{L1} to MRP_{L2}, generating a new point *C* on the firm's demand for labor curve. Thus *A* and *C* are on the demand for labor curve, but *B* is not.

given by *B*. *A* and *C* are both on the firm's demand for labor curve (with machinery variable) D_L; *B* is not. Note that as constructed, the demand for labor curve is more elastic than either of the two marginal product of labor curves (which presume no change in the amount of machinery). Thus, when capital inputs are variable in the long run, there is a greater elasticity of demand because firms can substitute capital for labor in the production process.

The Market Demand Curve

Recall from §4.3 that the market demand curve for a product shows how much of the product consumers are willing to buy as the price of the product changes.

When we aggregated the individual demand curves of consumers to obtain the market demand curve for a product, we were concerned with a single industry. However, a factor input like skilled labor is demanded by firms in many different industries. To obtain the total market demand for labor curve, we must therefore first determine each industry's demand for labor, and then add the industry demand curves horizontally. The second step is straightforward. Adding industry demand curves for labor to obtain a market demand curve for labor is just like adding individual product demand curves to obtain the market demand curve for that product. So let's concentrate our attention on the more difficult first step.

Determining Industry Demand The first step—determining industry demand—takes into account the fact that both the level of output produced by the firm and its product price change as the prices of the inputs to production change. It is easiest to determine market demand when there is a single producer.

In that case, the marginal revenue product curve is the industry demand curve for the input. When there are many firms, however, the analysis is more complex because of the possible interaction among the firms. Consider, for instance, the demand for labor when output markets are perfectly competitive. Then, the marginal revenue product of labor is the product of the price of the good and the marginal product of labor (see equation 14.2), as shown by the curve MRP_{L1} in Figure 14.5.

Suppose initially that the wage rate for labor is \$15 per hour and that the firm demands 100 worker-hours of labor. Now the wage rate for this firm falls to \$10 per hour. If no other firms could hire workers at the lower wage, then our firm would hire 150 worker-hours of labor (by finding the point on the MRP_{L1} curve that corresponds to the \$10-per-hour wage rate). But if the wage rate falls for all firms in an industry, the industry as a whole will hire more labor. This will lead to more output from the industry, a shift to the right of the industry supply curve, and a lower market price for its product.

In Figure 14.5(a), when the product price falls, the original marginal revenue product curve shifts downward, from MRP_{L1} to MRP_{L2}. This shift results in a lower quantity of labor demanded by the firm—120 worker-hours rather than 150. Consequently, industry demand for labor will be lower than if only one firm were able to hire workers at the lower wage. Figure 14.5(b) illustrates this. The lighter line shows the horizontal sum of the individual firms' demands for labor that would result if product price did not change as the wage falls. The darker line shows the industry demand curve for labor, which takes into account the fact that product price will fall as all firms expand their output in response to the lower wage rate. When the wage rate is \$15 per hour, industry demand for labor is L_0 worker-hours. When it falls to \$10 per hour, industry demand

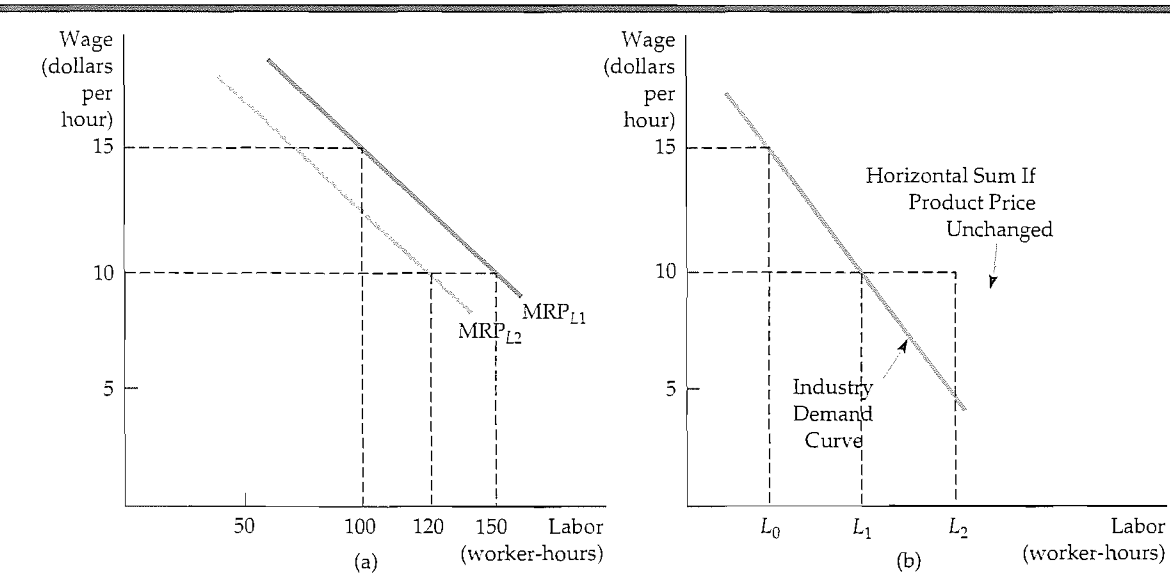

FIGURE 14.5 The Industry Demand for Labor

The demand curve for labor of a competitive firm, MRP_{L1} in (a), takes the product price as given. But as the wage rate falls from \$15 to \$10 per hour, the product price also falls. Thus the firm's demand curve shifts downward to MRP_{L2}. As a result, the industry demand curve, shown in (b), is more inelastic than the demand curve that would be obtained if the product price were assumed to be unchanged.

increases to L_1. Note that this is a smaller increase than L_2, which would occur if the product price were fixed. The aggregation of industry demand curves into the market demand curve for labor is the final step: To complete it we simply add the labor demanded in all industries.

The derivation of the market demand curve for labor (or for any other input) is essentially the same when the output market is noncompetitive. The only difference is that it is more difficult to predict the change in product price in response to a change in the wage rate because each firm in the market is likely to be pricing strategically rather than taking price as given.

EXAMPLE 14.1 The Demand for Jet Fuel

Throughout the 1970s and the early 1980s, fuel costs for U.S. airlines increased rapidly, in tandem with rising world oil prices. For example, whereas fuel costs made up 12.4 percent of total operating costs in 1971, its share of operating costs rose to about 30 percent in 1980. As we would expect, the amount of jet fuel used by airlines during this period fell as its price rose. Thus the output of the airline industry, as measured by the number of ton-miles, rose by 29.6 percent, while the amount of jet fuel consumed increased by only 8.8 percent. (One ton-mile is short for one ton of passengers, baggage, or freight transported one mile.) Jet fuel's share of operating costs fell during the late 1980s as oil prices fell, but it remained significant, still topping 15 percent during the 1990s.

Understanding the demand for jet fuel is important to managers of oil refineries, who must decide how much jet fuel to produce. It is also crucial to managers of airlines, who must project fuel purchases and costs when fuel prices rise.[2]

The effect of the increase in fuel costs on the airline industry depends on the ability of airlines either to cut fuel usage by reducing weight (by carrying less excess fuel) and flying slower (reducing drag and increasing engine efficiency) or to pass on their higher costs in customer prices. Thus the price elasticity of demand for jet fuel depends both on the ability to conserve fuel and on the elasticities of demand and supply of travel.

In §4.4, we define the price elasticity of demand as the percentage change in quantity demanded resulting from a 1-percent change in the price of a good.

To measure the short-run elasticity of demand for jet fuel, we use as the quantity of fuel demanded the number of gallons of fuel used by an airline in all markets within its domestic route network. The price of jet fuel is measured in dollars per gallon. A statistical analysis of demand must control for factors other than price that can explain why some firms demand more fuel than others. Some airlines, for example, use more fuel-efficient jet aircraft than others. A second factor is the length of the flights: The shorter the flight, the more fuel consumed per mile of travel. Both these factors were included in a statistical analysis that relates the quantity of fuel demanded to its price. Table 14.1 shows some short-run price elasticities. (They do not account for the introduction of new types of aircraft.)

TABLE 14.1 Short-Run Price Elasticity of Demand for Jet Fuel

AIRLINE	ELASTICITY	AIRLINE	ELASTICITY
American	−.06	Delta	−.15
Continental	−.09	TWA	−.10
Northwest	−.07	United	−.10

The jet fuel price elasticities for the airlines range in value from −.06 (for American) to −.15 (for Delta). Overall, the results show that the demand for jet fuel as an input to the production of airline flight-miles is very inelastic. This finding is not surprising: in the short run, there is no good substitute for jet fuel. The long-run elasticity of demand is higher, however, because airlines can eventually introduce more energy-efficient airplanes.

Figure 14.6 shows the short- and long-run demands for jet fuel. The short-run demand curve, MRP_{SR}, is much less elastic than the long-run demand curve because it takes time to substitute newer, more fuel-efficient airplanes for other planes when the price of fuel goes up.

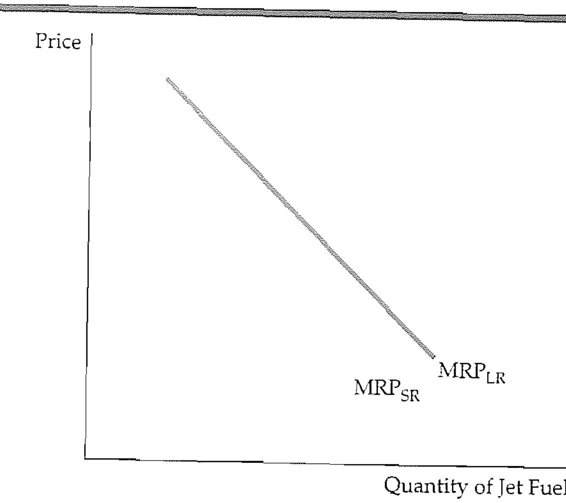

FIGURE 14.6 The Short- and Long-Run Demand for Jet Fuel

The short-run demand for jet fuel MRP_{SR} is more inelastic than the long-run demand MRP_{LR}. In the short run, airlines cannot reduce fuel consumption much when fuel prices increase. In the long run, however, they can take longer, more fuel-efficient routes and put more fuel-efficient planes into service.

The Supply of Inputs to a Firm

When the market for a factor input is perfectly competitive, a firm can purchase as much of that input as it wants at a fixed market price, which is determined by the intersection of the market demand and supply curves, as shown in Figure 14.7(a). The input supply curve facing a firm is then perfectly elastic. Thus, in Figure 14.7(b), a firm is buying fabric at \$10 per yard to sew into clothing. Because the firm is only a small part of the fabric market, it can buy all it wants without affecting the price.

[2] This example is drawn in part from Joseph M. Cigliano, "The Demand for Jet Fuel by the U.S. Domestic Trunk Airlines," *Business Economics* (September 1982): 32–36.

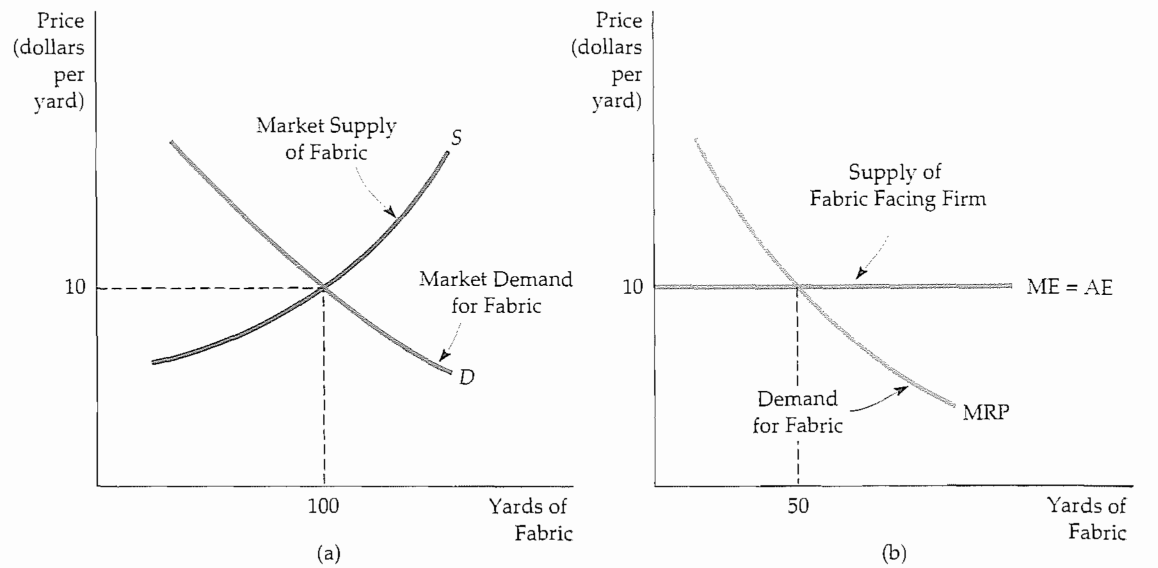

FIGURE 14.7 A Firm's Input Supply in a Competitive Factor Market

In a competitive factor market, a firm can buy any amount of the input it wants without affecting the price. Therefore the firm faces a perfectly elastic supply curve for that input. As a result, the quantity of the input purchased by the producer of the product is determined by the intersection of the input demand and supply curves. In (a) the industry quantity demanded and quantity supplied of fabric are equated at a price of $10 per yard. In (b) the firm faces a horizontal marginal expenditure curve at a price of $10 per yard of fabric, and chooses to buy 50 yards.

average expenditure curve Supply curve representing the price per unit that a firm pays for a good.

marginal expenditure curve Curve describing the incremental cost of purchasing one additional unit of a good.

Recall from Section 10.5 that the supply curve AE facing the firm in Figure 14.7(b) is its **average expenditure curve** (just as the demand curve facing a firm is its *average revenue curve*), because it represents the price per unit that the firm pays for the good. On the other hand, the **marginal expenditure curve** represents the firm's expenditure on an *additional unit* that it buys. (The marginal expenditure curve in a factor market is analogous to the marginal revenue curve in the output market.) When the factor market is competitive, the average expenditure and marginal expenditure curves are identical horizontal lines, just as the marginal and average revenue curves are identical (and horizontal) for a competitive firm in the output market.

How much of the input should a firm facing a competitive factor market purchase? As long as the marginal revenue product curve lies above the marginal expenditure curve, profit can be increased by purchasing more of the input because the benefit of an additional unit (MRP) exceeds the cost (ME). However, when the marginal revenue product curve lies below the marginal expenditure curve, some units yield benefits that are less than cost. Therefore, profit maximization requires that *marginal revenue product be equal to marginal expenditure:*

$$\text{ME} = \text{MRP} \tag{14.5}$$

When we considered the special case of a competitive output market, we saw that the firm bought inputs, such as labor, up to the point at which the marginal revenue product is equal to the price of the input w, as in equation (14.3). In the competitive case, therefore, the condition for profit maximization is that the price of the input be equal to marginal expenditure:

$$\text{ME} = w \tag{14.6}$$

In our example, the price of the fabric ($10 per yard) is determined in the competitive fabric market shown in Figure 14.7(a) at the intersection of the demand and supply curves. Figure 14.7(b) shows the amount of fabric purchased by a firm at the intersection of the marginal expenditure and marginal revenue product curves. When 50 yards of fabric are purchased, the marginal expenditure of $10 is equal to the marginal revenue from the sale of clothing made possible by the increased use of fabric in the production process. If less than 50 yards of fabric were purchased, the firm would be forgoing an opportunity to make additional profit from clothing sales. If more than 50 yards were purchased, the cost of the fabric would be greater than the additional revenue from the sale of the extra clothing.

The Market Supply of Inputs

The market supply curve for a factor input is usually upward sloping. We saw in Chapter 8 that the market supply for a good sold in a competitive market is usually upward sloping because the marginal cost of producing the good is typically increasing. This is also the case for fabric and other raw material inputs.

When the input is labor, however, people rather than firms are making supply decisions. In this case, utility maximization by workers rather than profit maximization by firms determines supply. In the discussion that follows, we use the analysis of income and substitution effects from Chapter 4 to show that although the market supply curve for labor can be upward sloping, it may also, as in Figure 14.8, be *backward bending*. In other words, a higher wage rate can lead to *less* labor being supplied.

In §8.5, we explain that the short-run market supply curve shows the amount of output that will be produced by firms in the market for every possible price.

In §4.2, we explain that an increase in the price of a good has two effects: the real purchasing power of each consumer decreases (the income effect) and the good becomes relatively expensive (the substitution effect).

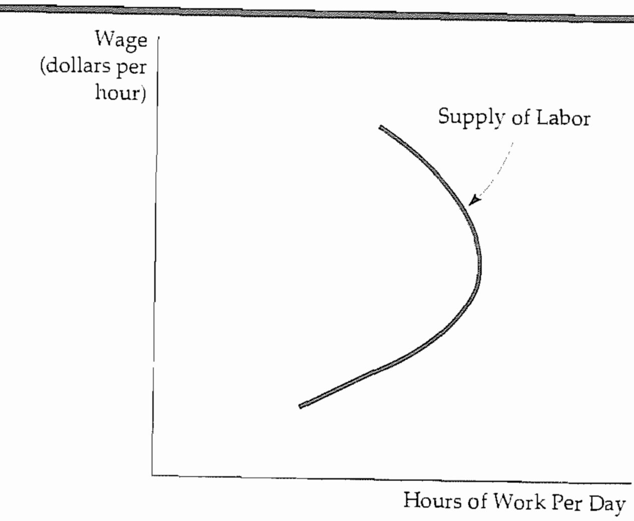

FIGURE 14.8 Backward-Bending Supply of Labor

When the wage rate increases, the hours of work supplied increase initially but can eventually decrease as individuals choose to enjoy more leisure and to work less. The backward-bending portion of the labor supply curve arises when the income effect of the higher wage (which encourages more leisure) is greater than the substitution effect (which encourages more work).

To see why a labor supply curve may be backward bending, divide the day into hours of work and hours of leisure. *Leisure* is a term that describes enjoyable nonwork activities, including sleeping and eating. *Work* benefits the worker only through the income that it generates. We also assume that a worker has the flexibility to choose how many hours per day to work.

The wage rate measures the price that the worker places on leisure time, because his or her wage measures the amount of money that the worker gives up to enjoy leisure. As the wage rate increases, therefore, the price of leisure also increases. This price change brings about both a substitution effect (a change in relative price with utility held constant) and an income effect (a change in utility with relative prices unchanged). There is a substitution effect because the higher price of leisure encourages workers to substitute work for leisure. An income effect occurs because the higher wage rate increases the worker's purchasing power. With higher income, the worker can buy more of many goods, one of which is leisure. If more leisure was chosen, it is because the income effect has encouraged the worker to work fewer hours. Income effects can be large because wages are the primary component of most people's income. When the income effect outweighs the substitution effect, the result is a backward-bending supply curve.

Figure 14.9 illustrates how a backward-bending supply curve for labor can result from the work-leisure decision for a typical weekday. The horizontal axis shows hours of leisure per day, the vertical axis income generated by work. (We assume there are no other sources of income.) Initially the wage rate is $10 per hour, and the budget line is given by PQ. Point P, for example, shows that if an individual worked a 24-hour day he would earn an income of $240.

The worker maximizes utility by choosing point A and by enjoying 16 hours of leisure per day (with 8 hours of work) and earning $80. When the wage rate increases to $20 per hour, the budget line rotates about the horizontal intercept to line RQ. (Only 24 hours of leisure are possible.) Now the worker maximizes utility at B by choosing 20 hours of leisure per day (with 4 hours of work), while earning $80. If only the substitution effect came into play, the higher wage rate would encourage the worker to work 12 hours (at C) instead of 8. However, the income effect works in the opposite direction. It overcomes the substitution effect and lowers the work day from 8 hours to 4.

In real life, a backward-bending labor supply curve might apply to a college student working during the summer to earn living expenses for the school year. As soon as a target level of earnings is reached, the student stops working and allocates more time to leisure. An increase in the wage rate will then lead to fewer hours worked because it enables the student to reach the target level of earnings faster.

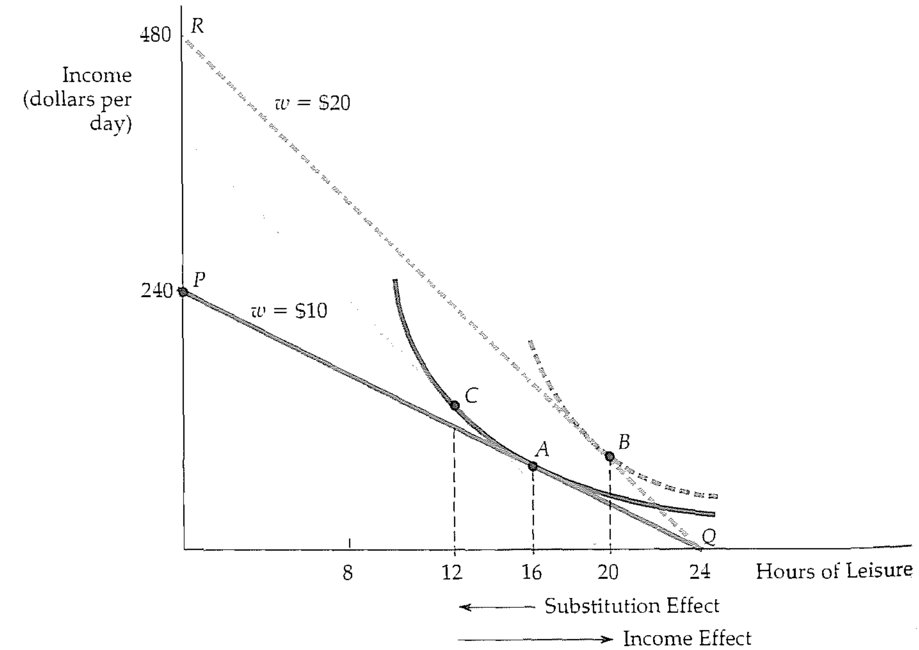

FIGURE 14.9 Substitution and Income Effects of a Wage Increase

When the wage rate increases from $10 to $20 per hour, the worker's budget line shifts from PQ to RQ. In response, the worker moves from A to B while decreasing work hours from 8 to 4. The reduction in hours worked arises because the income effect outweighs the substitution effect. In this case, the supply of labor curve is backward bending.

EXAMPLE 14.2 Labor Supply for One- and Two-Earner Households

One of the most dramatic changes in the labor market in the twentieth century has been the increase in women's participation in the labor force. Women made up only 29 percent of the labor force in 1950 but over 60 percent in 1999. Married women account for a substantial portion of this increase. The increased role of women in the labor market has also had a major impact on housing markets: Where to live and work has increasingly become a joint husband-and-wife decision.

The complex nature of the work choice was analyzed in a study that compared the work decisions of 94 unmarried females with the work decisions of heads of households and spouses in 397 families.[3] One way to describe the work decisions of the various family groups is to calculate labor supply elasticities. Each elasticity relates the numbers of hours worked not only to the wage earned by the head of the household but also to the wage of the other member of two-earner households. Table 14.2 summarizes the results.

When a higher wage rate leads to fewer hours worked, the labor supply curve is backward bending: The income effect, which encourages more leisure, outweighs the substitution effect, which encourages more work. The elasticity of labor supply is then negative. Table 14.2 shows that heads of one-earner families with children and two-earner families (with or without children) all have backward-bending labor supply curves, with elasticities ranging from −.002 to −.078. Most single-earner heads of households are on the upward-sloping portion of their labor supply curve, with the largest elasticity of .106 associated

[3] See Janet E. Kohlhase, "Labor Supply and Housing Demand for One- and Two-Earner Households," *Review of Economics and Statistics* 68 (1986): 48–56; and Ray C. Fair and Diane J. Macunovich, "Explaining the Labor Force Participation of Women 20–24" (unpublished, February 1997).

TABLE 14.2 Elasticities of Labor Supply (Hours Worked)

GROUP	HEAD'S HOURS WITH RESPECT TO HEAD'S WAGE	SPOUSE'S HOURS WITH RESPECT TO SPOUSE'S WAGE	HEAD'S HOURS WITH RESPECT TO SPOUSE'S WAGE
Unmarried males (no children)	.026		
Unmarried females (with children)	.106		
Unmarried females (no children)	.011		
One-earner family (with children)	−.078		
One-earner family (no children)	.007		
Two-earner family (with children)	−.002	−.086	−.004
Two-earner family (no children)	−.107	−.028	−.059

with single women with children. Married women (listed as spouses of heads of households) are also on the backward-bending portion of the labor supply curve, with elasticities of −.028 and −.086.

14.2 Equilibrium in a Competitive Factor Market

A competitive factor market is in equilibrium when the price of the input equates the quantity demanded to the quantity supplied. Figure 14.10(a) shows such an equilibrium for a labor market. At point *A*, the equilibrium wage rate is w_C and the equilibrium quantity supplied is L_C. Because they are well informed, all workers receive the identical wage and generate the identical marginal revenue product of labor wherever they are employed. If any worker had a wage lower than her marginal product, a firm would find it profitable to offer that worker a higher wage.

If the output market is also perfectly competitive, the demand curve for an input measures the benefit that consumers of the product place on the additional use of the input in the production process. The wage rate also reflects the cost to the firm and to society of using an additional unit of the input. Thus, at *A* in Figure 14.10(a), the marginal benefit of an hour of labor (its marginal revenue product MRP_L) is equal to its marginal cost (the wage rate w).

In §9.2, we explain that in a perfectly competitive market, efficiency is achieved because the sum of aggregate consumer and producer surplus is maximized.

When output and input markets are both perfectly competitive, resources are used efficiently because the difference between total benefits and total costs is maximized. Efficiency requires that the additional revenue generated by employing an additional unit of labor (the marginal revenue product of labor, MRP_L) equal the benefit to consumers of the additional output, which is given by the price of the product times the marginal product of labor, $(P)(MP_L)$.

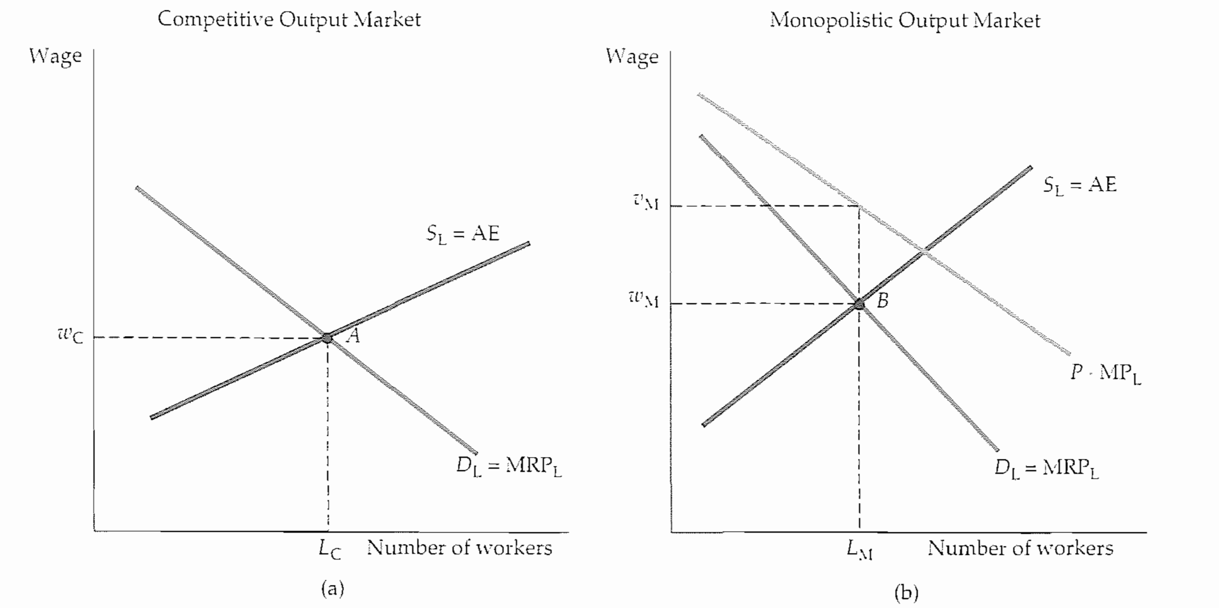

FIGURE 14.10 Labor Market Equilibrium

In a competitive labor market in which the output market is competitive, the equilibrium wage w_C is given by the intersection of the demand for labor (marginal revenue product) curve and the supply of labor (average expenditure) curve. This is point *A* in part (a) of the figure. Part (b) shows that when the producer has monopoly power, the marginal value of a worker v_M is greater than the wage w_M. Thus too few workers are employed. (Point *B* determines the quantity of labor that the firm hires and the wage rate paid.)

When the output market is not perfectly competitive, the condition $MRP_L = (P)(MP_L)$ no longer holds. Note in Figure 14.10(b) that the curve representing the product price multiplied by the marginal product of labor $[(P)(MP_L)]$ lies above the marginal revenue product curve $[(MR)(MP_L)]$. Point *B* is the equilibrium wage w_M and the equilibrium labor supply L_M. But because the price of the product is a measure of the value to consumers of each additional unit of output that they buy, $(P)(MP_L)$ is the value that consumers place on additional units of labor. Therefore, when L_M laborers are employed, the marginal cost to the firm w_M is less than the marginal benefit to consumers v_M. Although the firm is maximizing its profit, its output is below the efficient level and it uses less than the efficient level of the input. Economic efficiency would be increased if more laborers were hired and, consequently, more output produced. (The gains to consumers would outweigh the firm's lost profit.)

Economic Rent

The concept of economic rent helps explain how factor markets work. When discussing output markets in the long run, in Chapter 8, we defined economic rent as the payments received by a firm over and above the minimum cost of producing its output. For a factor market, *economic rent is the difference between the payments made to a factor of production and the minimum amount that must be spent to obtain the use of that factor*. Figure 14.11 illustrates the concept of economic rent as applied to a competitive labor market. The equilibrium price of labor is w^*, and the quantity of labor supplied is L^*. The supply of labor curve is the upward-sloping average expenditure curve, and the demand for labor is the downward-

In §8.6, we explain that economic rent is the amount that firms are willing to pay for an input less the minimum amount necessary to buy it.

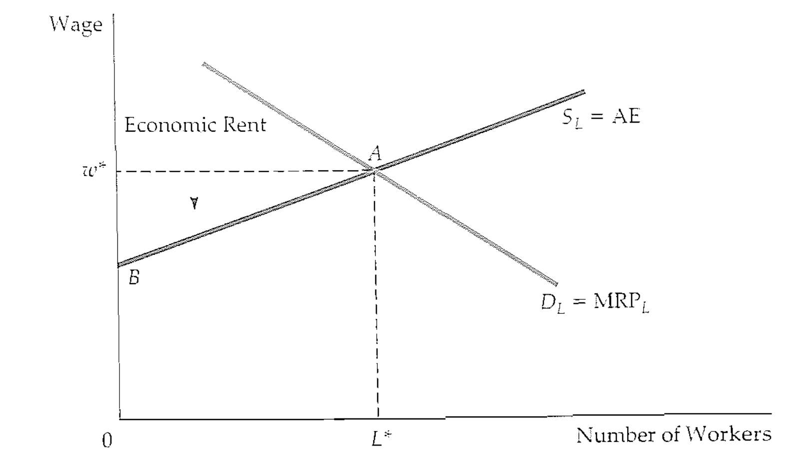

FIGURE 14.11 Economic Rent

The economic rent associated with the employment of labor is the excess of wages paid above the minimum amount needed to hire workers. The equilibrium wage is given by A, at the intersection of the labor supply and labor demand curves. Because the supply curve (AE) is upward sloping, some workers would have accepted jobs for a wage less than w^*. The green-shaded area ABw^* is the economic rent received by all workers.

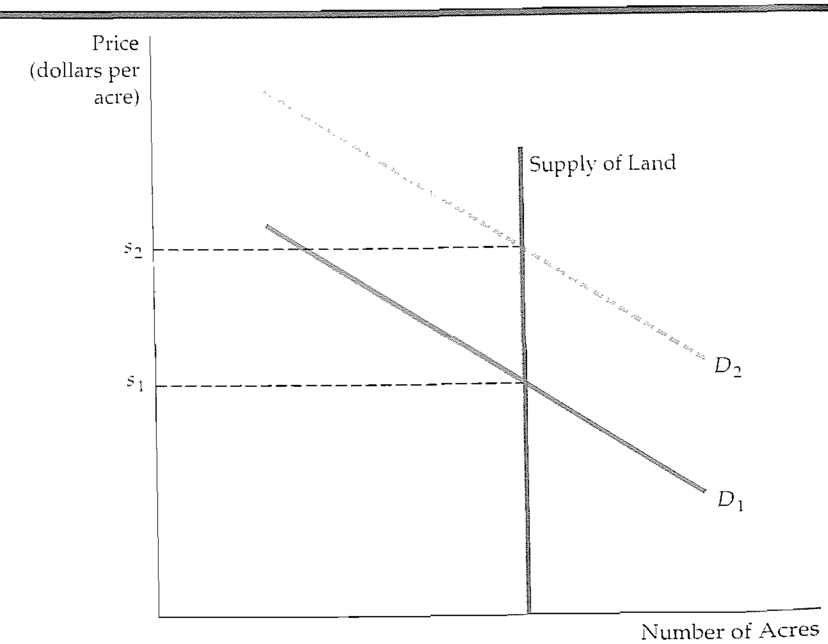

FIGURE 14.12 Land Rent

When the supply of land is perfectly inelastic, the market price of land is determined at the point of intersection with the demand curve. The entire value of the land is then an economic rent. When demand is given by D_1, the economic rent per acre is given by s_1, and when demand increases to D_2, rent per acre increases to s_2.

sloping marginal revenue product curve. Because the supply curve tells us how much labor will be supplied at each wage rate, the minimum expenditure needed to employ L^* units of labor is given by the tan-shaded area AL^*0B, below the supply curve to the left of the equilibrium labor supply L^*.

In perfectly competitive markets, all workers are paid the wage w^*. This wage is required to get the last "marginal" worker to supply his or her labor, but all other workers earn rents because their wage is greater than the wage that would be needed to get them to work. Because total wage payments are equal to the rectangle $0w^*AL^*$, the economic rent earned by labor is given by the area ABw^*.

Note that if the supply curve were perfectly elastic, economic rent would be zero. Rents arise only when supply is somewhat inelastic. And when supply is perfectly inelastic, all payments to a factor of production are economic rents because the factor will be supplied no matter what price is paid.

As Figure 14.12 shows, one example of an inelastically supplied factor is land. The supply curve is perfectly inelastic because land for housing (or for agriculture) is fixed, at least in the short run. With land inelastically supplied, its price is determined entirely by demand. The demand for land is given by D_1, and its price per unit is s_1. Total land rent is given by the green-shaded rectangle. But when the demand for land increases to D_2, the rental value per unit of land increases to s_2; now total land rent includes the blue-shaded area as well. Thus, an increase in the demand for land (a shift to the right in the demand curve) leads both to a higher price per acre and to a higher economic rent.

EXAMPLE 14.3 Pay in the Military

The U.S. Army has had a personnel problem for many years. During the Civil War, roughly 90 percent of the armed forces were unskilled workers involved in ground combat. But since then the nature of warfare has evolved. Ground combat forces now make up only 16 percent of the armed forces. Meanwhile, changes in technology have led to a severe shortage in skilled technicians, trained pilots, computer analysts, mechanics, and others needed to operate sophisticated military equipment. Why has such a shortage developed? Why has the military been unable to keep skilled personnel? An economic study provides some answers.[4]

The rank structure of the army has remained essentially unchanged over the years. Among the officer ranks, pay increases are determined primarily by the number of years of service. Consequently, officers with differing skill levels and abilities are usually paid similar salaries. Moreover, some skilled workers are underpaid relative to what they could receive in the private sector. As a result, skilled workers who join the army because of attractive salaries find that their marginal revenue products are eventually higher than their wages. Some remain in the army, but many leave.

This study of army pay applies to all of the armed forces. Figure 14.13 shows the inefficiency that can result from the military pay policy. The equilibrium wage rate w^* is the wage that equates the demand for labor to the supply. Because of inflexibility in its pay structure, however, the military pays the wage w_0, which is below the equilibrium wage. At w_0, demand is greater than supply, and there

[4] Walter Y. Oi, "Paying Soldiers: On a Wage Structure for a Large Internal Labor Market" (unpublished, undated paper).

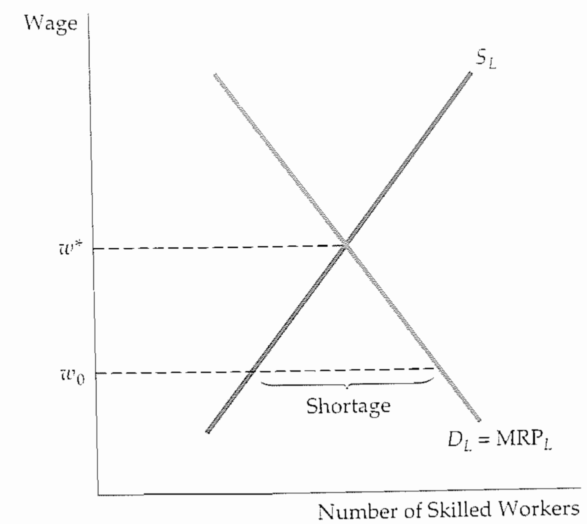

FIGURE 14.13 The Shortage of Skilled Military Personnel

When the wage w^* is paid to military personnel, the labor market is in equilibrium. When the wage is kept below w^*, at w_0, there is a shortage of personnel because the quantity of labor demanded is greater than the quantity supplied.

is a shortage of skilled labor. By contrast, competitive labor markets pay more productive workers higher wages than their less productive counterparts. But how can the military attract and keep a skilled labor force?

The military's choice of wage structure affects the nation's ability to maintain an effective fighting force. In response to its personnel problems, the military has begun to change its salary structure by expanding the number and size of its reenlistment bonuses. Selective reenlistment bonuses targeted at skilled jobs for which there are shortages can be an effective recruiting device. The immediate bonuses create more of an incentive than the uncertain promise of higher wages in the future. As the demand for skilled military jobs increases, we can expect the armed forces to make greater use of reenlistment bonuses and other market-based incentives.

14.3 Factor Markets with Monopsony Power

In §10.5, we explain that a buyer has monopsony power when his purchasing decision can affect the price of the product.

In some factor markets, individual buyers of factors have monopsony power. For example, we saw in Chapter 10 that U.S. automobile companies have considerable monopsony power as buyers of parts and components. GM, Ford, and Daimler-Chrysler buy large quantities of brakes, radiators, tires, and other parts and can negotiate lower prices than smaller purchasers might pay. Likewise, IBM has monopsony power in the market for disk drives because it purchases so many drives for its computers.

Throughout this section, we will assume that the output market is perfectly competitive. Also, because a single buyer is easier to visualize than several buyers who all have some monopsony power, we will restrict our attention to pure monopsony.

Marginal and Average Expenditure

Recall from Section 10.5 that when you are deciding how much of a good to purchase, you keep increasing the number of units purchased until the additional value from the last unit purchased—the *marginal value*—is just equal to the cost of that unit—the *marginal expenditure*. In perfect competition, the price that you pay for the good—the *average expenditure*—is equal to the marginal expenditure. However, when you have monopsony power, the marginal expenditure is greater than the average expenditure, as Figure 14.14 shows.

The factor supply curve facing the monopsonist is the market supply curve. (It shows how much of the factor suppliers are willing to sell as its price increases.) Because the monopsonist pays the same price for each unit, the supply curve is its *average expenditure curve*. The average expenditure curve is upward sloping because the decision to buy an extra unit raises the price that must be paid for all units, not just the last one. For a profit-maximizing firm, however, the *marginal expenditure curve* is relevant in deciding how much to buy. Recall from Chapter 10 that the marginal expenditure curve lies above the average expenditure curve: When the firm increases the price of the factor to hire more units, it must pay *all* units that higher price, not just the last unit hired.

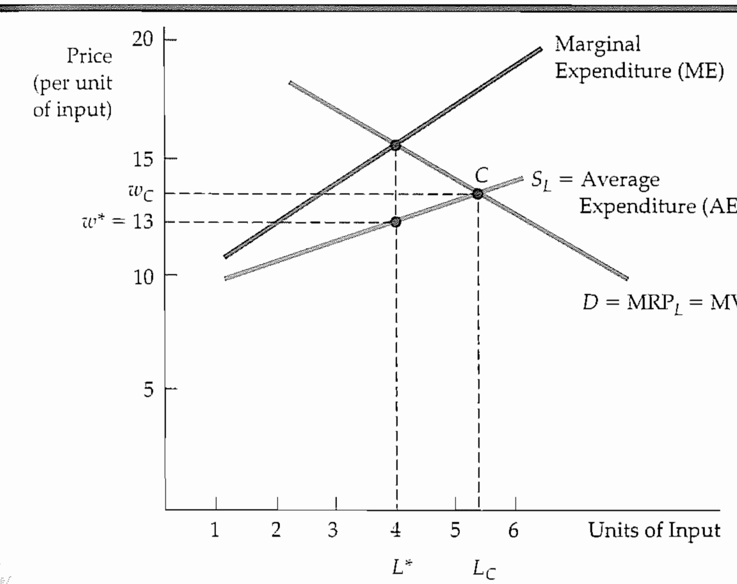

FIGURE 14.14 Marginal and Average Expenditure

When the buyer of an input has monopsony power, the marginal expenditure curve lies above the average expenditure curve because the decision to buy an extra unit raises the price that must be paid for all units, not just for the last one. The number of units of input purchased is given by L^*, at the intersection of the marginal revenue product and marginal expenditure curves. The corresponding wage rate w^* is lower than the competitive wage w_C.

The Input Purchasing Decision of the Firm

How much of the input should the firm buy? As we saw earlier, it should buy up to the point where marginal expenditure equals marginal revenue product. Here the benefit from the last unit bought (MRP) is just equal to the cost (ME). Figure 14.14 illustrates this principle for a labor market. Note that the monopsonist hires L^* units of labor; at that point, ME = MRP_L. The wage rate w^* that workers are paid is given by finding the point on the average expenditure or supply curve with L^* units of labor.

As we showed in Chapter 10, a buyer with monopsony power maximizes net benefit (utility less expenditure) from a purchase by buying up to the point where marginal value (MV) is equal to marginal expenditure:

$$MV = ME$$

For a firm buying a factor input, MV is just the marginal revenue product of the factor MRP. Thus, we have (as in the case of a competitive factor market)

$$ME = MRP \qquad (14.7)$$

Note from Figure 14.14 that the monopsonist hires less labor than a firm or group of firms with no monopsony power. In a competitive labor market, L_C workers would be hired: At that level, the quantity of labor demanded (given by the marginal revenue product curve) is equal to the quantity of labor supplied (given by the average expenditure curve). Note also that the monopsonistic firm will be paying its workers a wage w^* that is less than the wage w_C that would be paid in a competitive market.

Monopsony power can arise in different ways. One source can be the specialized nature of a firm's business. If the firm buys a component that no one else buys, it is likely to be a monopsonist in the market for that component. Another source can be a business's location—it may be the only major employer within an area. Monopsony power can also arise when the buyers of a factor form a cartel to limit purchases of the factor, so they can buy it at less than the competitive price. (But as we explained in Chapter 10, this is a violation of the antitrust laws.)

Few firms in our economy are pure monopsonists. But firms (or individuals) often have some monopsony power, because their purchases account for a large portion of the market. The government is a monopsonist when it hires volunteer soldiers or buys missiles, aircraft, and other specialized military equipment. A mining firm or other company that is the only major employer in a community also has monopsony power in the local labor market. Even in these cases, however, monopsony power may be limited because the government competes to some extent with other firms that offer similar jobs. Likewise the mining firm competes to some extent with companies in nearby communities.

EXAMPLE 14.4 Monopsony Power in the Market for Baseball Players

In the United States, major league baseball is exempt from the antitrust laws, the result of a Supreme Court decision and the policy of Congress not to apply the antitrust laws to labor markets.[5] This antitrust exemption allowed baseball team owners (before 1975) to operate a monopsonistic cartel. Like all cartels, it depended on an agreement among owners. This agreement involved an annual draft of players and a *reserve clause* that effectively tied each player to one team for life, thereby eliminating most interteam competition for players. Once a player was drafted by a team, he could not play for another team unless rights were sold to that team. As a result, baseball owners had monopsony power in negotiating new contracts with their players: The only alternative to signing an agreement was to give up the game or play it outside the United States.

During the 1960s and early 1970s, baseball players' salaries were far below the market value of their marginal products (determined in part by the incremental attention that better hitting or pitching might achieve). For example, if the players' market had been perfectly competitive, those players receiving a salary of about $42,000 in 1969 would have instead received a salary of $300,000 in 1969 dollars (which is $1.5 million in year 2000 dollars).

Fortunately for the players, and unfortunately for the owners, there was a strike in 1972 followed by a lawsuit by one player (Curt Flood of the St. Louis Cardinals) and an arbitrated labor-management agreement. This process eventually led in 1975 to an agreement by which players could become free agents after playing for a team for six years. The reserve clause was no longer in effect, and a highly monopsonistic labor market became much more competitive.

The result was an interesting experiment in labor market economics. Between 1975 and 1980, the market for baseball players adjusted to a new post-reserve clause equilibrium. Before 1975, expenditures on players' contracts made up approximately 25 percent of all team expenditures, but by 1980 those expenditures had increased to 40 percent. Moreover, the average player's salary doubled in real terms. By 1992, the average baseball player was earning $1,014,942—a very large increase from the monopsonistic wages of the 1960s. In 1969, for example, the average baseball salary was approximately $42,000. Adjusted for inflation, this is about $200,000 in year 1999 dollars.

Salaries for baseball players continued to grow during the 1990s. Average salaries in 1997 were $1,383,578, and some players earned far more. In 1999, for example, Mo Vaughan and Randy Johnson, the two highest-paid players, earned $13,333,333 and $13,250,000, respectively.

[5] This example builds on an analysis of the structure of baseball players' salaries by Roger Noll, who has kindly supplied us with the relevant data.

EXAMPLE 14.5 Teenage Labor Markets and the Minimum Wage

Increases in the national minimum wage rate (which was $4.50 in early 1996 and $5.15 in 1999) were controversial, raising the question of whether the cost of any unemployment that might be generated would be outweighed by the benefit of higher incomes to those whose wage has been increased.[6] A study of

In §9.3, we explain that setting a minimum wage in a perfectly competitive market can create unemployment and a deadweight loss.

[6] See Example 1.3 for an initial discussion of the minimum wage, and Section 9.3 for an analysis of its effects on employment.

the effects of the minimum wage on employment in fast-food restaurants in New Jersey added to that controversy.[7]

Some states have minimum wages above the Federal level. In April 1992 the New Jersey minimum wage was increased from $4.25 to $5.05 per hour. Using a survey of 410 fast-food restaurants, David Card and Alan Krueger found that employment had actually *increased* by 13 percent after the minimum wage went up. What is the explanation for this surprising result? One possibility is that restaurants responded to the higher minimum wage by reducing fringe benefits, which usually take the form of free and reduced-price meals for employees. A related explanation is that employers responded by providing less on-the-job training and by offering lower wages for those with experience who had previously been paid more than the minimum wage.

An alternative explanation for the increased New Jersey employment holds that the labor market for teenage (and other) unskilled workers is not highly competitive. If so, the analysis of Chapter 9 does not apply. If the unskilled fast-food labor market were monopsonistic, for example, we would expect a different effect from the increased minimum wage. Suppose that the wage of $4.25 was the wage that fast-food employers with monopsony power in the labor market would offer their workers even if there were no minimum wage. Suppose also that $5.10 would be the wage enjoyed by workers if the labor market were fully competitive. As Figure 14.14 shows, the increase in the minimum wage would not only raise the wage, but would also increase the employment level (from L^* to L_C).

Does the fast-food study show that employers have monopsony power in this labor market? The evidence suggests no. If firms do have monopsony power but the fast-food market is competitive, then the increase in the minimum wage should have no effect on the price of fast food. Because the market for fast food is so competitive, firms paying the higher minimum wage would be forced to absorb the higher wage cost themselves. The study suggests, however, that prices did increase after the introduction of the higher minimum wage.

The Card-Krueger analysis of the minimum wage remains hotly debated. A number of authors have argued that the New Jersey study was atypical; they point out that most studies do show that a higher minimum reduces employment, as we suggested in Chapter 9.[8] Where do we go from here? Perhaps a better characterization of low-wage labor markets requires a more sophisticated theory (e.g., the efficiency wage theory discussed in Chapter 17). In any case, new empirical analyses should shed more light on the effects of the minimum wage.

[7] David Card and Alan Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *Quarterly Journal of Economics*, 1998. See also David Card and Alan Krueger, "A Reanalysis of the Effect of the New Jersey Minimum Wage on the Fast-Food Industry with Representative Payroll Data," Working Paper No. 6386, Cambridge, MA: National Bureau of Economic Research, 1998; and Madeline Zadodvy, "Why Minimum Wage Hikes May Not Reduce Employment," Federal Reserve Bank of Atlanta, Economic Review, Second Quarter, 1998.

[8] For example, see Donald Deere, Kevin M. Murphy, and Finis Welch, "Employment and the 1990–1991 Minimum Wage Hike," *American Economic Review Papers and Proceedings* 85 (May 1995): 232–37; and David Neumark and William Wascher, "The New Jersey-Pennsylvania Minimum Wage Experiment: A Reevaluation Using Payroll Records," *American Economic Review*, forthcoming.

14.4 Factor Markets with Monopoly Power

Just as buyers of inputs can have monopsony power, sellers of inputs can have monopoly power. In the extreme, the seller of an input may be a monopolist, as when a firm has a patent to produce a computer chip that no other firm can duplicate. Because the most important example of monopoly power in factor markets involves labor unions, we will concentrate most of our attention there. In the subsections that follow, we show how a labor union, which is a monopolist in the sale of labor services, might increase the well-being of its members and substantially affect nonunionized workers.

In §10.2, we explain that a seller of a product has some monopoly power if it can profitably charge a price greater than marginal cost.

Monopoly Power over the Wage Rate

Figure 14.15 shows a demand for labor curve in a market with no monopsony power: It aggregates the marginal revenue products of firms that compete to buy labor. The labor supply curve describes how union members would supply labor *if* the union exerted no monopoly power. In that case, the labor market would be competitive, and L^* workers would be hired at a wage of w^*, where demand D_L equals supply S_L.

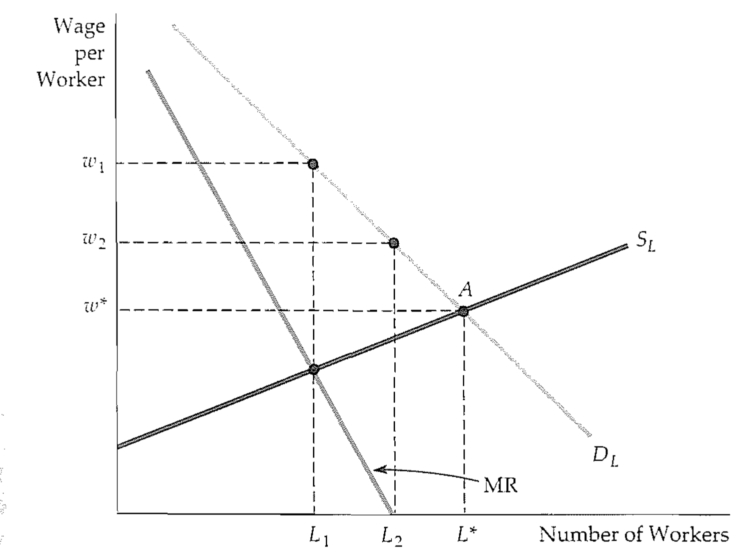

FIGURE 14.15 Monopoly Power of Sellers of Labor

When a labor union is a monopolist, it chooses among points on the buyer's demand for labor curve D_L. The seller can maximize the number of workers hired, at L^*, by agreeing that workers will work at wage w^*. The quantity of labor L_1 that maximizes the rent earned by employees is determined by the intersection of the marginal revenue and supply of labor curves; union members will receive a wage rate of w_1. Finally, if the union wishes to maximize total wages paid to workers, it should allow L_2 union members to be employed at a wage rate of w_2: At that point, the marginal revenue to the union will be zero.

Because of its monopoly power, however, the union can choose any wage rate and the corresponding quantity of labor supplied, just as a monopolist seller of output chooses price and the corresponding quantity of output. If the union wanted to maximize the number of workers hired, it would choose the competitive outcome at A. However, if the union wished to obtain a higher than competitive wage, it could restrict its membership to L_1 workers. As a result, the firm would pay a wage rate of w_1. While union members who work would be better off, those who cannot find jobs would be worse off.

Is a policy of restrictive union membership worthwhile? If the union wishes to maximize the economic rent that its workers receive, the answer is yes. By restricting membership, the union would be acting like a monopolist, which restricts output in order to maximize profit. To a firm, profit is the revenue it receives less its opportunity costs. To a union, rent represents the wages its members earn as a group in excess of their opportunity cost. To maximize rent, the union must choose the number of workers hired so that the marginal revenue to the union (the additional wages earned) is equal to the extra cost of inducing workers to work. This cost is a *marginal* opportunity cost because it is a measure of what an employer has to offer an additional worker to get him or her to work for the firm. However, the wage that is necessary to encourage additional workers to take jobs is given by the supply of labor curve S_L.

In §7.1, we explain that opportunity cost is the cost associated with opportunities that are forgone by not putting a firm's resources to their highest-value use.

The rent-maximizing combination of wage rate and number of workers is given by the intersection of the MR and S_L curves. We have chosen the wage-employment combination of w_1 and L_1 with the rent-maximization premise in mind. The shaded area below the demand for labor curve, above the supply of labor curve and to the left of L_1, represents the economic rent that all workers receive.

A rent-maximizing policy might benefit nonunion workers if they can find nonunion jobs. However, if these jobs are not available, rent maximization could create too sharp a distinction between winners and losers. An alternative objective is to maximize the aggregate wages that all union members receive. Look again at the example in Figure 14.15. To achieve this goal, the number of workers hired is increased from L_1 until the marginal revenue to the union is equal to zero. Because any further employment decreases total wage payments, aggregate wages are maximized when the wage is equal to w_2 and the number of workers is equal to L_2.

Unionized and Nonunionized Workers

When the union uses its monopoly power to increase members' wages, fewer unionized workers are hired. Because these workers either move to the nonunion sector or choose initially not to join the union, it is important to understand what happens in the nonunionized part of the economy.

Assume that the total supply of unionized and nonunionized workers is fixed. In Figure 14.16, the market supply of labor in both sectors is given by S_L. The demand for labor by firms in the unionized sector is given by D_U, the demand in the nonunionized sector by D_{NU}. Total market demand is the horizontal sum of the demands in the two sectors and is given by D_L.

Suppose the union chooses to increase the wage rate of its workers above the competitive wage w^*, to w_U. At that wage rate, the number of workers hired in the unionized sector falls by an amount ΔL_U, as shown on the horizontal axis. As these workers find employment in the nonunionized sector, the wage rate in the

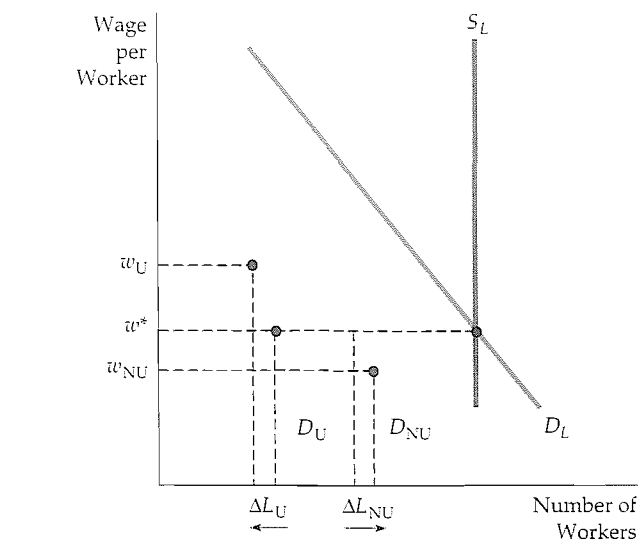

FIGURE 14.16 Wage Discrimination in Unionized and Nonunionized Sectors

When a monopolistic union raises the wage in the unionized sector of the economy from w^* to w_U, employment in that sector falls, as shown by the movement along the demand curve D_U. For the total supply of labor, given by S_L, to remain unchanged, the wage in the nonunionized sector must fall from w^* to w_{NU}, as shown by the movement along the demand curve D_{NU}.

nonunionized sector adjusts until the labor market is in equilibrium. At the new wage rate in the nonunionized sector, w_{NU}, the additional number of workers hired in the nonunionized sector, ΔL_{NU}, is equal to the number of workers who left the unionized sector.

Figure 14.16 shows an adverse consequence of a union strategy directed toward raising union wages: Nonunionized wages fall. Unionization can improve working conditions and provide useful information to workers and management. But when the demand for labor is not perfectly inelastic, union workers are helped at the expense of nonunion workers.

Bilateral Monopoly in the Labor Market

The adverse effects of union wage policies by a monopolistic union depend to some extent on our assumption that the input market is otherwise competitive. We now consider the consequences of union wage policies when the buyers of labor also have monopsony power.

As we saw in Chapter 10, a *bilateral monopoly* is a market in which a monopolist sells to a monopsonist. In a labor market, a bilateral monopoly might arise when representatives from a union meet to negotiate wages with companies that hire a certain type of worker. Figure 14.17 shows a typical bilateral bargaining situation. The S_L curve represents the supply curve for skilled labor. The firm's demand curve for labor is given by the marginal revenue product curve D_L.

In §10.6, we explain that a bilateral monopoly is a market in which there is only one buyer and one seller.

If the union had no monopoly power, the monopsonist would make its hiring decision on the basis of its marginal expenditure curve ME, choosing to hire 20

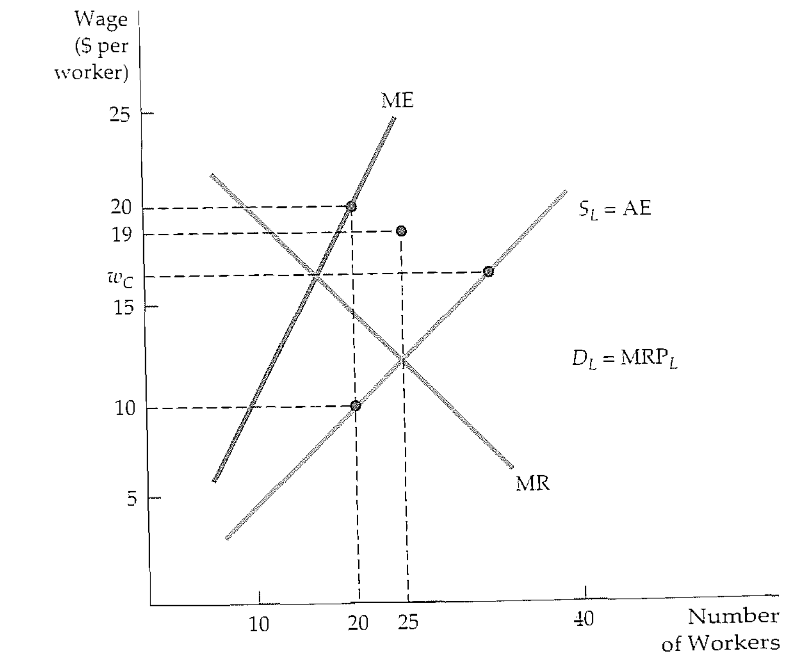

FIGURE 14.17 Bilateral Monopoly

When the seller of labor is a monopolist and the buyer a monopsonist, the negotiated wage rate will be between a high of $19 (determined by the intersection of the marginal revenue and average expenditure curves) and a low of $10 (determined by the intersection of the marginal revenue product curve and marginal expenditure curves).

workers and paying them $10 per hour. When 20 workers are hired, the marginal revenue product of labor is equal to the marginal expenditure of the firm.

The seller of labor faces a demand curve D_L that describes the firm's hiring plans as the wage rate varies. The union chooses a point on the demand curve that maximizes its members' wages. Remember that the wage paid to all workers falls as the number hired increases. The marginal revenue curve MR, therefore, describes the additional wages that the union gets for its members as the number of employees hired increases.

The supply curve S_L tells the union the minimum payment necessary to encourage workers to offer their labor to firms in the industry. Suppose the union wishes to maximize the economic rent of its members. To do so, it views the supply curve as the marginal cost of labor. To maximize the rent that is earned, the union chooses a wage of $19 because $19 is the wage that equates the marginal revenue (the marginal increase in wages) to the marginal cost (the increase in the minimum wages needed to hire the labor). At $19, firms would hire 25 workers.

In summary, firms are willing to pay a wage of $10 and hire 20 workers, but the union is demanding a wage of $19 and wants the firm to hire 25 workers. What happens in this case? The result depends on the bargaining strategies of the two parties. If the union can make a credible threat to strike, it might secure a wage closer to $19 than to $10. If the firms can make a credible threat to hire nonunion labor, they might secure a wage closer to $10. If both parties can make credible threats, the resulting agreement might be close to the competitive outcome (wage w_C) of about $15 in Figure 14.17.[9]

EXAMPLE 14.6 The Decline of Private-Sector Unionism

For several decades, both the membership and bargaining power of labor unions have been declining.[10] A decline in union monopoly power can lead to different responses by union negotiators and can also affect the wage rate and level of employment. During the 1970s, most of the impact was on union wages: Although levels of employment did not change much, the differential between union and nonunion wages decreased substantially. We would have expected the same pattern to occur in the 1980s because of heavily publicized wage freezes and the rapid growth of two-tier wage provisions in which newer union members are paid less than more experienced counterparts.

Surprisingly, however, the union-management bargaining process changed during this period. From 1979 to 1984, the level of unionized employment fell from 27.8 percent to 19.0 percent. Yet the union-nonunion wage differential remained stable—and in fact grew wider in some industries. For example, the union wage rate in mining, forestry, and fisheries declined only from 25 percent higher than the nonunion wage in 1979 to 24 percent higher in 1984. On the other hand, the union wage rate in manufacturing increased slightly from approximately 14 percent higher than the nonunion wage in 1979 to 16 percent in 1984. This same pattern has continued throughout the 1990s. As Figure 14.18 shows, by 1998 unionized employment had fallen to 14 percent of total employment. The union-nonunion wage differential remained essentially unchanged.

One explanation for this pattern of wage-employment responses is a change in union strategy—a move to maximize the wage rate for its members rather than the total wages paid to all union members. However, the demand for unionized employees has probably become increasingly elastic over time as firms find it easier to substitute capital for skilled labor in the production process. Faced with an elastic demand for its services, the union would have little choice but to maintain the wage rate of its members and allow employment levels to fall. Of course, the substitution of nonunion for union workers

[9] There is no guarantee that monopoly power and monopsony power will cancel each other, or that the total number of workers hired will be near the competitive level. Why? Because both the monopolist and the monopsonist want to limit the number of workers hired. In "Unions and Monopoly Profits," *Review of Economics and Statistics* 67 (1985): 34–42, Thomas Karier shows that unions reduce profits in highly concentrated industries but have little or no effect on profits in more competitive industries.

[10] This example is based on Richard Edwards and Paul Swaim, "Union-Nonunion Earnings Differentials and the Decline of Private-Sector Unionism," *American Economic Review* 76 (May 1986): 97–102.

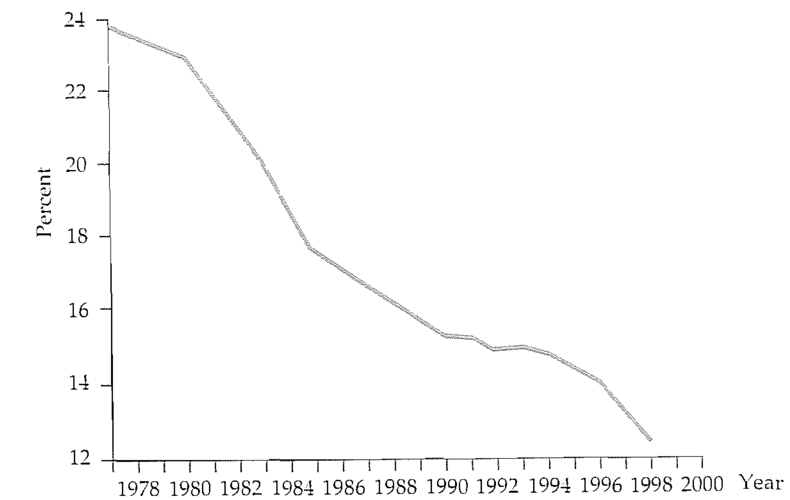

FIGURE 14.18 Union Workers as a Percentage of Total

The percentage of workers that are unionized has been declining steadily over the past two decades.

Source: Bureau of Labor Statistics, "Employment and Earnings," January issue; *The Economist*, June 12, 1999.

may cause further losses in the bargaining power of labor unions. How this will affect the differential between union and nonunion wages remains to be seen.

EXAMPLE 14.7 Wage Inequality—Have Computers Changed the Labor Market?

In Example 2.2, we explained how the rapid growth in the demand for skilled relative to unskilled labor has been partly responsible for the growing inequality in the distribution of income in the United States. What is the underlying source of that change in relative demand? Is it the decline in private-sector unionism and the failure of the minimum wage to keep up with inflation? Or is it the increasing role that computers now play in the labor market? A recent study, which focuses on the wages of college relative to high school graduates, provides some answers.[11]

From 1950 to 1980, the relative wages of college graduates (the ratio of their average wages to those of high school graduates) hardly changed. In contrast, the relative wage grew rapidly from 1980 to 1995. This pattern is not consistent with what one would expect if the decline of unionism (and/or changes in the minimum wage) was the primary reason for the growing inequality. A clue to what happened is given by the dramatic increase in the use of computers by workers during the past 20 years. In 1984, 25.1 percent of all workers used computers; that figure increased to 46.6 percent by 1993 (and is now closer to 60 percent).

While computer use increased for all workers, the largest increases were registered by workers with college degrees—an increase from 42 to 70 percent. For those without a high school degree, the increase was only 5 percentage points (from 5 to 10 percent); for those with high school degrees, the increase was 16 percentage points (from 19 to 35 percent).

A further analysis of data on jobs and wages confirms the importance of computers. Education and computer use have gone hand-in-hand to increase the demand for skilled workers. The wages of college graduates who use computers (relative to high school graduates) grew by about 11 percent from 1983 to 1993; for noncomputer users, wages grew by less than 4 percent. A statistical analysis shows that, overall, the spread of computer technology is responsible for nearly half the increase in relative wages during this period. Furthermore, the growth in the demand for skilled workers has occurred primarily within industries where computers have become increasingly useful.

Is this increase in the relative wages of skilled workers necessarily a bad thing? At least one economist suggests that the answer is no.[12] It is true that the growing inequality can disadvantage low-wage workers, whose limited opportunities might lead them to drop out of the labor force; in the extreme, they might even turn to crime. However, it can also motivate workers, whose opportunities for upward mobility through high-wage jobs have never been better.

Consider the circumstances facing men and women who are deciding whether to complete high school or college. Again, we'll take the average wage of someone who completed high school as the norm. First, the bad news. In the period 1993 to 1997, high school dropouts who had been out of school less than 10 years earned 29 percent less than high school graduates. Although their real wages have increased, high-school dropouts are relatively worse off today than they were 30 years ago, when the comparable differential was 19 percent. The good news is that over the 1993 to 1997 period, the average weekly wage for college graduates (out of school less than 10 years) was 96 percent higher than for high school graduates. The college graduation wage premium has more than doubled over the past 30 years and provides strong motivation for college students to finish their studies.

[11] David H. Autor, Lawrence Katz, and Alan B. Krueger, "Computing Inequality: Have Computers Changed the Labor Market?" National Bureau of Economic Research, Working Paper No. 5956, March, 1997.

[12] Finis Welch, "In Defense of Inequality," *American Economic Association Papers and Proceedings* 89, No. 2 (May 1999): 1–17

SUMMARY

1. In a competitive input market, the demand for an input is given by the marginal revenue product, the product of the firm's marginal revenue, and the marginal product of the input.
2. A firm in a competitive labor market will hire workers to the point at which the marginal revenue product of labor is equal to the wage rate. This principle is analogous to the profit-maximizing output condition that production be increased to the point at which marginal revenue is equal to marginal cost.
3. The market demand for an input is the horizontal sum of industry demands for the input. But industry demand is not the horizontal sum of the demands of all the firms in the industry. To determine industry demand, one must take into account the fact that the market price of the product will change in response to changes in the price of an input.
4. When factor markets are competitive, the buyer of an input assumes that its purchases will have no effect on its price. As a result, the firm's marginal expendi-

ture and average expenditure curves are both perfectly elastic.

5. The market supply of a factor such as labor need not be upward sloping. A backward-bending labor supply curve can result if the income effect associated with a higher wage rate (more leisure is demanded because it is a normal good) is greater than the substitution effect (less leisure is demanded because its price has gone up).
6. Economic rent is the difference between the payments to factors of production and the minimum payment that would be needed to employ them. In a labor market, rent is measured by the area below the wage level and above the marginal expenditure curve.
7. When a buyer of an input has monopsony power, the marginal expenditure curve lies above the average expenditure curve, which reflects the fact that the monopsonist must pay a higher price to attract more of the input into employment.
8. When the input seller is a monopolist such as a labor union, the seller chooses the point on the marginal revenue product curve that best suits its objective. Maximization of employment, economic rent, and wages are three plausible objectives for labor unions.
9. When a monopolistic union bargains with a monopsonistic employer, the wage rate depends on the nature of the bargaining process. There is little reason to believe that the competitive outcome will be achieved.

QUESTIONS FOR REVIEW

1. Why is a firm's demand for labor curve more inelastic when the firm has monopoly power in the output market than when the firm is producing competitively?
2. Why might a labor supply curve be backward bending?
3. How is a computer company's demand for computer programmers a derived demand?
4. Compare the hiring choices of a monopsonistic and a competitive employer of workers. Which will hire more workers, and which will pay the higher wage? Explain.
5. Rock musicians sometimes earn several million dollars per year. Can you explain such large incomes in terms of economic rent?
6. What happens to the demand for one input when the use of a complementary input increases?
7. For a monopsonist, what is the relationship between the supply of an input and the marginal expenditure on it?
8. Currently the National Football League has a system for drafting college players by which each player is picked by only one team. The player must sign with that team or not play in the league. What would happen to the wages of newly drafted and more experienced football players if the draft system were repealed, and all teams could compete for college players?
9. Why are wages and employment levels indeterminate when the union has monopoly power and the firm has monopsony power?

EXERCISES

1. Assume that workers whose incomes are less than \$10,000 currently pay no federal income taxes. Suppose a new government program guarantees each worker \$5000, whether or not he or she earns any income. For all earned income up to \$10,000, the worker must pay a 50-percent tax. Draw the budget line facing the worker under this new program. How is the program likely to affect the labor supply curve of workers?
2. Using your knowledge of marginal revenue product, explain the following:
   a. A famous tennis star is paid \$200,000 for appearing in a 30-second television commercial. The actor who plays his doubles partner is paid \$500.
   b. The president of an ailing savings and loan is paid *not* to stay in his job for the last two years of his contract.
   c. A jumbo jet carrying 400 passengers is priced higher than a 250-passenger model even though both aircraft cost the same to manufacture.
3. The demands for the factors of production listed below have increased. What can you conclude about changes in the demands for the related consumer goods? If demands for the consumer goods remain unchanged, what other explanation is there for an increase in derived demands for these items?
   a. Computer memory chips
   b. Jet fuel for passenger planes
   c. Paper used for newsprint
   d. Aluminum used for beverage cans
4. Suppose there are two groups of workers, unionized and nonunionized. Congress passes a law that requires all workers to join the union. What do you expect to happen to the wage rates of formerly nonunionized workers? Of those workers who were originally unionized? What have you assumed about the union's behavior?
5. Suppose a firm's production function is given by $Q = 12L - L^2$, for $L = 0$ to 6, where L is labor input per day and Q is output per day. Derive and draw the firm's demand for labor curve if the firm's output sells for \$10 in a competitive market. How many workers will the firm hire when the wage rate is \$30 per day? \$60 per day? (*Hint:* The marginal product of labor is $12 - 2L$.)
6. The only legal employer of military soldiers in the United States is the federal government. If the government uses its knowledge of its monopsonistic position, what criteria will it employ when figuring how many soldiers to recruit? What happens if a mandatory draft is implemented?
7. The demand for labor by an industry is given by the curve $L = 1200 - 10w$, where L is the labor demanded per day and w is the wage rate. The supply curve is given by $L = 20w$. What is the equilibrium wage rate and quantity of labor hired? What is the economic rent earned by workers?
8. Using the same information in Exercise 7, suppose now that the only labor available is controlled by a monopolistic labor union that wishes to maximize the rent earned by union members. What will be the quantity of labor employed and the wage rate? How does your answer compare with your answer to Exercise 7? Discuss. (*Hint:* The union's marginal revenue curve is given by $L = 1200 - 10w$.)

CHAPTER 15

Investment, Time, and Capital Markets

Chapter Outline

List of Examples

In Chapter 14 we saw that in competitive markets, firms decide how much to purchase each month by comparing the marginal revenue product of each factor to its cost. The decisions of all firms determine the market demand for each factor, and the market price is the price that equates the quantity demanded with the quantity supplied. For factor inputs such as labor and raw materials, this picture is reasonably complete, but not so for capital. The reason is that capital is *durable:* it can last and contribute to production for years after it is purchased.

Firms sometimes rent capital much the way they hire workers. For example, a firm might rent office space for a monthly fee, just as it hires a worker for a monthly wage. But more often, capital expenditures involve the purchases of factories and equipment that are expected to last for years. This introduces the element of *time*. When a firm decides whether to build a factory or purchase machines, it must compare the outlays it would have to make *now* with the additional profit that the new capital will generate *in the future*. To make this comparison, it must address the following question: *How much are future profits worth today?* This problem does not arise when hiring labor or purchasing raw materials. To make those choices, the firm need only compare its *current* expenditure on the factor—e.g., the wage or the price of steel—with the factor's *current* marginal revenue product.

In this chapter, we will learn how to calculate the current value of future flows of money. This is the basis for our study of the firm's investment decisions. Most of these decisions involve comparing an outlay today with profits that will be received in the future; we will see how firms can make this comparison and determine whether the outlay is warranted. Often, the future profits resulting from a capital investment may be higher or lower than anticipated. We will see how firms can take this kind of uncertainty into account.

We will also examine other intertemporal decisions that firms sometimes face. For example, producing a depletable resource, such as coal or oil, now means that less will be available to produce in the future. How should a producer take this into account? And how long should a timber company let trees grow before harvesting them for lumber?

In §14.1, we explain that in a competitive factor market, the demand for each factor is given by its marginal revenue product—i.e., the additional revenue earned from an incremental unit of the factor.

The answers to these investment and production decisions depend in part on the *interest rate* that one pays or receives when borrowing or lending money. We will discuss the factors that determine interest rates and explain why interest rates on government bonds, corporate bonds, and savings accounts differ.

15.1 Stocks versus Flows

Recall from §6.1 that a firm's production function involves flows of inputs and outputs: It turns certain amounts of labor and capital each year into an amount of output that same year.

Before proceeding, we must be clear about how to measure capital and other factor inputs that firms purchase. Capital is measured as a *stock*, i.e., as a quantity of plant and equipment that the firm owns. For example, if a firm owns an electric motor factory worth \$10 million, we say that it has a *capital stock* worth \$10 million. Inputs of labor and raw materials, on the other hand, are measured as *flows*, as is the output of the firm. For example, this same firm might use 20,000 worker-hours of labor and 50,000 pounds of copper *per month* to produce 8000 electric motors *per month*. (The choice of monthly units is arbitrary; we could just as well have expressed these quantities in weekly or annual terms—for example, 240,000 worker-hours of labor per year, 600,000 pounds of copper per year, and 96,000 motors per year.)

Let's look at this producer of electric motors in more detail. Both variable cost and the rate of output are flows. Suppose the wage rate is \$15 per hour and the price of copper is 80 cents per pound. Then the variable cost is (20,000)(\$15) + (50,000)(\$0.80) = \$340,000 *per month*. Average variable cost, on the other hand, is a cost *per unit:*

$$\frac{\$340{,}000 \text{ per month}}{8000 \text{ units per month}} = \$42.50 \text{ per unit}$$

Suppose the firm sells its motors for \$52.50 each. Then its average profit is \$52.50 − \$42.50 = \$10.00 per unit, and its total profit is \$80,000 *per month*. (Note that total profit is also a flow.) To make and sell these motors, however, the firm needs capital—namely, the factory that it built for \$10 million. *Thus the firm's \$10 million capital stock allows it to earn a flow of profit of \$80,000 per month.*

Was the \$10 million investment in this factory a sound decision? To answer this question, we must translate the \$80,000 per month profit flow into a number that we can compare with the factory's \$10 million cost. Suppose the factory is expected to last for 20 years. Then, simply put, the problem is: What is the value today of \$80,000 per month for the next 20 years? If that value is greater than \$10 million, the investment was a good one.

A profit of \$80,000 per month for 20 years comes to (\$80,000)(20)(12) = \$19.2 million. That would make the factory seem like an excellent investment. But is \$80,000 five years—or 20 years—from now worth \$80,000 today? No, because money today can be invested—in a bank account, a bond, or other interest-bearing assets—to yield more money in the future. As a result, \$19.2 million received over the next 20 years is worth *less* than \$19.2 million today.

15.2 Present Discounted Value

interest rate Rate at which one can borrow or lend money.

We will return to our \$10 million electric motor factory in Section 15.4, but first we must address a basic problem: *How much is \$1 paid in the future worth today?* The answer depends on the **interest rate:** the rate at which one can borrow or lend money.

Suppose the annual interest rate is R. (Don't worry about which interest rate this actually is; later, we'll discuss the various types of interest rates.) Then \$1 today can be invested to yield $(1 + R)$ dollars a year from now. Therefore, $1 + R$ dollars is the *future value* of \$1 today. Now, what is the value today, i.e., the *present discounted value* (PDV), of \$1 paid one year from now? The answer is easy: Because $1 + R$ dollars one year from now is worth $(1 + R)/(1 + R) = \$1$ today, *\$1 a year from now is worth* $\$1/(1 + R)$ *today*. This is the amount of money that will yield \$1 after one year if invested at the rate R.

What is the value today of \$1 paid *two* years from now? If \$1 were invested today at the interest rate R, it would be worth $1 + R$ dollars after one year, and $(1 + R)(1 + R) = (1 + R)^2$ dollars at the end of two years. Because $(1 + R)^2$ dollars two years from now is worth \$1 today, \$1 two years from now is worth $\$1/(1 + R)^2$ today. Similarly, \$1 paid three years from now is worth $\$1/(1 + R)^3$ today, and \$1 paid n years from now is worth $\$1/(1 + R)^n$ today.[1]

We can summarize this as follows:

$$\text{PDV of \$1 paid after 1 year} = \frac{\$1}{(1 + R)}$$

$$\text{PDV of \$1 paid after 2 years} = \frac{\$1}{(1 + R)^2}$$

$$\text{PDV of \$1 paid after 3 years} = \frac{\$1}{(1 + R)^3}$$

$$\vdots$$

$$\text{PDV of \$1 paid after } n \text{ years} = \frac{\$1}{(1 + R)^n}$$

Table 15.1 shows, for different interest rates, the present value of \$1 paid after 1, 2, 5, 10, 20, and 30 years. Note that for interest rates above 6 or 7 percent, \$1 paid 20 or 30 years from now is worth very little today. But this is not the case for low interest rates. For example, if R is 3 percent, the PDV of \$1 paid 20 years from now is about 55 cents. In other words, if 55 cents were invested now at the rate of 3 percent, it would yield about \$1 after 20 years.

Valuing Payment Streams

We can now determine the present value of a stream of payments over time. For example, consider the two payment streams in Table 15.2. Stream *A* comes to \$200: \$100 paid now and \$100 a year from now. Stream *B* comes to \$220: \$20 paid now, \$100 a year from now, and \$100 two years from now. Which payment stream would you prefer to receive? The answer depends on the interest rate.

[1] We are assuming that the annual rate of interest R is constant from year to year. Suppose the annual interest rate were expected to change, so that R_1 is the rate in year 1, R_2 is the rate in year 2, and so forth. After two years, \$1 invested today would be worth $(1 + R_1)(1 + R_2)$, so that the PDV of \$1 received two years from now is $\$1/(1 + R_1)(1 + R_2)$. Similarly, the PDV of \$1 received n years from now is $\$1/(1 + R_1)(1 + R_2)(1 + R_3) \ldots (1 + R_n)$.

TABLE 15.1 PDV of $1 Paid in the Future

INTEREST RATE	1 YEAR	2 YEARS	5 YEARS	10 YEARS	20 YEARS	30 YEARS
0.01	$0.990	$0.980	$0.951	$0.905	$0.820	$0.742
0.02	0.980	0.961	0.906	0.820	0.673	0.552
0.03	0.971	0.943	0.863	0.744	0.554	0.412
0.04	0.962	0.925	0.822	0.676	0.456	0.308
0.05	0.952	0.907	0.784	0.614	0.377	0.231
0.06	0.943	0.890	0.747	0.558	0.312	0.174
0.07	0.935	0.873	0.713	0.508	0.258	0.131
0.08	0.926	0.857	0.681	0.463	0.215	0.099
0.09	0.917	0.842	0.650	0.422	0.178	0.075
0.10	0.909	0.826	0.621	0.386	0.149	0.057
0.15	0.870	0.756	0.497	0.247	0.061	0.015
0.20	0.833	0.694	0.402	0.162	0.026	0.004

TABLE 15.2 Two Payment Streams

	TODAY	1 YEAR	2 YEARS
Payment Stream *A*:	$100	$100	0
Payment Stream *B*:	$ 20	$100	$100

To calculate the present discounted value of these two streams, we compute and add the present values of each year's payment:

$$\text{PDV of Stream } A = \$100 + \frac{\$100}{(1+R)}$$

$$\text{PDV of Stream } B = \$20 + \frac{\$100}{(1+R)} + \frac{\$100}{(1+R)^2}$$

Table 15.3 shows the present values of the two streams for interest rates of 5, 10, 15, and 20 percent. As the table shows, the preferred stream depends on the interest rate. For interest rates of 10 percent or less, Stream *B* is worth more; for interest rates of 15 percent or more, Stream *A* is worth more. Why? Because less is paid out in Stream *A*, but is paid out sooner.

TABLE 15.3 PDV of Payment Streams

	$R = .05$	$R = .10$	$R = .15$	$R = .20$
PDV of Stream *A*:	$195.24	$190.91	$186.96	$183.33
PDV of Stream *B*:	205.94	193.55	182.58	172.78

EXAMPLE 15.1 The Value of Lost Earnings

In legal cases involving accidents, victims or their heirs (if the victim is killed) sue the injuring party (or an insurance company) to recover damages. In addition to compensating for pain and suffering, those damages include the future income that the injured or deceased person would have earned had the accident not occurred. To see how the present value of lost earnings can be calculated, let's examine an actual 1996 accident case. (The names and some of the data have been changed to preserve anonymity.)

Harold Jennings died in an automobile accident on January 1, 1996, at the age of 53. His family sued the driver of the other car for negligence. A major part of the damages they asked to be awarded was the present value of the earnings that Jennings would have received from his job as an airline pilot had he not been killed. The calculation of present value was typical of cases like this.

Had he worked in 1996, Jennings' salary would have been $85,000. The normal age of retirement for an airline pilot is age 60. To calculate the present value of Jennings' lost earnings, we must take several things into account. First, Jennings' salary would probably have increased over the years. Second, we cannot be sure that he would have lived to retirement had the accident not occurred; he might have died from some other cause. The PDV of his lost earnings until retirement at the end of 2003 is, therefore,

$$\text{PDV} = W_0 + \frac{W_0(1+g)(1-m_1)}{(1+R)} + \frac{W_0(1+g)^2(1-m_2)}{(1+R)^2} + \cdots + \frac{W_0(1+g)^7(1-m_7)}{(1+R)^7}$$

where W_0 is his salary in 1996, g is the annual percentage rate at which his salary is likely to have grown (so that $W_0(1+g)$ would be his salary in 1997, $W_0(1+g)^2$ his salary in 1998, etc.), and $m_1, m_2, \ldots, m_7$ are *mortality rates*, i.e., the probabilities that he would have died from some other cause by 1997, 1998, . . . , 2003.

To calculate this PDV, we need to know the mortality rates $m_1, \ldots, m_7$, the expected rate of growth of Jennings' salary g, and the interest rate R. Mortality data are available from insurance tables that provide death rates for men of

TABLE 15.4 Calculating Lost Wages

YEAR	$W_0(1+g)^t$	$(1-m_t)$	$1/(1+R)^t$	$W_0(1+g)^t(1-m_t)/(1+R)^t$
1996	$ 85,000	.991	1.000	$84,235
1997	91,800	.990	.917	83,339
1998	99,144	.989	.842	82,561
1999	107,076	.988	.772	81,671
2000	115,642	.987	.708	80,810
2001	124,893	.986	.650	80,044
2002	134,884	.985	.596	79,185
2003	145,675	.984	.547	78,409

similar age and race.[2] As a value for g, we can use 8 percent, the average rate of growth of wages for airline pilots over the past decade. Finally, for the interest rate we can use the rate on government bonds, which was about 9 percent. (We will say more about how one chooses the correct interest rate to discount future cash flows in Sections 15.4 and 15.5.) Table 15.4 shows the details of the present value calculation.

By summing the last column, we obtain a PDV of $650,254. If Jennings' family were successful in proving that the defendant was at fault, and if there were no other damage issues involved in the case, they could recover this amount as compensation.[3]

15.3 The Value of a Bond

bond Contract in which a borrower agrees to pay the bondholder (the lender) a stream of money.

A **bond** is a contract in which a borrower agrees to pay the bondholder (the lender) a stream of money. For example, a corporate bond (a bond issued by a corporation) might make "coupon" payments of $100 per year for the next ten years, and then a principal payment of $1000 at the end of the ten-year period.[4] How much would you pay for such a bond? To find out how much the bond is worth, we simply compute the present value of the payment stream:

$$\text{PDV} = \frac{\$100}{(1+R)} + \frac{\$100}{(1+R)^2} + \cdots + \frac{\$100}{(1+R)^{10}} + \frac{\$1000}{(1+R)^{10}} \qquad \textbf{(15.1)}$$

Again, the present value depends on the interest rate. Figure 15.1 shows the value of the bond—the present value of its payment stream—for interest rates up to 20 percent. Note that the higher the interest rate, the lower the value of the bond. At an interest rate of 5 percent, the bond is worth about $1386, but at an interest rate of 15 percent, its value is only $749.

Perpetuities

perpetuity Bond paying out a fixed amount of money each year, forever.

A **perpetuity** is a bond that pays out a fixed amount of money each year, *forever*. How much is a perpetuity that pays $100 per year worth? The present value of the payment stream is given by the infinite summation:

$$\text{PDV} = \frac{\$100}{(1+R)} + \frac{\$100}{(1+R)^2} + \frac{\$100}{(1+R)^3} + \frac{\$100}{(1+R)^4} + \cdots$$

Fortunately, it isn't necessary to calculate and add up all these terms to find the value of this perpetuity; the summation can be expressed in terms of a simple formula:[5]

[2] See, for example, the *Statistical Abstract of the United States*, 1998, Table 130.

[3] Actually, this sum should be reduced by the amount of Jennings' wages which would have been spent on his own consumption and which would not therefore have benefitted his wife or children.

[4] In the United States, the coupon payments on most corporate bonds are made in semiannual installments. To keep the arithmetic simple, we will assume that they are made annually.

[5] Let x be the PDV of $1 per year in perpetutity, so $x = 1/(1+R) + 1/(1+R)^2 + \ldots$. Then $x(1+R) = 1 + 1/(1+R) + 1/(1+R)^2 + \ldots$, so $x(1+R) = 1 + x$, $xR = 1$, and $x = 1/R$.

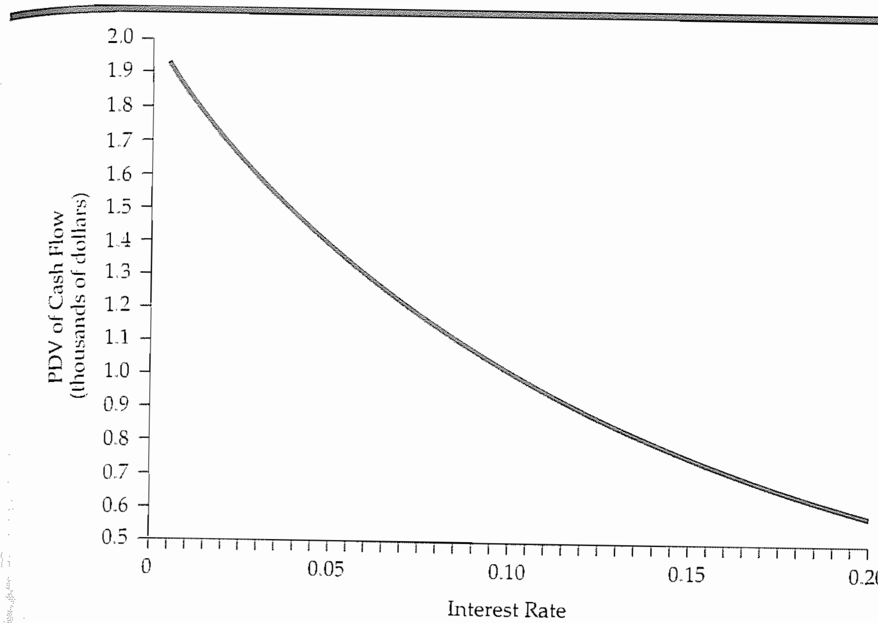

FIGURE 15.1 Present Value of the Cash Flow from a Bond

Because most of the bond's payments occur in the future, the present discounted value declines as the interest rate increases. For example, when the interest rate is 5 percent, the PDV of a 10-year bond paying $100 per year on a principal of $1000 is $1386.

$$\text{PDV} = \$100/R \qquad \textbf{(15.2)}$$

So if the interest rate is 5 percent, the perpetuity is worth $100/(.05) = $2000, but if the interest rate is 20 percent, the perpetuity is worth only $500.

The Effective Yield on a Bond

Many corporate and most government bonds are traded in the *bond market*. The value of a traded bond can be determined directly by looking at its market price, since this is what buyers and sellers agree that the bond is worth.[6] Thus we usually know the value of a bond, but to compare the bond with other investment opportunities, we would like to determine the interest rate consistent with that value.

effective yield (or rate of return) Percentage return that one receives by investing in a bond.

Effective Yield Equations (15.1) and (15.2) show how the values of two different bonds depend on the interest rate used to discount future payments. These equations can be "turned around" to relate the interest rate to the bond's value. This is particularly easy to do for the perpetuity. Suppose the market price—and thus the value—of the perpetuity is P. Then from equation (15.2), $P = \$100/R$, and $R = \$100/P$. So if the price of the perpetuity is $1000, we know that the interest rate is $R = \$100/\$1000 = 0.10$, or 10 percent. This interest rate is called the **effective yield,** or **rate of return.** It is the percentage return that one receives by investing in a bond, which in this case is a perpetuity.

[6] The prices of actively traded corporate and U.S. government bonds are shown in newspapers, such as the *Wall Street Journal* and *New York Times*, and on financial market Web sites such as www.thestreet.com and www.schwab.com.

For the ten-year coupon bond in equation (15.1), calculating the effective yield is a bit more complicated. If the price of the bond is P, we write equation (15.1) as

$$P = \frac{\$100}{(1+R)} + \frac{\$100}{(1+R)^2} + \frac{\$100}{(1+R)^3} + \cdots + \frac{\$100}{(1+R)^{10}} + \frac{\$1000}{(1+R)^{10}}$$

Given the price P, this equation must be solved for R. Although there is no simple formula to express R in terms of P in this case, there are methods (sometimes available on hand-held calculators) for calculating R numerically. Figure 15.2, which plots the same curve that is in Figure 15.1, shows how R depends on P for this ten-year coupon bond. Note that if the price of the bond is \$1000, the effective yield is 10 percent. If the price rises to \$1300, the effective yield drops to about 6 percent. If the price falls to \$700, the effective yield rises to over 16 percent.

Yields can differ considerably among different bonds. Corporate bonds generally yield more than government bonds, and as Example 15.2 shows, the bonds of some corporations yield much more than the bonds of others. One of the most important reasons for this is that different bonds carry different degrees of risk. The U.S. government is less likely to *default* (fail to make interest or principal payments) on its bonds than is a private corporation. And some corporations are financially stronger and therefore less likely to default than others. As we saw in Chapter 5, the more risky an investment, the greater the return that an investor demands. As a result, riskier bonds have higher yields.

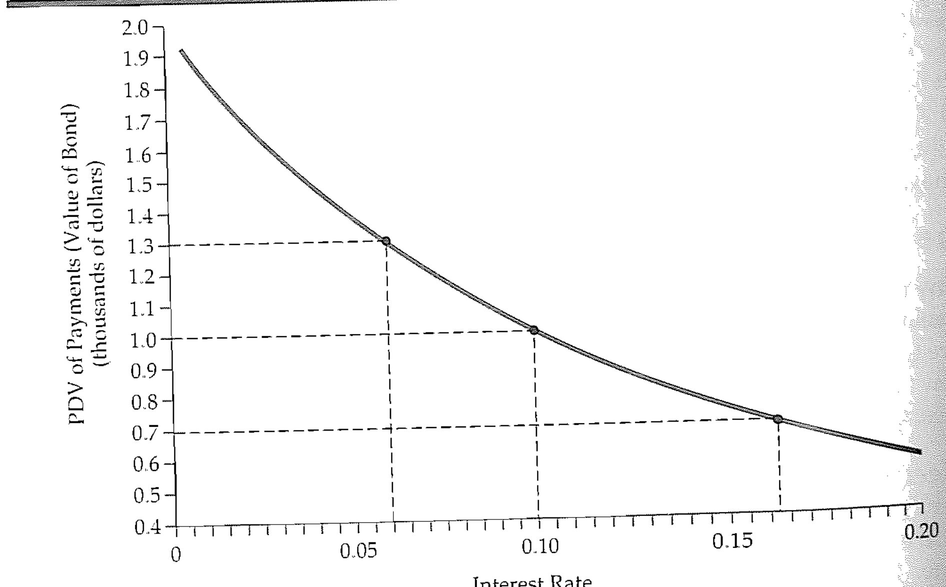

FIGURE 15.2 Effective Yield on a Bond

The effective yield is the interest rate that equates the present value of the bond's payment stream with the bond's market price. The figure shows the present value of the payment stream as a function of the interest rate. The effective yield can thus be found by drawing a horizontal line at the level of the bond's price. For example, if the price of this bond were \$1000, its effective yield would be about 10 percent. If the price were \$1300, the effective yield would be about 6 percent; if the price were \$700, it would be 16.2 percent.

EXAMPLE 15.2 The Yields on Corporate Bonds

To see how corporate bond yields are calculated—and how they can differ from one corporation to another—let's examine the yields for two coupon bonds: one issued by IBM and the other by the Polaroid Corporation. Each has a *face value* of \$100, which means that when the bond matures, the holder receives a principal payment of that amount. Each bond makes a "coupon" (i.e., interest) payment every six months.

We calculate the bond yields using the closing prices on July 23, 1999. The following information on the bonds appeared on the bond page of the newspapers on July 24.

For IBM:

IBM	$5\frac{3}{8}$ 09	5.8	30	92	$-1\frac{1}{2}$

For Polaroid:

Polaroid	$11\frac{1}{2}$ 06	10.8	80	106	$-\frac{5}{8}$

What do these numbers mean? For IBM, $5\frac{3}{8}$ refers to the coupon payments *over one year*. The bond pays \$2.6875 every six months, for a total of \$5.375 per year. The number 09 means that the bond matures in 2009 (at which time the holder will receive \$100 in principal). The next number, 5.8, is the annual coupon divided by the bond's closing price (i.e., 5.375/92). The number 30 refers to the number of these IBM bonds traded that day. The number 92 is the closing price for the bond. Finally, the -1.5 means that the closing price was one and one half points lower than the preceding day's close.[7]

What is the yield on this bond? For simplicity, we'll assume that the coupon payments are made annually instead of every six months. (The error that this introduces is small.) Because the bond matures in 2009, payments will be made for 2009 − 1999 = 10 years. The yield is given by the following equation:

$$92 = \frac{5.375}{(1+R)} + \frac{5.375}{(1+R)^2} + \frac{5.375}{(1+R)^3} + \cdots + \frac{5.375}{(1+R)^{10}} + \frac{100}{(1+R)^{10}}$$

This equation must be solved for R. You can check (by substituting and seeing whether the equation is satisfied) that the solution is $R^* = 6.5$ percent.

The yield on the Polaroid bond is found in the same way. This bond makes coupon payments of \$11.50 per year, matures in the year 2006, and had a closing price of 106. Because the bond has seven years to mature, the equation for its yield is:

$$106 = \frac{11.5}{(1+R)} + \frac{11.5}{(1+R)^2} + \frac{11.5}{(1+R)^3} + \cdots + \frac{11.5}{(1+R)^7} + \frac{100}{(1+R)^7}$$

The solution to this equation is $R^* = 10.2$ percent.

[7] These bonds actually have a face value of \$1000, not \$100. The prices and coupon payments are listed as though the face value were \$100 to save space. To get the actual prices and payments, just multiply the numbers that appear in the newspaper by 10.

Why was the yield on the Polaroid bond higher than that on the IBM bond? Because it was riskier. In 1999, Polaroid's sales and profits had been shrinking, the company was burdened by considerable debt, and its future was highly uncertain. Given Polaroid's more risky financial situation, investors required a higher return before they would buy its bonds.

15.4 The Net Present Value Criterion for Capital Investment Decisions

One of the most common and important decisions that firms make is to invest in new capital. Millions of dollars may be invested in a factory or machines that will last—and affect profits—for many years. The future cash flows that the investment will generate are often uncertain. And once the factory has been built, the firm usually cannot disassemble and resell it to recoup its investment—it becomes a sunk cost.

In §7.1, we explain that a sunk cost is an expenditure that has been made and cannot be recovered.

How should a firm decide whether a particular capital investment is worthwhile? It should calculate the present value of the future cash flows that it expects to receive from the investment, and compare it with the cost of the investment. This method is known as the **net present value (NPV) criterion:**

net present value (NPV) criterion Rule holding that one should invest if the present value of the expected future cash flow from an investment is larger than the cost of the investment.

NPV criterion: Invest if the present value of the expected future cash flows from an investment is larger than the cost of the investment.

Suppose a capital investment costs C and is expected to generate profits over the next 10 years of amounts $\pi_1, \pi_2, \ldots, \pi_{10}$. We then write the net present value as

$$NPV = -C + \frac{\pi_1}{(1+R)} + \frac{\pi_2}{(1+R)^2} + \cdots + \frac{\pi_{10}}{(1+R)^{10}} \qquad (15.3)$$

where R is the **discount rate** that we use to discount the future stream of profits. (R might be a market interest rate or some other rate; we will discuss how to choose it shortly.) Equation (15.3) describes the net benefit to the firm from the investment. The firm should make the investment only if that net benefit is positive—i.e., *only if* $NPV > 0$.

discount rate Rate used to compare the value of a dollar received in the future to the value of a dollar received today.

Determining the Discount Rate What discount rate should the firm use? The answer depends on the alternative ways that the firm could use its money. For example, instead of this investment, the firm might invest in another piece of capital that generates a different stream of profits. Or it might invest in a bond that yields a different return. As a result, we can think of R as the firm's **opportunity cost of capital.** Had the firm not invested in this project, it could have earned a return by investing in something else. *The correct value for R is therefore the return that the firm could earn on a "similar" investment.*

opportunity cost of capital Rate of return that one could earn by investing in an alternate project with similar risk.

By "similar" investment, we mean one with the same *risk*. As we saw in Chapter 5, the more risky an investment, the greater the return one expects to receive from it. Therefore, the opportunity cost of investing in this project is the return that one could earn from another project or asset of similar riskiness.

We'll see how to evaluate the riskiness of an investment in the next section. For now, let's assume that this project has *no risk* (i.e., the firm is sure that the future profit flows will be π_1, π_2, etc.). Then the opportunity cost of the investment is the *risk-free* return—e.g., the return one could earn on a government bond. If the project is expected to last for 10 years, the firm could use the annual interest rate on a 10-year government bond to compute the NPV of the project, as in equation (15.3).[8] If the NPV is zero, the benefit from the investment would just equal the opportunity cost, so the firm should be indifferent between investing and not investing. If the NPV is greater than zero the benefit exceeds the opportunity cost, so the investment should be made.[9]

The Electric Motor Factory

In Section 15.1, we discussed a decision to invest \$10 million in a factory to produce electric motors. This factory would enable the firm to use labor and copper to produce 8000 motors per month for 20 years at a cost of \$42.50 each. The motors could be sold for \$52.50 each, for a profit of \$10 per unit, or \$80,000 per month. We will assume that after 20 years, the factory will be obsolete but can be sold for scrap for \$1 million. Is this a good investment? To find out, we must calculate its net present value.

We will assume for now that the \$42.50 production cost and the \$52.50 price at which the motors can be sold are certain, so that the firm is sure that it will receive \$80,000 per month, or \$960,000 per year, in profit. We also assume that the \$1 million scrap value of the factory is certain. The firm should therefore use a risk-free interest rate to discount future profits. Writing the cash flows in millions of dollars, the NPV is

$$NPV = -10 + \frac{.96}{(1+R)} + \frac{.96}{(1+R)^2} + \frac{.96}{(1+R)^3} + \cdots + \frac{.96}{(1+R)^{20}} + \frac{1}{(1+R)^{20}} \qquad (15.4)$$

Figure 15.3 shows the NPV as a function of the discount rate R. Note that at the rate R^*, which is about 7.5 percent, the NPV is equal to zero. (The rate R^* is sometimes referred to as the *internal rate of return* on the investment.) For discount rates below 7.5 percent, the NPV is positive, so the firm should invest in the factory. For discount rates above 7.5 percent, the NPV is negative, and the firm should not invest.

Real versus Nominal Discount Rates

In the example above, we assumed that future cash flows are certain, so that the discount rate R should be a risk-free interest rate, such as the rate on U.S. government bonds. Suppose that rate happened to be 9 percent. Does that mean the NPV is negative and the firm should not invest?

[8] This is an approximation. To be precise, the firm should use the rate on a one-year bond to discount π_1, the rate on a two-year bond to discount π_2, etc.

[9] This NPV rule is incorrect when the investment is irreversible, subject to uncertainty, and can be delayed. For a treatment of irreversible investment, see Avinash Dixit and Robert Pindyck, *Investment under Uncertainty* (Princeton, NJ: Princeton University Press, 1994).

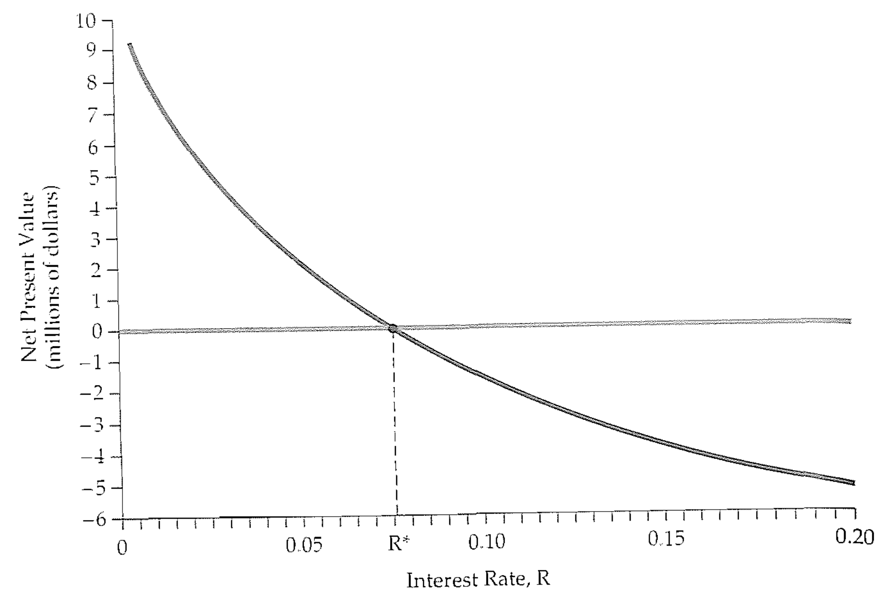

FIGURE 15.3 Net Present Value of a Factory

The NPV of a factory is the present discounted value of all the cash flows involved in building and operating it. Here it is the PDV of the flow of future profits less the current cost of construction. The NPV declines as the interest rate increases. At interest rate R^*, the NPV is zero.

To answer this question, we must distinguish between real and nominal discount rates, and between real and nominal cash flows. Let's begin with the cash flows. In Chapter 1, we discussed real versus nominal prices. We explained that whereas the real price is *net of inflation*, the nominal price includes inflation. In our example, we assumed that the electric motors coming out of our factory could be sold for \$52.50 each over the next 20 years. We said nothing, however, about the effect of inflation. Is the \$52.50 a real price, i.e., net of inflation, or does it include inflation? As we will see, the answer to this question can be critical.

Let's assume that the \$52.50 price—and the \$42.50 production cost—are in real terms. This means that if we expect a 5-percent annual rate of inflation, the nominal price of the motors will increase from \$52.50 in the first year to (1.05)(52.50) = \$55.13 in the second year, to (1.05)(55.13) = \$57.88 in the third year, and so on. Therefore, our profit of \$960,000 per year is also in real terms.

Now let's turn to the discount rate. *If the cash flows are in real terms, the discount rate must also be in real terms.* Why? Because the discount rate is the opportunity cost of the investment. If inflation is not included in the cash flows, it should not be included in the opportunity cost either.

In our example, the discount rate should therefore be the real interest rate on government bonds. The nominal interest rate (9 percent) is the rate that we see in the newspapers; it includes inflation. *The real interest rate is the nominal rate minus the expected rate of inflation.*[10] If we expect inflation to be 5 percent per year on average, the real interest rate would be 9 − 5 = 4 percent. This is the discount

[10] People may have different views about future inflation and may therefore have different estimates of the real interest rate.

rate that should be used to calculate the NPV of the investment in the electric motor factory. Note from Figure 15.3 that at this rate the NPV is clearly positive, so the investment should be undertaken.

When the NPV rule is used to evaluate investments, the numbers in the calculations may be in real or in nominal terms, as long as they are consistent. If cash flows are in real terms, the discount rate should also be in real terms. If a nominal discount rate is used, the effect of future inflation must also be included in the cash flows.

Negative Future Cash Flows

Factories and other production facilities can take several years to build and equip. The cost of the investment will also be spread out over several years, instead of occurring only at the outset. In addition, some investments are expected to result in *losses*, rather than profits, for the first few years. (For example, demand may be low until consumers learn about the product, or costs may start high and fall only when managers and workers have moved down the learning curve.) Negative future cash flows create no problem for the NPV rule; they are simply discounted, just like positive cash flows.

For example, suppose that our electric motor factory will take a year to build: \$5 million is spent right away, and another \$5 million is spent next year. Also, suppose the factory is expected to *lose* \$1 million in its first year of operation and \$0.5 million in its second year. Afterward, it will earn \$0.96 million a year until year 20, when it will be scrapped for \$1 million, as before. (All these cash flows are in real terms.) Now the net present value is

$$\text{NPV} = -5 - \frac{5}{(1+R)} - \frac{1}{(1+R)^2} - \frac{.5}{(1+R)^3} + \frac{.96}{(1+R)^4} + \frac{.96}{(1+R)^5} + \cdots + \frac{.96}{(1+R)^{20}} + \frac{1}{(1+R)^{20}} \qquad (15.5)$$

Suppose the real interest rate is 4 percent. Should the firm build this factory? You can confirm that the NPV is positive, so this project is a good investment.

15.5 Adjustments for Risk

We have seen that a risk-free interest rate is an appropriate discount rate for future cash flows that are certain. For most projects, however, future cash flows are far from certain. At our electric motor factory, for example, we would expect uncertainty over future copper prices, over the future demand and the price of motors, and even over future wage rates. Thus the firm cannot know what its profits from the factory will be over the next 20 years. Its best estimate of profits might be \$960,000 per year, but actual profits may turn out to be higher or lower. How should the firm take this uncertainty into account when calculating the net present value of the project?

A common practice is to increase the discount rate by adding a **risk premium** to the risk-free rate. The idea is that the owners of the firm are risk averse, which makes future cash flows that are risky worth less than those that are certain. Increasing the discount rate takes this into account by reducing the present value of those future cash flows. But how large should the risk premium be? As we will see, the answer depends on the nature of the risk.

risk premium Amount of money that a risk-averse individual will pay to avoid taking a risk.

Diversifiable versus Nondiversifiable Risk

diversifiable risk Risk that can be eliminated either by investing in many projects or by holding the stocks of many companies.

nondiversifiable risk Risk that cannot be eliminated by investing in many projects or by holding the stocks of many companies.

Adding a risk premium to the discount rate must be done with care. If the firm's managers are operating in the stockholders' interests, they must distinguish between two kinds of risk—*diversifiable* and *nondiversifiable*.[11] **Diversifiable risk** can be eliminated by investing in many projects or by holding the stocks of many companies. **Nondiversifiable risk** cannot be eliminated in this way. *Only nondiversifiable risk affects the opportunity cost of capital and should enter into the risk premium.*

To understand this, recall from Chapter 5 that diversifying can eliminate many risks. For example, I cannot know whether the result of a coin flip will be heads or tails. But I can be reasonably sure that out of a thousand coin flips, roughly half will be heads. Similarly, an insurance company that sells me life insurance cannot know how long I will live. But by selling life insurance to thousands of people, it can be reasonably sure about the percentage of those who will die each year.

Much the same is true about capital investment decisions. Although the profit flow from a single investment may be very risky, overall risk will be much less if the firm invests in dozens of projects (as most large firms do). Furthermore, even if the company invests in only one project, stockholders can easily diversify by holding the stocks of a dozen or more different companies, or by holding a mutual fund that invests in many stocks. Thus stockholders—the owners of the firm—can eliminate diversifiable risk.

Because investors can eliminate diversifiable risk, they cannot expect to earn a return higher than the risk-free rate by bearing it: No one will pay you for bearing a risk that there is no need to bear. And indeed, assets that have only diversifiable risk tend on average to earn a return close to the risk-free rate. Now, remember that the discount rate for a project is the opportunity cost of *investing in that project rather than in some other project or asset with similar risk characteristics.* Therefore, if the project's only risk is diversifiable, the opportunity cost is the risk-free rate. *No risk premium should be added to the discount rate.*

What about nondiversifiable risk? First, let's be clear about how it can arise. For a life insurance company, the possibility of a major war poses nondiversifiable risk. Because a war may increase mortality rates sharply, the company cannot expect that an "average" number of its customers will die each year, no matter how many customers it has. As a result, most insurance policies, whether for life, health, or property, do not cover losses resulting from acts of war.

For capital investments, nondiversifiable risk arises because a firm's profits tend to depend on the overall economy. When economic growth is strong, corporate profits tend to be higher. (For our electric motor factory, the demand for motors is likely to be strong, so profits increase.) On the other hand, profits tend to fall in a recession. Because future economic growth is uncertain, diversification cannot eliminate all risk. Investors should (and indeed can) earn higher returns by bearing this risk.

To the extent that a project has nondiversifiable risk, the opportunity cost of investing in that project is higher than the risk-free rate, and a risk premium must be included in the discount rate. Let's see how the size of that risk premium can be determined.

[11] Diversifiable risk is also called *nonsystematic* risk and nondiversifiable risk is called *systematic* risk. Adding a simple risk premium to the discount rate may not always be the correct way of dealing with risk. See, for example, Richard Brealey and Stewart Myers, *Principles of Corporate Finance* (New York: McGraw-Hill, 1999).

The Capital Asset Pricing Model

Capital Asset Pricing Model (CAPM) Model in which the risk premium for a capital investment depends on the correlation of the investment's return with the return on the entire stock market.

The **Capital Asset Pricing Model (CAPM)** measures the risk premium for a capital investment by comparing the expected return on that investment with the expected return on the entire stock market. To understand the model, suppose, first, that you invest in the entire stock market (say, through a mutual fund). In that case, your investment would be completely diversified and you would bear no diversifiable risk. You would, however, bear nondiversifiable risk because the stock market tends to move with the overall economy. (The stock market reflects expected future profits, which depend in part on the economy.) As a result, the expected return on the stock market is higher than the risk-free rate. Denoting the expected return on the stock market by r_m and the risk-free rate by r_f, the risk premium on the market is $r_m - r_f$. This is the additional expected return you get for bearing the nondiversifiable risk associated with the stock market.

Now consider the nondiversifiable risk associated with one asset, such as a company's stock. We can measure that risk in terms of the extent to which the return on the asset tends to be *correlated* with (i.e., move in the same direction as) the return on the stock market as a whole. For example, one company's stock might have almost no correlation with the market as a whole. On average, the price of that stock would move independently of changes in the market, so it would have little or no nondiversifiable risk. The return on that stock should therefore be about the same as the risk-free rate. Another stock, however, might be highly correlated with the market. Its price changes might even amplify changes in the market as a whole. That stock would have substantial nondiversifiable risk, perhaps more than the stock market as a whole. If so, its return on average will exceed the market return r_m.

The CAPM summarizes this relationship between expected returns and the risk premium by the following equation:

$$r_i - r_f = \beta(r_m - r_f) \tag{15.6}$$

where r_i is the expected return on an asset. The equation says that the risk premium on the asset (its expected return less the risk-free rate) is proportional to the risk premium on the market. The constant of proportionality, β, is called the **asset beta.** It measures the sensitivity of the asset's return to market movements and, therefore, the asset's nondiversifiable risk. If a 1-percent rise in the market tends to result in a 2-percent rise in the asset price, the beta is 2. If a 1-percent rise in the market tends to result in a 1-percent rise in the asset price, the beta is 1. And if a 1-percent rise in the market tends to result in no change in the price of the asset, the beta is zero. As equation (15.6) shows, the larger beta, the greater the expected return on the asset. Why? Because the asset's nondiversifiable risk is greater.

asset beta A constant that measures the sensitivity of an asset's return to market movements and, therefore, the asset's nondiversifiable risk.

Given beta, we can determine the correct discount rate to use in computing an asset's net present value. That discount rate is the expected return on the asset or on another asset with the same risk. It is therefore the risk-free rate plus a risk premium to reflect nondiversifiable risk:

$$\text{Discount rate} = r_f + \beta(r_m - r_f) \tag{15.7}$$

Over the past 60 years, the risk premium on the stock market, $(r_m - r_f)$, has been about 8 percent on average. If the real risk-free rate were 4 percent and beta were 0.6, the correct discount rate would be $0.04 + 0.6(0.08) = 0.09$, or 9 percent.

company cost of capital Weighted average of the expected return on a company's stock and the interest rate that it pays for debt.

If the asset is a stock, its beta can usually be estimated statistically.[12] When the asset is a new factory, however, determining its beta is more difficult. Many firms therefore use the company cost of capital as a (nominal) discount rate. The **company cost of capital** is a weighted average of the expected return on the company's stock (which depends on the beta of the stock) and the interest rate that it pays for debt. This approach is correct as long as the capital investment in question is typical for the company as a whole. It can be misleading, however, if the capital investment has much more or much less nondiversifiable risk than the company as a whole. In that case, it may be better to make a reasoned guess as to how much the revenues from the investment are likely to depend on the overall economy.

EXAMPLE 15.3 Capital Investment in the Disposable Diaper Industry

In Example 13.6, we discussed the disposable diaper industry, which has been dominated by Procter & Gamble, with about a 60-percent market share, and Kimberly-Clark, with another 30 percent. We explained that their continuing R&D (research and development) expenditures have given these firms a cost advantage that deters entry. Now we'll examine the capital investment decision of a potential entrant.

Suppose you are considering entering this industry. To take advantage of scale economies, both in production and in advertising and distribution, you would need to build three plants at a cost of $60 million each, with the construction cost spread over three years. When operating at capacity, the plants would produce a total of 2.5 billion diapers per year. These would be sold at wholesale for about 16 cents per diaper, yielding revenues of about $400 million per year. You can expect your variable production costs to be about $290 million per year, for a net revenue of $110 million per year.

You will have other expenses, however. Using the experience of P&G and Kimberly-Clark as a guide, you can expect to spend about $60 million in R&D before start-up to design an efficient manufacturing process, and another $20 million in R&D during each year of production to maintain and improve that process. Finally, once you are operating at full capacity, you can expect to spend another $50 million per year for a sales force, advertising, and marketing, for a net operating profit of $40 million per year. The plants will last for 15 years and will then be obsolete.

Is the investment a good idea? To find out, let's calculate its net present value. Table 15.5 shows the relevant numbers. We assume that production begins at 33 percent of capacity when the plant is completed in 2001, takes two years to reach full capacity, and continues through the year 2016. Given the net cash flows, the NPV is calculated as

$$\text{NPV} = -120 - \frac{93.4}{(1+R)} - \frac{56.6}{(1+R)^2} + \frac{40}{(1+R)^3} + \frac{40}{(1+R)^4} + \cdots + \frac{40}{(1+R)^{15}}$$

[12] You can estimate beta by running a linear regression of the return on the stock against the excess return on the market, $r_m - r_f$. You would find, for example, that the beta for Intel Corporation is about 1.4, the beta for Eastman Kodak is about 0.8, and the beta for General Motors is about 0.5.

TABLE 15.5 Data for NPV Calculation ($ millions)

	PRE-2001	2001	2002	2003	...	2016
Sales		133.3	266.7	400.0	...	400.0
LESS						
Variable cost		96.7	193.3	290.0	...	290.0
Ongoing R&D		20.0	20.0	20.0	...	20.0
Sales force, ads, and promotion		50.0	50.0	50.0	...	50.0
Operating profit		− 33.4	3.4	40.0	...	40.0
LESS						
Construction cost	60.0	60.0	60.0			
Initial R&D	60.0					
NET CASH FLOW	− 120.0	− 93.4	− 56.6	40.0	...	40.0
	Discount Rate:	0.05	0.10	0.15		
	NPV:	80.5	− 16.9	− 75.1		

Table 15.5 shows the NPV for discount rates of 5, 10, and 15 percent.

Note that the NPV is positive for a discount rate of 5 percent, but it is negative for discount rates of 10 or 15 percent. What is the correct discount rate? First, we have ignored inflation, so the discount rate should be in *real* terms. Second, the cash flows are risky—we don't know how efficient our plants will be, how effective our advertising and promotion will be, or even what the future demand for disposable diapers will be. Some of this risk is nondiversifiable. To calculate the risk premium, we will use a beta of 1, which is typical for a producer of consumer products of this sort. Using 4 percent for the real risk-free interest rate and 8 percent for the risk premium on the stock market, our discount rate should be

$$R = 0.04 + 1(0.08) = 0.12$$

At this discount rate, the NPV is clearly negative, so the investment does not make sense. We will not enter the industry; P&G and Kimberly-Clark can breathe a sigh of relief. You should not be surprised, however, that these firms can make money in this market while we cannot. Their experience, years of earlier R&D (they need not spend $60 million on R&D before building new plants), and brand name recognition give them a competitive advantage that a new entrant would find hard to overcome.

15.6 Investment Decisions by Consumers

We have seen how firms value future cash flows and thereby decide whether to invest in long-lived capital. Consumers face similar decisions when they purchase durable goods, such as cars or major appliances. Unlike the decision to purchase food, entertainment, or clothing, the decision to buy a durable good involves comparing a flow of *future* benefits with the *current* purchase cost.

Suppose you are deciding whether to buy a new car. If you keep the car for six or seven years, most of the benefits (and costs of operation) will occur in the future. You must therefore compare the future flow of net benefits from owning the car (the benefit of having transportation less the cost of insurance, maintenance, and gasoline) with the purchase price. Likewise, when deciding whether to buy a new air conditioner, you must compare its price with the present value of the flow of net benefits (the benefit of a cool room less the cost of electricity to operate the unit).

These problems are analogous to the problem of a firm that must compare a future flow of profits with the current cost of plant and equipment when making a capital investment decision. We can therefore analyze these problems just as we analyzed the firm's investment problem. Let's do this for a consumer's decision to buy a car.

The main benefit from owning a car is the flow of transportation services it provides. The value of those services differs from consumer to consumer. Let's assume our consumer values the service at S dollars per year. Let's also assume that the total operating expense (insurance, maintenance, and gasoline) is E dollars per year, that the car costs \$20,000, and that after six years its resale value will be \$4000. The decision to buy the car can then be framed in net present value terms:

$$\text{NPV} = -20{,}000 + (S - E) + \frac{(S - E)}{(1 + R)} + \frac{(S - E)}{(1 + R)^2} + \cdots + \frac{(S - E)}{(1 + R)^6} + \frac{4000}{(1 + R)^6} \qquad (15.8)$$

What discount rate R should the consumer use? The consumer should apply the same principle that a firm does: The discount rate is the opportunity cost of money. If the consumer already has \$20,000 and does not need a loan, the correct discount rate is the return that could be earned by investing the money in another asset—say, a savings account or a government bond. On the other hand, if the consumer is in debt, the discount rate would be the borrowing rate that he or she is already paying. Because this rate is likely to be much higher than the interest rate on a bond or savings account, the NPV of the investment will be smaller.

Consumers must often make trade-offs between up-front versus future payments. An example is the decision of whether to buy or lease a new car. Suppose you can buy a new Toyota Corolla for \$15,000 and after six years, sell it for \$6,000. Alternatively, you could lease the car for \$300 per month for three years, and at the end of the three years, return the car. Which is better—buying or leasing? The answer depends on the interest rate. If the interest rate is very low, buying the car is preferable because the present value of the future lease payments is high. If the interest rate is high, leasing is preferred, because the present value of the future lease payments is low. You will have an opportunity to examine this problem in more detail in Exercise 15.9.

EXAMPLE 15.4 Choosing an Air Conditioner and a New Car

Buying a new air conditioner involves making a trade-off. Some air conditioners cost less but are less efficient—they consume a lot of electricity relative to their cooling power. Others cost more but are more efficient. Should you buy an inefficient air conditioner that costs less now but will cost more to operate in the future, or an efficient one that costs more now but will cost less to operate?

Let's assume you are comparing air conditioners of equivalent cooling power, so that they yield the same flow of benefits. We can then compare the present discounted values of their costs. Assuming an eight-year lifetime and no resale, the PDV of the costs of buying and operating air conditioner i is

$$\text{PDV} = C_i + OC_i + \frac{OC_i}{(1 + R)} + \frac{OC_i}{(1 + R)^2} + \cdots + \frac{OC_i}{(1 + R)^8}$$

where C_i is the purchase price of air conditioner i and OC_i is its average annual operating cost.

The preferred air conditioner depends on your discount rate. If you have little free cash and must borrow, you should use a high discount rate. Because this would make the present value of the future operating costs smaller, you would probably choose a less expensive but relatively inefficient unit. If you have plenty of free cash, so that your opportunity cost of money (and thus your discount rate) is low, you would probably buy the more expensive unit.

An econometric study of household purchases of air conditioners shows that consumers tend to trade off capital costs and expected future operating costs in just this way, although the discount rates that people use are high—about 20 percent for the population as a whole.[13] (American consumers seem to behave myopically by overdiscounting future savings.) The study also shows that consumers' discount rates vary inversely with their incomes. For example, people with above-average incomes used discount rates of about 9 percent, while those in the bottom quarter of the income distribution used discount rates of 39 percent or more. We would expect this result because higher-income people are likely to have more free cash available and therefore have a lower opportunity cost of money.

Buying a new car involves a similar trade-off. One car might cost less than another but offer lower fuel efficiency and require more maintenance and repairs, so that expected future operating costs are higher. As with air conditioners, a consumer can compare two or more cars by calculating and comparing the PDV of the purchase price and expected average annual operating cost for each. An econometric study of automobile purchases found that consumers indeed trade off the purchase price and expected operating costs in this way.[14] It found the average discount rate for all consumers to be in the range of 11 to 17 percent. These discount rate estimates are somewhat lower than those for air conditioners, and probably reflect the widespread availability of auto loans.

*15.7 Intertemporal Production Decisions—Depletable Resources

Production decisions often have *intertemporal* aspects—production today affects sales or costs in the future. The learning curve, which we discussed in Chapter 7, is an example of this. By producing today, the firm gains experience that lowers

Recall from §7.6 that with a learning curve, the firm's cost of production falls over time as managers and workers become more experienced and more effective at using available plant and equipment.

[13] See Jerry A. Hausman, "Individual Discount Rates and the Purchase and Utilization of Energy-Using Durables," *Bell Journal of Economics* 10 (Spring 1979): 33–54.

[14] See Mark K. Dreyfus and W. Kip Viscusi, "Rates of Time Preference and Consumer Valuations of Automobile Safety and Fuel Efficiency," *Journal of Law and Economics* 38 (April 1995): 79–105.

future costs. In this case, production today is partly an investment in future cost reduction, and the value of this must be taken into account when comparing costs and benefits. Another example is the production of a depletable resource. When the owner of an oil well pumps oil today, less oil is available for future production. This must be taken into account when deciding how much to produce.

Production decisions in cases like these involve comparisons between costs and benefits today with costs and benefits in the future. We can make those comparisons using the concept of present discounted value. We'll look in detail at the case of a depletable resource, although the same principles apply to other intertemporal production decisions.

The Production Decision of an Individual Resource Producer

Suppose your rich uncle gives you an oil well. The well contains 1000 barrels of oil that can be produced at a constant average and marginal cost of $10 per barrel. Should you produce all the oil today, or should you save it for the future?[15]

You might think that the answer depends on the profit you can earn if you remove the oil from the ground. After all, why not remove the oil if its price is greater than the cost of extraction? However, this ignores the opportunity cost of using up the oil today so that it is not available for the future.

The correct answer, then, depends not on the current profit level but on how fast you expect the price of oil to rise. Oil in the ground is like money in the bank; you should keep it in the ground only if it earns a return at least as high as the market interest rate. If you expect the price of oil to remain constant or rise very slowly, you would be better off extracting and selling all of it now and investing the proceeds. But if you expect the price of oil to rise rapidly, you should leave it in the ground.

How fast must the price rise for you to keep the oil in the ground? The value of each barrel of oil in your well is equal to the price of oil less the $10 cost of extracting it. (This is the profit you can obtain by extracting and selling each barrel.) This value must rise at least as fast as the rate of interest for you to keep the oil. Your production decision rule is therefore: *Keep all your oil if you expect its price less its extraction cost to rise faster than the rate of interest. Extract and sell all of it if you expect price less cost to rise at less than the rate of interest.* What if you expect price less cost to rise at exactly the rate of interest? Then you would be indifferent between extracting the oil and leaving it in the ground. Letting P_t be the price of oil this year, P_{t+1} the price next year, and c the cost of extraction, we can write this production rule as follows:

If $(P_{t+1} - c) > (1 + R)(P_t - c)$, keep the oil in the ground.

If $(P_{t+1} - c) < (1 + R)(P_t - c)$, sell all the oil now.

If $(P_{t+1} - c) = (1 + R)(P_t - c)$, makes no difference.

Given our expectation about the growth rate of oil prices, we can use this rule to determine production. But how fast should we expect the market price of oil to rise?

[15] For most real oil wells, marginal and average cost are not constant, and it would be extremely costly to extract all the oil in a short time. We will ignore this complication.

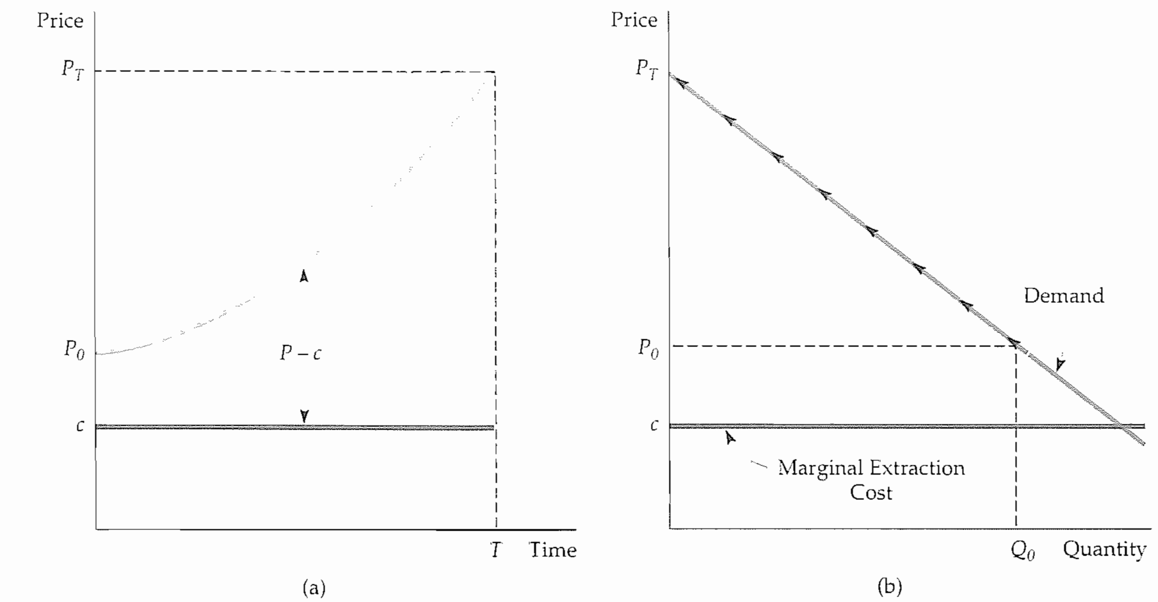

FIGURE 15.4 Price of an Exhaustible Resource

In (a), the price is shown rising over time. Units of a resource in the ground must earn a return commensurate with that on other assets. Therefore, in a competitive market, price less marginal production cost will rise at the rate of interest. Part (b) shows the movement up the demand curve as price rises.

The Behavior of Market Price

Suppose there were no OPEC cartel and the oil market consisted of many competitive producers with oil wells like our own. We could then determine how fast oil prices are likely to rise by considering the production decisions of other producers. If other producers want to earn the highest possible return, they will follow the production rule we stated above. This means that *price less marginal cost must rise at exactly the rate of interest.*[16] To see why, suppose price less cost were to rise faster than the rate of interest. In that case, no one would sell any oil. Inevitably, this would drive up the current price. If, on the other hand, price less cost were to rise at a rate less than the rate of interest, everyone would try to sell all of their oil immediately, which would drive the current price down.

Figure 15.4 illustrates how the market price must rise. The marginal cost of extraction is c, and the price and total quantity produced are initially P_0 and Q_0. Part (a) shows the net price, $P - c$, rising at the rate of interest. Part (b) shows that as price rises, the quantity demanded falls. This continues until time T, when all the oil has been used up and the price P_T is such that demand is just zero.

User Cost

We saw in Chapter 8 that a competitive firm always produces up to the point at which price is equal to marginal cost. However, in a competitive market for an exhaustible resource, price *exceeds* marginal cost (and the difference between price and marginal cost rises over time). Does this conflict with what we learned in Chapter 8?

[16] This result is called the Hotelling rule because it was first demonstrated by Harold Hotelling in "The Economics of Exhaustible Resources," *Journal of Political Economy* 39 (April 1931): 137–75.

No, once we recognize that the *total* marginal cost of producing an exhaustible resource is greater than the marginal cost of extracting it from the ground. There is an additional opportunity cost because producing and selling a unit today makes it unavailable for production and sale in the future. We call this opportunity cost the **user cost of production.** In Figure 15.4, user cost is the difference between price and marginal production cost. It rises over time because as the resource remaining in the ground becomes scarcer, the opportunity cost of depleting another unit becomes higher.

user cost of production Opportunity cost of producing and selling a unit today and so making it unavailable for production and sale in the future.

Resource Production by a Monopolist

What if the resource is produced by a *monopolist* rather than a competitive industry? Should price less marginal cost still rise at the rate of interest?

Suppose a monopolist is deciding between keeping an incremental unit of a resource in the ground, or producing and selling it. The value of that unit is the *marginal revenue* less the marginal cost. The unit should be left in the ground if its value is expected to rise faster than the rate of interest; it should be produced and sold if its value is expected to rise at *less* than the rate of interest. Since the monopolist controls total output, it will produce so that marginal revenue less marginal cost—i.e., the value of an incremental unit of resource—rises at exactly the rate of interest:

In §10.1, we explain that a monopolist maximizes its profit by choosing an output at which marginal revenue is equal to marginal cost.

$$(\text{MR}_{t+1} - c) = (1 + R)(\text{MR}_t - c)$$

Note that this rule also holds for a competitive firm. For a competitive firm, however, marginal revenue equals the market price p.

For a monopolist facing a downward-sloping demand curve, price is greater than marginal revenue. Therefore if marginal revenue less marginal cost rises at the rate of interest, *price* less marginal cost will rise at less than the rate of interest. We thus have the interesting result that a monopolist is *more conservationist* than a competitive industry. In exercising monopoly power, the monopolist starts out charging a higher price and depletes the resource more slowly.

EXAMPLE 15.5 How Depletable Are Depletable Resources?

Resources like oil, natural gas, coal, uranium, copper, iron, lead, zinc, nickel, and helium are all depletable: Because there is a finite amount of each in the earth's crust, the production and consumption of each will ultimately cease. Nonetheless, some resources are more depletable than others.

For oil, natural gas, and helium, known and potentially discoverable in-ground reserves are equal to only 50 to 100 years of current consumption. For these resources, the user cost of depletion can be a significant component of the market price. Other resources, such as coal and iron, have a proved and potential reserve base equal to several hundred or even thousands of years of current consumption. For these resources, the user cost is very small.

The user cost for a resource can be estimated from geological information about existing and potentially discoverable reserves, and from knowledge of the demand curve and the rate at which that curve is likely to shift out over time in response to economic growth. If the market is competitive, user cost can be determined from the economic rent earned by the owners of resource-bearing lands.

TABLE 15.6 User Cost as a Fraction of Competitive Price

RESOURCE	USER COST/COMPETITIVE PRICE
Crude oil	.4 to .5
Natural gas	.4 to .5
Uranium	.1 to .2
Copper	.2 to .3
Bauxite	.05 to .2
Nickel	.1 to .2
Iron ore	.1 to .2
Gold	.05 to .1

Table 15.6 shows estimates of user cost as a fraction of the competitive price for crude oil, natural gas, uranium, copper, bauxite, nickel, iron ore, and gold.[17] Note that only for crude oil and natural gas is user cost a substantial component of price. For the other resources, it is small and in some cases almost negligible. Moreover, although most of these resources have experienced sharp price fluctuations, user cost had almost nothing to do with those fluctuations. For example, oil prices changed because of OPEC and political turmoil in the Persian Gulf, natural gas prices because of changes in government price controls, uranium and bauxite because of cartelization during the 1970s, and copper because of major changes in demand.

Resource depletion, then, has not been very important as a determinant of resource prices over the past few decades. Much more important have been market structure and changes in market demand. But the role of depletion should not be ignored. Over the long term, it will be the ultimate determinant of resource prices.

15.8 How Are Interest Rates Determined?

We have seen how market interest rates are used to help make capital investment and intertemporal production decisions. But what determines interest rate levels? Why do they fluctuate over time? To answer these questions, remember that an interest rate is the price that borrowers pay lenders to use their funds. Like any market price, interest rates are determined by supply and demand—in this case the supply and demand for loanable funds.

The *supply of loanable funds* comes from households that wish to save part of their incomes in order to consume more in the future (or make bequests to their heirs). For example, some households have high incomes now but expect to earn less after retirement. Saving lets them spread their consumption more

[17] These numbers are based on Michael J. Mueller, "Scarcity and Ricardian Rents for Crude Oil," *Economic Inquiry* 23 (1985): 703–24; Kenneth R. Stollery, "Mineral Depletion with Cost as the Extraction Limit: A Model Applied to the Behavior of Prices in the Nickel Industry," *Journal of Environmental Economics and Management* 10 (1983): 151–65; Robert S. Pindyck, "On Monopoly Power in Extractive Resource Markets," *Journal of Environmental Economics and Management* 14 (1987): 128–42; and Martin L. Weitzman, "Pricing the Limits to Growth from Mineral Depletion," *Quarterly Journal of Economics* (May 1999).

evenly over time. In addition, because they receive interest on the money they lend, they can consume more in the future in return for consuming less now. As a result, the higher the interest rate, the greater the incentive to save. The supply of loanable funds is therefore an upward-sloping curve, labeled S in Figure 15.5.

The *demand for loanable funds* has two components. First, some households want to consume more than their current incomes, either because their incomes are low now but are expected to grow, or because they want to make a large purchase (e.g., a house) that must be paid for out of future income. These households are willing to pay interest in return for not having to wait to consume. However, the higher the interest rate, the greater the cost of consuming rather than waiting, so the less willing these households will be to borrow. The household demand for loanable funds is therefore a declining function of the interest rate. In Figure 15.5, it is the curve labeled D_H.

The second source of demand for loanable funds is firms that want to make capital investments. Remember that firms will invest in projects with NPVs that are positive because a positive NPV means that the expected return on the project exceeds the opportunity cost of funds. That opportunity cost—the discount rate used to calculate the NPV—is the interest rate, perhaps adjusted for risk. Often firms borrow to invest because the flow of profits from an investment comes in the future while the cost of an investment must usually be paid now. The desire of firms to invest is thus an important source of demand for loanable funds.

As we saw earlier, however, the higher the interest rate, the lower the NPV of a project. If interest rates rise, some investment projects that had positive NPVs will now have negative NPVs and will therefore be cancelled. Overall, because firms' willingness to invest falls when interest rates rise, their demand for loanable funds also falls. The demand for loanable funds by firms is thus a downward-sloping curve; in Figure 15.5, it is labeled D_F.

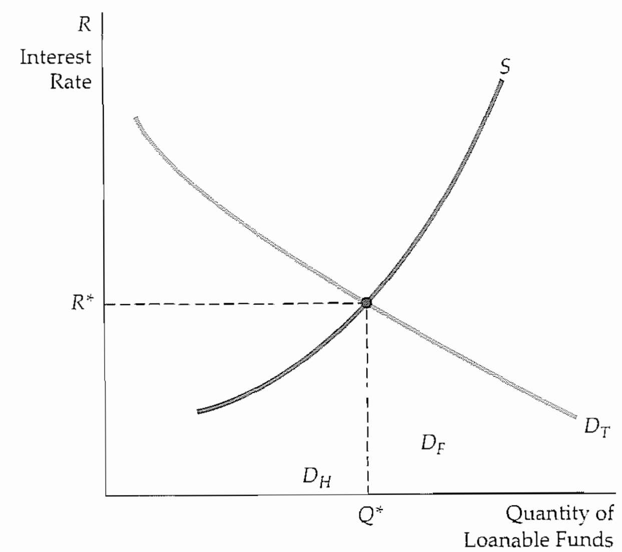

FIGURE 15.5 Supply and Demand for Loanable Funds

Market interest rates are determined by the demand and supply of loanable funds. Households supply funds in order to consume more in the future; the higher the interest rate, the more they supply. Households and firms both demand funds, but the higher the interest rate, the less they demand. Shifts in demand or supply cause changes in interest rates.

The total demand for loanable funds is the sum of household demand and firm demand; in Figure 15.5 it is the curve D_T. This total demand curve together with the supply curve determine the equilibrium interest rate. In Figure 15.5, that rate is R^*.

Figure 15.5 can also help us understand why interest rates change. Suppose the economy goes into a recession. Firms will expect lower sales and lower future profits from new capital investments. The NPVs of projects will fall, and firms' willingness to invest will decline, as will their demand for loanable funds. D_F, and therefore D_T, will shift to the left, and the equilibrium interest rate will fall. Or suppose the federal government spends much more money than it takes through taxes—i.e., that it runs a large deficit. It will have to borrow to finance the deficit, shifting the total demand for loanable funds D_T to the right, so that R increases. The monetary policies of the Federal Reserve are another important determinant of interest rates. The Federal Reserve can create money, shifting the supply of loanable funds to the right and reducing R.

A Variety of Interest Rates

Figure 15.5 aggregates individual demands and supplies as though there were a single market interest rate. In fact, households, firms, and the government lend and borrow under a variety of terms and conditions. As a result, there is a wide range of "market" interest rates. Here we briefly describe some of the more important rates that are quoted in the newspapers and sometimes used for capital investment decisions.

- **Treasury Bill Rate** A Treasury bill is a short-term (one year or less) bond issued by the U.S. government. It is a pure *discount bond*—i.e., it makes no coupon payments but instead is sold at a price less than its redemption value at maturity. For example, a three-month Treasury bill might be sold for \$98. In three months, it can be redeemed for \$100; it thus has an effective three-month yield of about 2 percent and an effective annual yield of about 8 percent.[18] The Treasury bill rate can be viewed as a short-term, risk-free rate.
- **Treasury Bond Rate** A Treasury bond is a longer-term bond issued by the U.S. government for more than one year and typically for 10 to 30 years. Rates vary, depending on the maturity of the bond.
- **Discount Rate** Commercial banks sometimes borrow for short periods from the Federal Reserve. These loans are called *discounts*, and the rate that the Federal Reserve charges on them is the discount rate.
- **Commercial Paper Rate** Commercial paper refers to short-term (six months or less) discount bonds issued by high-quality corporate borrowers. Because commercial paper is only slightly riskier than Treasury bills, the commercial paper rate is usually less than 1 percent higher than the Treasury bill rate.
- **Prime Rate** This is the rate (sometimes called the *reference rate*) that large banks post as a reference point for short-term loans to their biggest corporate borrowers. As we saw in Example 12.4, this rate does not fluctuate from day to day as other rates do.

[18] To be exact, the three-month yield is $(100/98) - 1 = 0.0204$, and the annual yield is $(100/98)^4 - 1 = 0.0842$, or 8.42 percent.

- **Corporate Bond Rate** Newspapers and government publications report the average annual yields on long-term (typically 20-year) corporate bonds in different risk categories (e.g., high-grade, medium-grade, etc.). These average yields indicate how much corporations are paying for long-term debt. However, as we saw in Example 15.2, the yields on corporate bonds can vary considerably, depending on the financial strength of the corporation and the time to maturity for the bond.

SUMMARY

1. A firm's holding of capital is measured as a stock, but inputs of labor and raw materials are flows. Its stock of capital enables a firm to earn a flow of profits over time.
2. When a firm makes a capital investment, it spends money now in order to earn profits in the future. To decide whether the investment is worthwhile, the firm must determine the present value of future profits by discounting them.
3. The present discounted value (PDV) of \$1 paid one year from now is $\$1/(1 + R)$, where R is the interest rate. The PDV of \$1 paid n years from now is $\$1/(1 + R)^n$.
4. A bond is a contract in which a lender agrees to pay the bondholder a stream of money. The value of the bond is the PDV of that stream. The effective yield on a bond is the interest rate that equates that value with the bond's market price. Bond yields differ because of differences in riskiness and time to maturity.
5. Firms can decide whether to undertake a capital investment by applying the net present value (NPV) criterion: Invest if the present value of the expected future cash flows from an investment is larger than the cost of the investment.
6. The discount rate that a firm uses to calculate the NPV for an investment should be the opportunity cost of capital—i.e., the return the firm could earn on a similar investment.
7. When calculating NPVs, if cash flows are in nominal terms (i.e., include inflation), the discount rate should also be nominal; if cash flows are in real terms (i.e., are net of inflation), a real discount rate should be used.
8. An adjustment for risk can be made by adding a risk premium to the discount rate. However, the risk premium should reflect only nondiversifiable risk. Using the Capital Asset Pricing Model (CAPM), the risk premium is the "asset beta" for the project multiplied by the risk premium on the stock market as a whole. The "asset beta" measures the sensitivity of the project's return to movements in the market.
9. Consumers are also faced with investment decisions that require the same kind of analysis as those of firms. When deciding whether to buy a durable good like a car or a major appliance, the consumer must consider the present value of future operating costs.
10. An exhaustible resource in the ground is like money in the bank and must earn a comparable return. Therefore, if the market is competitive, price less marginal extraction cost will grow at the rate of interest. The difference between price and marginal cost is called *user cost*—it is the opportunity cost of depleting a unit of the resource.
11. Market interest rates are determined by the demand and supply of loanable funds. Households supply funds so they can consume more in the future. Households, firms, and the government demand funds. Changes in demand or supply cause changes in interest rates.

QUESTIONS FOR REVIEW

1. A firm uses cloth and labor to produce shirts in a factory that it bought for \$10 million. Which of its factor inputs are measured as flows and which as stocks? How would your answer change if the firm had leased a factory instead of buying one? Is its output measured as a flow or a stock? What about its profit?
2. Suppose the interest rate is 10 percent. If \$100 is invested at this rate today, how much will it be worth after one year? After two years? After five years? What is the value today of \$100 paid one year from now? Paid two years from now? Paid five years from now?
3. You are offered the choice of two payment streams: (a) \$100 paid one year from now and \$100 paid two years from now; (b) \$80 paid one year from now and \$130 paid two years from now. Which payment stream would you prefer if the interest rate is 5 percent? If it is 15 percent?
4. How do investors calculate the present value of a bond? If the interest rate is 5 percent, what is the present value of a perpetuity that pays \$1000 per year forever?
5. What is the *effective yield* on a bond? How does one calculate it? Why do some corporate bonds have higher effective yields than others?
6. What is the net present value (NPV) criterion for investment decisions? How does one calculate the NPV of an investment project? If all the cash flows for a project are certain, what discount rate should be used to calculate NPV?
7. What is the difference between a real discount rate and a nominal discount rate? When should a real discount rate be used in an NPV calculation and when should a nominal rate be used?
8. How is a risk premium used to account for risk in NPV calculations? What is the difference between diversifiable and nondiversifiable risk? Why should only nondiversifiable risk enter into the risk premium?
9. What is meant by the "market return" in the Capital Asset Pricing Model (CAPM)? Why is the market return greater than the risk-free interest rate? What does an asset's "beta" measure in the CAPM? Why should high-beta assets have a higher expected return than low-beta assets?
10. Suppose you are deciding whether to invest \$100 million in a steel mill. You know the expected cash flows for the project, but they are risky—steel prices could rise or fall in the future. How would the CAPM help you select a discount rate for an NPV calculation?
11. How does a consumer trade off current and future costs when selecting an air conditioner or other major appliance? How could this selection be aided by an NPV calculation?
12. What is meant by the "user cost" of producing an exhaustible resource? Why does price minus extraction cost rise at the rate of interest in a competitive exhaustible resource market?
13. What determines the supply of loanable funds? The demand for loanable funds? What might cause the supply or demand for loanable funds to shift, and how would that affect interest rates?

EXERCISES

1. Suppose the interest rate is 10 percent. What is the value of a coupon bond that pays \$80 per year for each of the next five years and then makes a principal repayment of \$1000 in the sixth year? Repeat for an interest rate of 15 percent.
2. A bond has two years to mature. It makes a coupon payment of \$100 after one year and both a coupon payment of \$100 and a principal repayment of \$1000 after two years. The bond is selling for \$966. What is its effective yield?
3. Equation (15.5) shows the net present value of an investment in an electric motor factory. Half of the \$10 million cost is paid initially and the other half after a year. The factory is expected to lose money during its first two years of operation. If the discount rate is 4 percent, what is the NPV? Is the investment worthwhile?
4. The market interest rate is 10 percent and is expected to stay at that level. Consumers can borrow and lend all they want at this rate. Explain your choice in each of the following situations:
   a. Would you prefer a \$500 gift today or a \$540 gift next year?
   b. Would you prefer a \$100 gift now or a \$500 loan without interest for four years?
   c. Would you prefer a \$250 rebate on an \$8000 car or one year of financing for the full price of the car at 5 percent interest?
   d. You have just won a million-dollar lottery and will receive \$50,000 a year for the next 20 years. How much is this worth to you today?
   e. You win the "honest million" jackpot. You can have \$1 million today or \$50,000 per year for eternity (a right that can be passed on to your heirs). Which do you prefer?
   f. In the past, adult children had to pay taxes on gifts of over \$10,000 from their parents, but parents could loan money to their children interest-free. Why did some people call this unfair? To whom were the rules unfair?
5. Ralph is trying to decide whether to go to graduate school. If he spends two years in graduate school, paying \$10,000 tuition each year, he will get a job that will pay \$50,000 per year for the rest of his working life. If he does not go to school, he will go into the workforce immediately. He will then make \$20,000 per year for the next three years, \$30,000 for the following three years, and \$50,000 per year every year after that. If the interest rate is 10 percent, is graduate school a good financial investment?

6. Suppose your uncle gave you an oil well like the one described in Section 15.7. (Marginal production cost is constant at $10.) The price of oil is currently $20 but is controlled by a cartel that accounts for a large fraction of total production. Should you produce and sell all your oil now or wait to produce? Explain your answer.

*7. You are planning to invest in fine wine. Each case costs $100, and you know from experience that the value of a case of wine held for t years is $100t^{1/2}$. One hundred cases of wine are available for sale, and the interest rate is 10 percent.
   a. How many cases should you buy, how long should you wait to sell them, and how much money will you receive at the time of their sale?
   b. Suppose that at the time of purchase, someone offers you $130 per case immediately. Should you take the offer?
   c. How would your answers change if the interest rate were only 5 percent?

8. Reexamine the capital investment decision in the disposable diaper industry (Example 15.3) from the point of view of an incumbent firm. If P&G or Kimberly-Clark were to expand capacity by building three new plants, they would not need to spend $60 million on R&D before start-up. How does this advantage affect the NPV calculations in Table 15.5? Is the investment profitable at a discount rate of 12 percent?

9. Suppose you can buy a new Toyota Corolla for $15,000 and sell it for $6,000 after six years. Alternatively, you can lease the car for $300 per month for three years and return it at the end of the three years. For simplification, assume that lease payments are made yearly instead of monthly—i.e., that they are $3600 per year for each of three years.
   a. If the interest rate, r, is 4 percent, is it better to lease or buy the car?
   b. Which is better if the interest rate is 12 percent?
   c. At what interest rate would you be indifferent between buying and leasing the car?

10. A consumer faces the following decision: She can buy a computer for $1000 and pay $10 per month for Internet access for three years, or she can receive a $400 rebate on the computer (so that its cost is $600) but agree to pay $25 per month for three years for Internet access. For simplification, assume that the consumer pays the access fees yearly (i.e., $10 per month = $120 per year).
   a. What should the consumer do if the interest rate is 3 percent?
   b. What if the interest rate is 17 percent?
   c. At what interest rate is the consumer indifferent between the two options?

PART 4

CHAPTERS

Information, Market Failure, and the Role of Government

MUCH of the analysis of the first three parts of this book has focused on positive questions—how consumers and firms behave and how that behavior affects different market structures. Part IV takes a more normative approach. Here we will describe the goal of economic efficiency, show when markets generate efficient outcomes, and explain when they fail and thus require government intervention.

Chapter 16 discusses general equilibrium analysis, in which the interactions among related markets are taken into account. This chapter also analyzes the conditions that are required for an economy to be efficient and shows when and why a perfectly competitive market is efficient. Chapter 17 examines an important source of market failure—incomplete information. We show that when some economic participants have better information than others, markets may fail to allocate goods efficiently or may not even exist. We also show how sellers can avoid problems of asymmetric information by giving potential buyers signals about product quality. Finally, Chapter 18 discusses two additional sources of market failure: externalities and public goods. We show that although these failures can sometimes be resolved through private bargaining, at other times they require government intervention. We also discuss a number of remedies for market failures, such as pollution taxes and tradeable emission permits.

CHAPTER 16

General Equilibrium and Economic Efficiency

Chapter Outline

List of Examples

For the most part, we have studied individual markets in isolation. But markets are often interdependent: Conditions in one can affect prices and outputs in others either because one good is an input to the production of another good or because two goods are substitutes or complements. In this chapter, we see how a *general equilibrium analysis* can be used to take these interrelationships into account.

We also expand the concept of economic efficiency that we introduced in Chapter 9, and we discuss the benefits of a competitive market economy. To do this, we first analyze economic efficiency, beginning with the exchange of goods among people or countries. We then use this analysis of exchange to discuss whether the outcomes generated by an economy are equitable. To the extent that these outcomes are deemed inequitable, government can help redistribute income.

We then go on to describe the conditions that an economy must satisfy if it is to produce and distribute goods efficiently. We explain why a perfectly competitive market system satisfies those conditions. We also show why free international trade can expand the production possibilities of a country and make its consumers better off. Most markets, however, are not perfectly competitive, and many deviate substantially from that ideal. In the final section of the chapter (as a preview to our detailed discussion of market failure in Chapters 17 and 18), we discuss some key reasons why markets may fail to work efficiently.

16.1 General Equilibrium Analysis

So far our discussions of market behavior have been largely based on **partial equilibrium analysis.** When determining the equilibrium prices and quantitites in a market using partial equilibrium analysis, we presume that activity in one market has little or no effect on other markets. For example, in Chapters 2 and 9, we presumed that the wheat market was largely independent of the markets for related products, such as corn and soybeans.

partial equilibrium analysis Determination of equilibrium prices and quantities in a market independent of effects from other markets.

Often a partial equilibrium analysis is sufficient to understand market behavior. However, market interrelationships can be important. In Chapter 2, for example, we saw how a change in the price of one good can affect the demand for another if they are complements or substitutes. In Chapter 8, we saw that an increase in a firm's input demand can cause both the market price of the input and the product price to rise.

general equilibrium analysis Simultaneous determination of the prices and quantities in all relevant markets, taking feedback effects into account.

Unlike partial equilibrium analysis, **general equilibrium analysis** *determines the prices and quantities in all markets simultaneously*, and it explicitly takes feedback effects into account. A *feedback effect* is a price or quantity adjustment in one market caused by price and quantity adjustments in related markets. Suppose, for example, that the U.S. government taxes oil imports. This would immediately shift the supply curve for oil to the left (by making foreign oil more expensive) and raise the price of oil. But the effect of the tax would not end there. The higher price of oil would increase the demand for and then the price of natural gas. The higher natural gas price would in turn cause oil demand to rise (shift to the right) and increase the oil price even more. The oil and the natural gas markets would continue to interact until eventually an equilibrium would be reached in which the quantity demanded and quantity supplied were equated in both markets.

In practice, a complete general equilibrium analysis, which evaluates the effects of a change in one market on *all* other markets, is not feasible. Instead, we confine ourselves to two or three markets that are closely related. For example, when looking at a tax on oil, we might also look at markets for natural gas, coal, and electricity.

Two Interdependent Markets—Moving to General Equilibrium

To study the interdependence of markets, let's examine the competitive markets for videocassette rentals and movie theater tickets. The two markets are closely related because the widespread ownership of videocassette recorders has given most consumers the option of watching movies at home as well as at the theater. Changes in pricing policies that affect one market are likely to affect the other, which in turn causes feedback effects in the first market.

Figure 16.1 shows the supply and demand curves for videos and movies. In part (a), the price of movie tickets is initially \$6.00; the market is in equilibrium at the intersection of D_M and S_M. In part (b), the video market is also in equilibrium with a price of \$3.00.

Now suppose that the government places a tax of \$1 on each movie ticket purchased. The effect of this tax is determined on a partial equilibrium basis by shifting the supply curve for movies upward by \$1, from S_M to S^*_M in Figure 16.1(a). Initially, this shift causes the prices of movies to increase to \$6.35 and the quantity of movie tickets sold to fall from Q_M to Q'_M. This is as far as a partial equilibrium analysis takes us. But we can go further with a general equilibrium analysis by doing two things: (1) looking at the effects of the movie tax on the market for videos, and (2) seeing whether there are any feedback effects from the video market to the movie market.

In §2.3, we explain that two goods are substitutes if an increase in the price of one leads to an increase in the quantity demanded of the other.

The movie tax affects the market for videos because movies and videos are *substitutes*. A higher movie price shifts the demand for videos from D_V to D'_V in Figure 16.1(b). In turn, this shift causes the rental price of videos to increase from \$3.00 to \$3.50. Note that a tax on one product can affect the prices and sales of other products—something that policymakers should remember when designing tax policies.

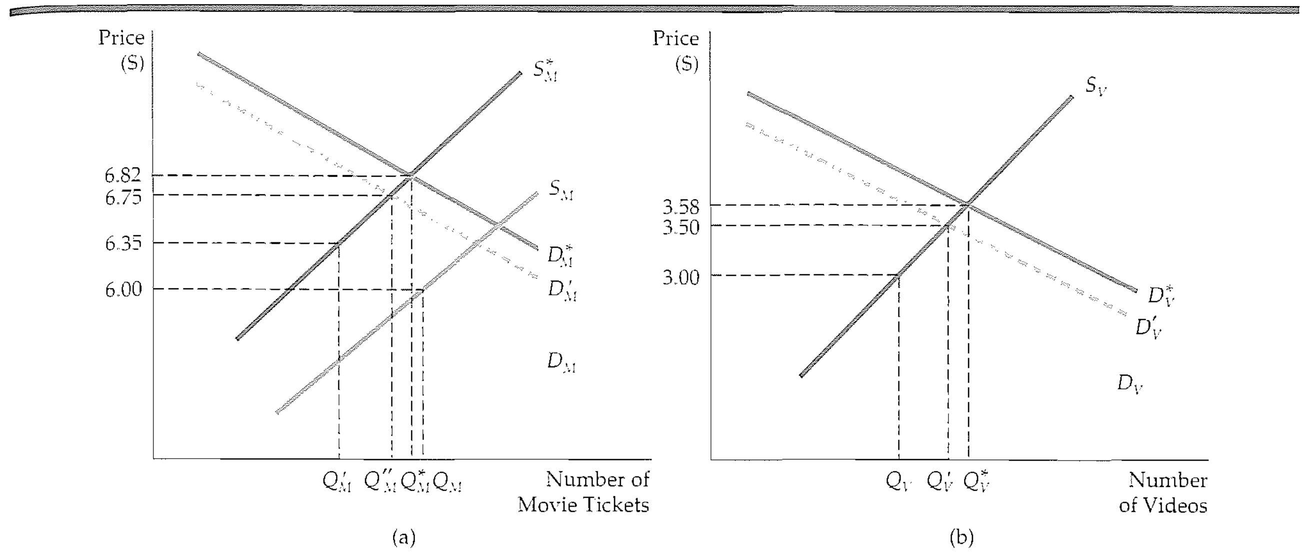

FIGURE 16.1 Two Interdependent Markets: (a) Movie Tickets and (b) Videocassette Rentals

When markets are interdependent, the prices of all products must be simultaneously determined. Here a tax on movie tickets shifts the supply of movies upward from S_M to S^*_M, as shown in (a). The higher price of movie tickets (\$6.35 rather than \$6.00) initially shifts the demand for videocassettes upward (from D_V to D'_V), causing the price of videos to rise (from \$3.00 to \$3.50), as shown in (b). The higher video price feeds back into the movie ticket market, causing demand to shift from D_M to D'_M and the price of movies to increase from \$6.35 to \$6.75. This continues until a general equilibrium is reached, as shown at the intersection of D^*_M and S^*_M in (a), with a movie ticket of \$6.82, and the intersection of D^*_V and S_V in (b), with a video price of \$3.58.

What about the market for movies? The original demand curve for movies presumed that the price of videos was unchanged at \$3.00. But because that price is now \$3.50, the demand for movies will shift upward, from D_M to D'_M in Figure 16.1(a). The new equilibrium price of movies (at the intersection of S^*_M and D'_M) is \$6.75, instead of \$6.35, and the quantity of movie tickets purchased has increased from Q'_M to Q''_M. Thus a partial equilibrium analysis would have underestimated the effect of the tax on the price of movies. The video market is so closely related to the market for movies that to determine the tax's full effect, we need a general equilibrium analysis.

The Attainment of General Equilibrium

Our analysis is not yet complete. The change in the market price of movies will generate a feedback effect on the price of videos that, in turn, will affect the price of movies, and so on. In the end, we must determine the equilibrium prices and quantities of *both* movies and videos *simultaneously*. The equilibrium movie price of \$6.82 is given in Figure 16.1(a) by the intersection of the equilibrium supply and demand curves for movie tickets (S^*_M and D^*_M). The equilibrium video price of \$3.58 is given in Figure 16.1(b) by the intersection of the equilibrium supply and demand curves for videos (S_V and D^*_V). These are the correct general equilibrium prices because the video market supply and demand curves have been drawn *on the assumption that the price of movie tickets is \$6.82*. Likewise, the movie ticket curves have been drawn *on the assumption that the price of videos is \$3.58*. In other words, both sets of curves are consistent with the prices in related markets, and we have no reason to expect that the supply and demand curves in either market will shift further.[1]

Note that even if we were only interested in the market for movies, it would be important to account for the videocassette market when determining the impact of a movie tax. In this example, partial equilibrium analysis would *understate* the effect on the tax, leading us to conclude that the tax will increase the price of movie tickets from $6.00 to $6.35. A general equilibrium analysis, however, shows us that the impact of the tax on the price of movie tickets is greater: It would in fact increase to $6.82.

Movies and videocassettes are substitute goods. By drawing diagrams analogous to those in Figure 16.1, you should be able to convince yourself that if the goods in question are *complements*, a partial equilibrium analysis will *overstate* the impact of a tax. Think about gasoline and automobiles, for example. A tax on gasoline will cause its price to go up, but this will reduce demand for automobiles, which in turn reduces the demand for gasoline, causing its price to fall somewhat.

Recall from §2.3 that two goods are complements if an increase in the price of one leads to a decrease in the quantity demanded of the other.

EXAMPLE 16.1 The Interdependence of International Markets

Because Brazil and the United States compete in the world soybean market, Brazilian regulation of its own soybean market can significantly affect the U.S. market, which in turn can have feedback effects on the Brazilian market. This led to unexpected results when Brazil adopted a regulatory policy aimed at increasing short-run domestic supplies and long-run exports of soybeans.[2]

During the late 1960s and early 1970s, the Brazilian government limited the export of soybeans, causing the price in Brazil to fall. It hoped that making soybeans cheaper in Brazil would both encourage the domestic sale of soybeans and stimulate the domestic demand for soybean products. Eventually the export controls were to be removed, and Brazilian exports were expected to increase.

This expectation was based on a partial equilibrium analysis of the Brazilian soybean market. In fact, the reduction in Brazilian exports increased the price and production of soybeans in the United States, and U.S. exports as well. This made it more difficult for Brazil to export soybeans, even after the controls were removed.

Figure 16.2 shows the consequences of the program. The bottom two lines show Brazilian soybean exports; the top two lines refer to U.S. exports. Actual exports are shown as purple and blue lines. The estimated levels of U.S. and Brazilian exports *had the Brazilian government regulations not gone into effect* are shown as red and green lines, respectively. (The lines diverge from approximately 1970 forward because that is when the major export controls were put into effect.) The figure shows that soybean exports from Brazil would have

[1] To find the general equilibrium prices (and quantities) in practice, we must simultaneously find two prices that equate quantity demanded and quantity supplied in all related markets. For our two markets, this would mean finding the solution to four equations (supply of movie tickets, demand for movie tickets, supply of videos, and demand for videos).

[2] This example presents a simplified version of the analysis in Gary W. Williams and Robert L. Thompson, "Brazilian Soybean Policy: The International Effects of Intervention," *American Journal of Agricultural Economics* 66 (1984): 488–98

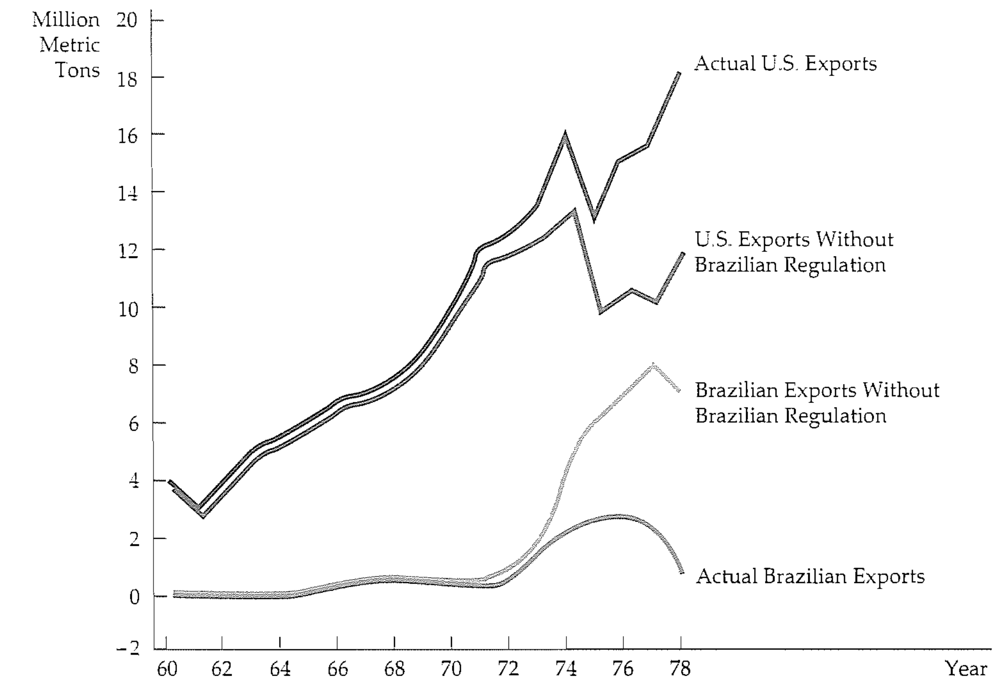

FIGURE 16.2 Soybean Exports—Brazil and the United States

World competition in the soybean market makes the Brazilian and U.S. export markets highly interactive. As a result of the general equilibrium nature of these markets, regulations to stimulate Brazil's market were counterproductive in the long run. Brazil's actual exports of soybeans were lower (and U.S. exports higher) than they would have been without the Brazilian regulation.

been higher, and exports from the United States lower, without the regulatory program. In 1977, for example, Brazilian soybean exports were 73 percent lower than they would have been had the government not intervened. Between 1973 and 1978, however, U.S. soybean exports were over 30 percent higher than they would otherwise have been.

Thus Brazilian soybean policy was misguided and hurt Brazil in the long run. Policymakers failed to take into account the effect of their actions on the U.S. production and export of soybeans.

16.2 Efficiency in Exchange

In Chapter 9 we saw that a competitive market is efficient because it maximizes consumer and producer surplus. To examine the concept of economic efficiency in more detail, we begin with an **exchange economy,** analyzing the behavior of two consumers who can trade either of two goods between themselves. (The analysis also applies to trade between two countries.) Suppose the two goods are initially allocated so that both consumers can make themselves better off by trading with each other. In this case, the initial allocation of goods is economically *inefficient. In an* **efficient allocation** *of goods, no one can be made better off without making someone else worse off.* The term *Pareto efficiency* is sometimes used

exchange economy Market in which two or more consumers trade two goods among themselves.

efficient allocation Allocation of goods in which no one can be made better off unless someone else is made worse off.

synonymously, to credit Italian economist Vilfredo Pareto, who developed the concept of efficiency in exchange. In the subsections that follow, we show why mutually beneficial trades result in an efficient allocation of goods.

The Advantages of Trade

As a rule, voluntary trade between two people or two countries is mutually beneficial.[3] To see how trade makes people better off, let's look in detail at a two-person exchange. Our analysis is based on two important assumptions:

1. Both people know each other's preferences.
2. Exchanging goods is costless.

Suppose James and Karen have 10 units of food and 6 units of clothing between them. Table 16.1 shows that initially James has 7 units of food and 1 unit of clothing, and Karen 3 units of food and 5 units of clothing. To decide whether a trade would be advantageous, we need to know their preferences for food and clothing. Suppose that because Karen has a lot of clothing and little food, her marginal rate of substitution (MRS) of food for clothing is 3: To get 1 unit of food, she will give up 3 units of clothing. However, James's MRS of food for clothing is only 1/2: He will give up only 1/2 a unit of clothing to get 1 unit of food.

In §3.1, we explain that the marginal rate of substitution is the maximum amount of one good that the consumer is willing to give up to obtain one unit of another good.

There is thus room for mutually advantageous trade because James values clothing more highly than Karen does, whereas Karen values food more highly than James does. To get another unit of food, Karen would be willing to trade up to 3 units of clothing. But James will give up 1 unit of food for 1/2 unit of clothing. The actual terms of the trade depend on the bargaining process. Among the possible outcomes are a trade of 1 unit of food by James for anywhere between 1/2 and 3 units of clothing from Karen.

Suppose Karen offers James 1 unit of clothing for 1 unit of food, and James agrees. Both will be better off. James will have more clothing, which he values more than food, and Karen will have more food, which she values more than clothing. Whenever two consumers' MRSs are different, there is room for mutually beneficial trade because the allocation of resources is inefficient: trading will make both consumers better off. Conversely, to achieve economic efficiency, the two consumers' MRSs must be equal.

TABLE 16.1 The Advantage of Trade

INDIVIDUAL	INITIAL ALLOCATION	TRADE	FINAL ALLOCATION
James	7F, 1C	− 1F, + 1C	6F, 2C
Karen	3F, 5C	+ 1F, − 1C	4F, 4C

[3] There are several situations in which trade may not be advantageous. First, limited information may lead people to believe that trade will make them better off when in fact it will not. Second, people may be coerced into making trades, either by physical threats or by the threat of future economic reprisals. Third, as we saw in Chapter 13, barriers to free trade can sometimes provide a strategic advantage to a country.

This important result also holds when there are many goods and consumers: *An allocation of goods is efficient only if the goods are distributed so that the marginal rate of substitution between any pair of goods is the same for all consumers.*

The Edgeworth Box Diagram

If trade is beneficial, which trades can occur? Which of those trades will allocate goods efficiently among customers? How much better off will consumers then be? We can answer these questions for any two-person, two-good example by using a diagram called an **Edgeworth box** named after political economist F. Y. Edgeworth.

Edgeworth box Diagram showing all possible allocations of either two goods between two people or of two inputs between two production processes.

Figure 16.3 shows an Edgeworth box in which the horizontal axis describes the number of units of food and the vertical axis the units of clothing. The length of the box is 10 units of food, the total quantity of food available; its height is 6 units of clothing, the total quantity of clothing available.

In the Edgeworth box, each point describes the market baskets of *both* consumers. James's holdings are read from the origin at O_J and Karen's holdings in the reverse direction from the origin at O_K. For example, point *A* represents the initial allocation of food and clothing. Reading on the horizontal axis from left to right at the bottom of the box, we see that James has 7 units of food, and reading upward along the vertical axis on the left of the diagram, we see that he has 1 unit of clothing. For James, therefore, *A* represents 7F and 1C. This leaves 3F and 5C for Karen. Karen's allocation of food (3F) is read from right to left at the top of the box diagram beginning at O_K; we read her allocation of clothing (5C) from top to bottom at the right of the box diagram.

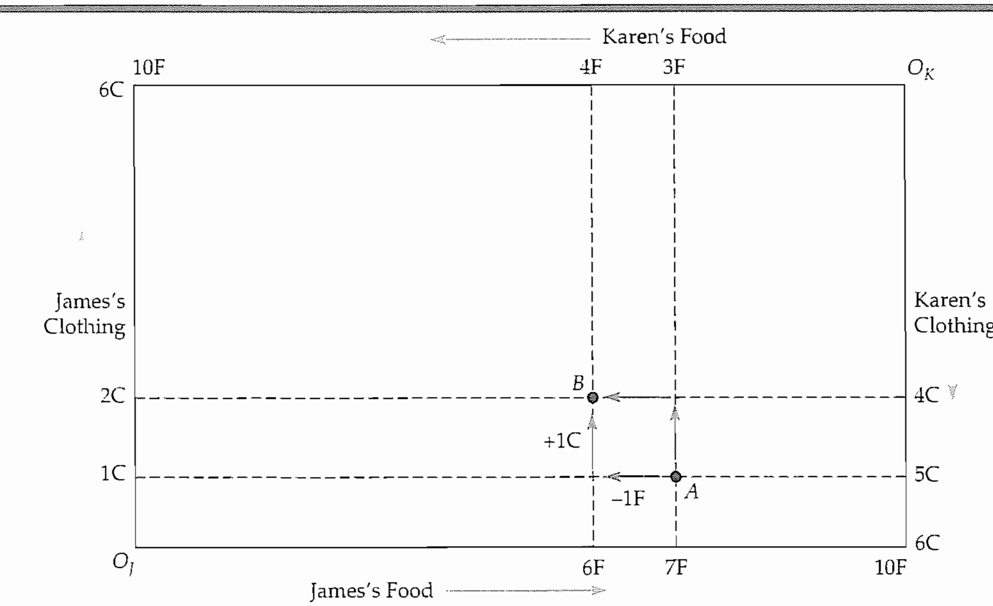

FIGURE 16.3 Exchange in an Edgeworth Box

Each point in the Edgeworth box simultaneously represents James's and Karen's market baskets of food and clothing. At *A*, for example, James has 7 units of food and 1 unit of clothing, and Karen 3 units of food and 5 units of clothing.

We can also see the effect of trade between Karen and James. James gives up 1F in exchange for 1C, moving from A to B. Karen gives up 1C and obtains 1F, also moving from A to B. Point B thus represents the market baskets of both James and Karen after the mutually beneficial trade.

Efficient Allocations

A trade from A to B thus made both Karen and James better off. But is B an *efficient* allocation? The answer depends on whether James's and Karen's MRSs are the same at B, which depends in turn on the shape of their indifference curves. Figure 16.4 shows several indifference curves for both James and Karen. Because his allocations are measured from the origin O_J, James's indifference curves are drawn in the usual way. But for Karen, we have rotated the indifference curves 180 degrees, so that the origin is at the upper right-hand corner of the box. Karen's indifference curves are convex, just like James's; we simply see them from a different perspective.

Now that we are familiar with the two sets of indifference curves, let's examine the curves labeled U_J^1 and U_K^1 that pass through the initial allocation at A. Both James's and Karen's MRSs give the slope of their indifference curves at A. James's is equal to 1/2, Karen's to 3. The shaded area between these two indifference curves represents all possible allocations of food and clothing that would make both James and Karen better off than at A. In other words, it describes all possible mutually beneficial trades.

Starting at A, any trade that moved the allocation of goods outside the shaded area would make one of the two consumers worse off and should not occur. The move from A to B was mutually beneficial. But in Figure 16.4, B is not an efficient point because indifference curves U_J^2 and U_K^2 intersect. In this case, James's and

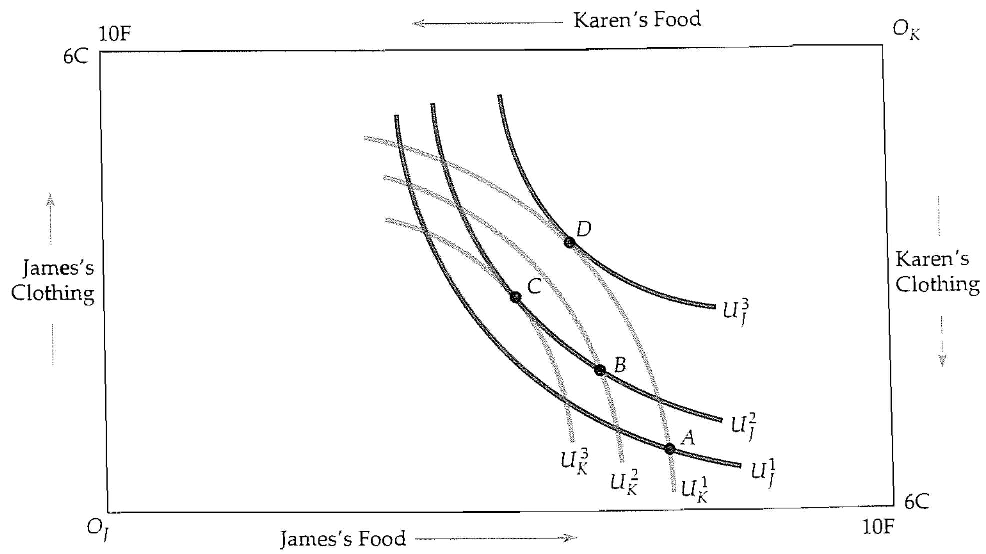

FIGURE 16.4 Efficiency in Exchange

The Edgeworth box illustrates the possibilities for both consumers to increase their satisfaction by trading goods. If A gives the initial allocation of resources, the shaded area describes all mutually beneficial trades.

Karen's MRSs are not the same and the allocation is not efficient. This illustrates an important point: *Even if a trade from an inefficient allocation makes both people better off, the new allocation is not necessarily efficient.*

Suppose that from B an additional trade is made, with James giving up another unit of food to obtain another unit of clothing and Karen giving up a unit of clothing for a unit of food. Point C in Figure 16.4 gives the new allocation. At C, the MRSs of both people are identical, which is why the indifference curves are tangent there. When the indifference curves are tangent, one person cannot be made better off without making the other person worse off. Therefore, C represents an efficient allocation.

Of course, C is not the only possible efficient outcome of a bargain between James and Karen. For example, if James is an effective bargainer, a trade might change the allocation of goods from A to D, where indifference curve U_J^3 is tangent to indifference curve U_K^1. This allocation would leave Karen no worse off than she was at A and James much better off. And because no further trade is possible, D is an efficient allocation. Thus C and D are both efficient allocations, although James prefers D to C and Karen C to D. In general, it is difficult to predict the allocation that will be reached in a bargain because the end result depends on the bargaining abilities of the people involved.

The Contract Curve

We have seen that from an initial allocation many possible efficient allocations can be reached through mutually beneficial trade. To find *all possible efficient allocations of food and clothing* between Karen and James, we look for *all points of tangency between each of their indifference curves.* Figure 16.5 shows the curve drawn through all such efficient allocations; it is called the **contract curve.**

contract curve Curve showing all efficient allocations of goods between two consumers, or of two inputs between two production functions.

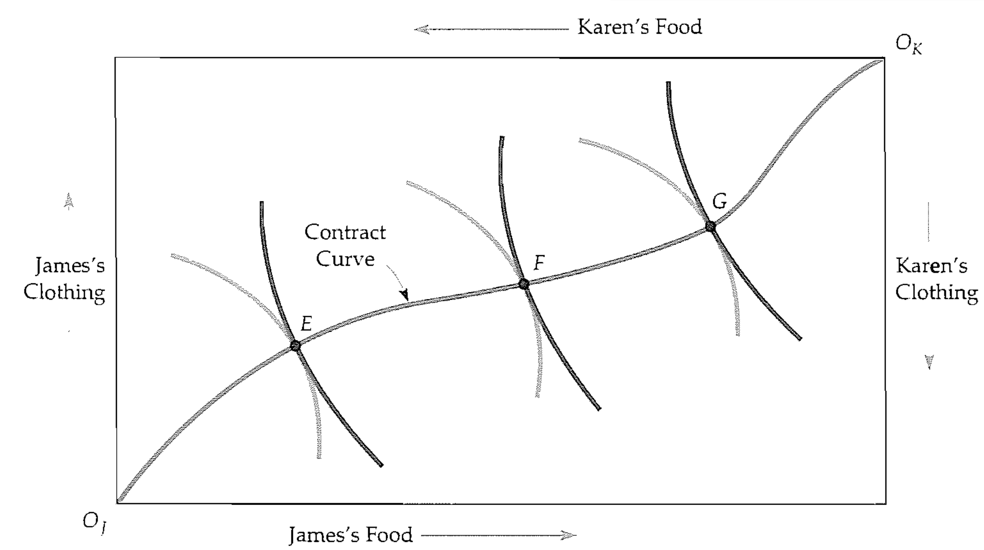

FIGURE 16.5 The Contract Curve

The contract curve contains all allocations for which consumers' indifference curves are tangent. Every point on the curve is efficient because one person cannot be made better off without making the other person worse off.

The contract curve shows all allocations from which no mutually beneficial trade can be made. *These allocations are efficient if there is no way to reallocate goods to make someone better off without making someone else worse off.* In Figure 16.5 three allocations labeled *E*, *F*, and *G* are Pareto efficient, although each involves a different distribution of food and clothing, because one person could not be made better off without making someone else worse off.

Several properties of the contract curve may help us understand the concept of efficiency in exchange. Once a point on a contract curve, such as *E*, has been chosen, there is no way to move to another point on the contract curve, say *F*, without making one person worse off (in this case, Karen). Without making further comparison between James's and Karen's preferences, we cannot compare allocations *E* and *F*. We simply know that both are efficient. In this sense, Pareto efficiency is a modest goal: It says that we should make all mutually beneficial exchanges, but it does not say which exchanges are best. Pareto efficiency can be a powerful concept, however. If a change will improve efficiency, it is in *everyone's* self-interest to support it.

We can frequently improve efficiency even when one aspect of a proposed change makes someone worse off. We need only include a second change, such that the *combined* set of changes leaves someone better off and no one worse off. Suppose, for example, that we propose to eliminate a quota on automobile imports into the United States. Although U.S. consumers would then enjoy lower prices and a greater selection of cars, some U.S. autoworkers would lose their jobs. But what if eliminating the quota were combined with federal tax breaks and job relocation subsidies for autoworkers? In that case, U.S. consumers would be better off (after accounting for the cost of the job subsidies) and U.S. autoworkers no worse off, and the result would increase efficiency.

Consumer Equilibrium in a Competitive Market

In a two-person exchange, the outcome can depend on the bargaining power of the two parties. Competitive markets, however, have many actual or potential buyers and sellers. Thus, if people do not like the terms of an exchange, they can look for another seller who offers better terms. As a result, each buyer and seller takes the price of the goods as fixed and decides how much to buy and sell at those prices. We can show how competitive markets lead to efficient exchange by using the Edgeworth box to mimic a competitive market. Suppose, for example, that there are many Jameses and many Karens. This allows us to think of each individual James and Karen as a price taker, even though we are working with only a two-person box diagram.

Figure 16.6 shows the opportunities for trade when we start at the allocation given by point *A* and when the prices of both food and clothing are equal to 1. (The actual prices do not matter; what matters is the price of food relative to the price of clothing.) When the prices of food and clothing are equal, each unit of food can be exchanged for 1 unit of clothing. As a result, the price line *PP'* in the diagram, which has a slope of -1, describes all possible allocations that exchange can achieve.

Suppose each James decides to buy 2 units of clothing and sell 2 units of food in exchange. This would move each James from *A* to *C* and increase satisfaction from indifference curve U_J^1 to U_J^2. Meanwhile, each Karen buys 2 units of food and sells 2 units of clothing. This would move each Karen from *A* to *C* as well, increasing satisfaction from indifference curve U_K^1 to U_K^2.

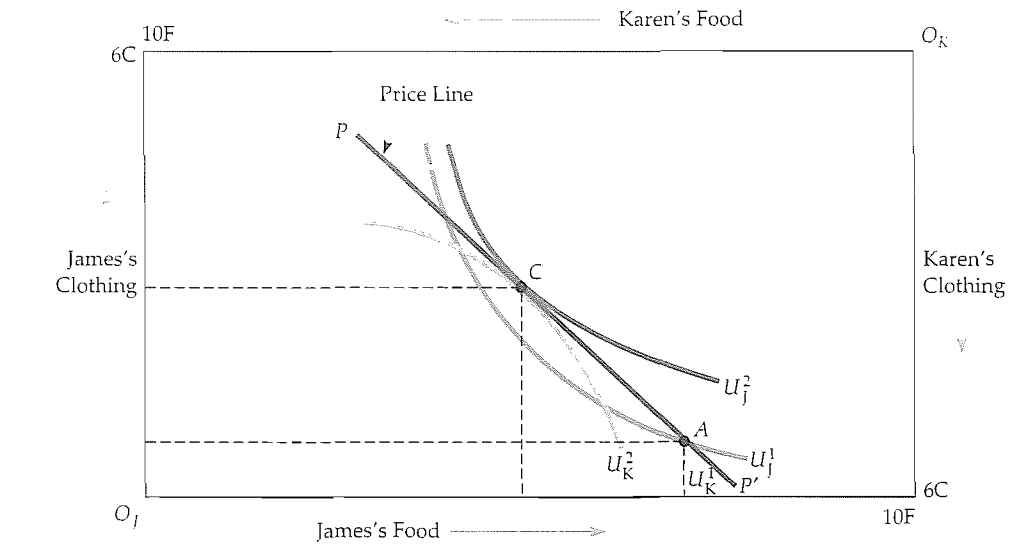

FIGURE 16.6 Competitive Equilibrium

In a competitive market the prices of the two goods determine the terms of exchange among consumers. If *A* is the initial allocation of goods and the price line *PP'* represents the ratio of prices, the competitive market will lead to an equilibrium at *C*, the point of tangency of both indifference curves. As a result, the competitive equilibrium is efficient.

We choose prices for the two goods so that the quantity of food demanded by each Karen is equal to the quantity of food that each James wishes to sell; likewise, the quantity of clothing demanded by each James is equal to the quantity of food that each Karen wishes to sell. As a result, the markets for food and clothing are in equilibrium. An *equilibrium is a set of prices at which the quantity demanded equals the quantity supplied in every market.* This is also a *competitive equilibrium* because all suppliers and demanders are price takers.

In §8.7, we explain that in a competitive equilibrium, price-taking firms maximize profit, and that the price of the product is such that the quantity demanded is equal to the quantity supplied.

Not all prices are consistent with an equilibrium. For example, if the price of food is 1 and the price of clothing 3, food must be exchanged for clothing on a 3-to-1 basis. But then each James will be unwilling to trade any food to get additional clothing because his MRS of clothing for food is only 1/2. Each Karen, on the other hand, would be happy to sell clothing to get more food, but has no one to trade with. The market is therefore in *disequilibrium* because the quantity demanded is not equal to the quantity supplied.

This disequilibrium should only be temporary. In a competitive market, prices will adjust if there is **excess demand** in some markets (the quantity demanded of one good is greater than the quantity supplied) and **excess supply** in others (the quantity supplied is greater than the quantity demanded). In our example, each Karen's quantity demanded for food is greater than each James's willingness to sell it, whereas each Karen's willingness to trade clothing is greater than each James's quantity demanded. As a result of this excess quantity demanded for food and excess quantity supplied of clothing, we can expect the price of food to increase relative to the price of clothing. As the price changes, so will the quantities demanded by all those in the market. Eventually, the prices will adjust until an equilibrium is reached. In our example, the price of both food

excess demand When the quantity demanded of a good exceeds the quantity supplied.

excess supply When the quantity supplied of a good exceeds the quantity demanded.

and clothing might be 2; we know from the previous analysis that when the price of clothing is equal to the price of food, the market will be in competitive equilibrium. (Recall that only relative prices matter; prices of 2 for clothing and food are equivalent to prices of 1 for each.)

Note the important difference between exchange with two people and an economy with many people. When only two people are involved, bargaining leaves an indeterminate outcome. However, when many people are involved, the prices of the goods are determined by the combined choices of demanders and suppliers of goods.

The Economic Efficiency of Competitive Markets

We can now understand one of the fundamental results of microeconomic analysis. We can see from point C in Figure 16.6 that *the allocation in a competitive equilibrium is economically efficient.* The key reason why this is so is that C must occur at the tangency of two indifference curves. If it does not, one of the Jameses or one of the Karens will not be achieving maximum satisfaction; he or she will be willing to trade to achieve a higher level of utility.

This result holds in an exchange framework and in a general equilibrium setting in which all markets are perfectly competitive. It is the most direct way of illustrating the workings of Adam Smith's famous *invisible hand,* because it tells us that the economy will automatically allocate resources efficiently without the need for regulatory control. It is the independent actions of consumers and producers, who take prices as given, that allows markets to function in an economically efficient manner. Not surprisingly, the invisible-hand result is often used as the norm against which the workings of all real-world markets are compared. For some, the invisible hand supports the normative argument for less government intervention; they argue that markets are highly competitive. For others, the invisible hand supports a more expansive role for government; they reply that intervention is needed to make markets more competitive.

Whatever one's view of government intervention, most economists consider the invisible-hand result important. In fact, the result that a competitive equilibrium is economically efficient is often described as the first theorem of **welfare economics.** (Welfare economics involves the normative evaluation of markets and economic policy.) Formally, the first theorem states the following:

welfare economics Normative evaluation of markets and economic policy.

> If everyone trades in the competitive marketplace, all mutually beneficial trades will be completed and the resulting equilibrium allocation of resources will be economically efficient.

Let's summarize what we know about a competitive equilibrium from the consumer's perspective:

1. Because the indifference curves are tangent, all marginal rates of substitution between consumers are equal.
2. Because each indifference curve is tangent to the price line, each person's MRS of clothing for food is equal to the ratio of the prices of the two goods.

Formally, if P_C and P_F are the two prices

$$MRS^J_{FC} = P_C/P_F = MRS^K_{FC} \quad (16.1)$$

To achieve an efficient allocation when there are many consumers (and many producers) is not easy. It can be done if all markets are perfectly competitive. But efficient outcomes can also be achieved by other means—for example, through a centralized system in which the government allocates all goods and services. The competitive solution is often preferred because it allocates resources with a minimum of information. All consumers must know their own preferences and the prices they face, but they need not know what is being produced or the demands of other consumers. Other allocation methods need more information, and as a result they become difficult and cumbersome to manage.

16.3 Equity and Efficiency

We have shown that different efficient allocations of goods are possible, and we have seen how a perfectly competitive economy generates an efficient allocation. But some allocations are likely to be more fair than others. How do we decide what is the most *equitable* allocation? That is a difficult question—economists and others disagree both about how to define *equity* and how to quantify it. Any such view would involve subjective comparisons of utility, and reasonable people could disagree about how to make these comparisons. In this section, we discuss this general point and then illustrate it in a particular case by showing that there is no reason to believe that the allocation associated with a competitive equilibrium will be equitable.

The Utility Possibilities Frontier

Recall that every point on the contract curve in our two-person exchange economy shows the levels of utility that James and Karen can achieve. In Figure 16.7 we put the information from the Edgeworth box in a different form. James's utility is measured on the horizontal axis and Karen's on the vertical axis. Any point in the Edgeworth box corresponds to a point in Figure 16.7 because every allocation generates utility for both people. Every movement to the right in Figure 16.7 represents an increase in James's utility, and every upward movement an increase in Karen's.

The **utility possibilities frontier** *represents all allocations that are efficient.* It shows the levels of satisfaction that are achieved when the two individuals have reached the contract curve. Point O_J is one extreme at which James has no goods and therefore zero utility, while point O_K is the opposite extreme at which Karen has no goods. All other points on the frontier, such as E, F, and G, correspond to points on the contract curve, so that one person cannot be made better off without making the other worse off. Point H, however, represents an inefficient allocation because any trade within the shaded area makes one or both parties better off. At L, both people would be better off, but L is not attainable because there is not enough of both goods to generate the levels of utility that the point represents.

utility possibilities frontier Curve showing all efficient allocations of resources measured in terms of the utility levels of two individuals.

It might seem reasonable to conclude that an allocation must be efficient to be equitable. Compare point H with F and E. Both F and E are efficient, and (relative to H) each makes one person better off without making the other worse off. We might agree, therefore, that it is inequitable to James or Karen or both for an economy to yield allocation H as opposed to F or E.

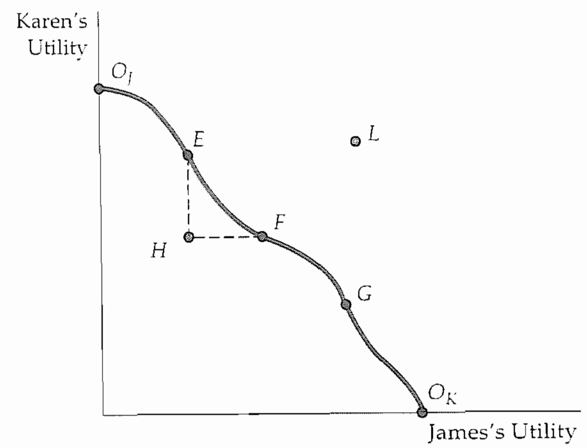

FIGURE 16.7 Utility Possibilities Frontier

The utility possibilities frontier shows the levels of satisfaction that each of two people achieve when they have traded to an efficient outcome on the contract curve. Points E, F, and G correspond to points on the contract curve and are efficient. Point H is inefficient because any trade within the shaded area will make one or both people better off.

But suppose H and G are the only possible allocations. Is G more equitable than H? Not necessarily. Compared with H, G yields more utility for James and less for Karen. Some people may feel that G is more equitable than H; others may feel the opposite. We can conclude, therefore, that *one inefficient allocation of resources may be more equitable than another efficient allocation.*

The problem is how to define an equitable allocation. Even if we restrict ourselves to all points on the utility possibilities frontier, we can still ask which of these points is the most equitable. *The answer depends on what one thinks equity entails* and, therefore, on the interpersonal comparisons of utility that one is willing to make.

social welfare function Weights applied to each individual's utility in determining what is socially desirable.

Social Welfare Functions In economics, we often use a **social welfare function** to describe the particular weights that are applied to each individual's utility in determining what is socially desirable. One social welfare function, the *utilitarian*, weights everyone's utility equally and consequently maximizes the total utility of all members of society. Each social welfare function can be associated with a particular view about equity. But some views do not explicitly weight individual utilities and cannot therefore be represented by a social welfare function. For example, a market-oriented view argues that the outcome of the competitive market process is equitable because it rewards those who are most able and who work the hardest. If E is the competitive equilibrium allocation, for example, E would be deemed to be more equitable than F, even though goods are less equally allocated.

When more than two people are involved, the meaning of the word equity becomes even more complex. The *Rawlsian* view[4] emphasizes that because the wealth they achieve will be taxed away, an equal distribution of resources may remove the incentive that spurs the most productive people to work hard. This

[4] See John Rawls, *A Theory of Justice* (New York: Oxford University Press, 1971).

TABLE 16.2 Four Views of Equity

1. Egalitarian—all members of society receive equal amounts of goods
2. Rawlsian—maximize the utility of the least-well-off person
3. Utilitarian—maximize the total utility of all members of society
4. Market-oriented—the market outcome is the most equitable

view allows inequalities, if they make the least-well-off person in society better off. According to Rawls, *the most equitable allocation maximizes the utility of the least-well-off person in society.* The Rawlsian perspective could be *egalitarian*—involving an equal allocation of goods among all members of society—but it need not be. Suppose that by rewarding more productive people more highly than less productive people, we can get the most productive people to work harder. This could produce more goods and services, some of which could then be reallocated to make the poorest members of society better off.

The four views of equity in Table 16.2 move roughly from most to least egalitarian. The egalitarian view explicitly requires equal allocations, while the Rawlsian puts a heavy weight on equality (otherwise some would be much worse off than others). The utilitarian is likely to require some difference between the best- and worst-off members of society. Finally, the market-oriented view may lead to substantial inequality in the allocations of goods and services.

Equity and Perfect Competition

A competitive equilibrium leads to a Pareto efficient outcome that may or may not be equitable. In fact, a competitive equilibrium could occur at any point on the contract curve, depending on the initial allocation. Imagine, for example, that the initial allocation gave all food and clothing to Karen. This would be at O_J in Figure 16.7, and Karen would have no reason to trade. Point O_J would then be a competitive equilibrium, as would point O_K and all intermediate points on the contract curve.

Because efficient allocations are not necessarily equitable, society must rely to some extent on government to achieve equity goals by redistributing income or goods among households. These goals can be reached through the tax system. For example, a progressive income tax whose funds are used for programs that benefit households proportionally to income will redistribute income from the wealthy to the poor. The government can also provide public services, such as medical aid to the poor (Medicaid), or it can transfer funds through such programs as Food Stamps.

The result that a competitive equilibrium can sustain every point on the contract curve is a fundamental result in microeconomics. It is important because it suggests an answer to a basic normative question: Is there a trade-off between equity and efficiency? In other words, must a society that wishes to achieve a more equitable allocation of resources necessarily operate in an economically inefficient manner? The answer, which is given by the *second theorem of welfare economics*, tells us that redistribution need not conflict with economic efficiency. Formally, the second theorem states the following:

> If individual preferences are convex, then every efficient allocation (every point on the contract curve) is a competitive equilibrium for some initial allocation of goods.

Literally, this theorem tells us that any equilibrium deemed to be equitable can be achieved by a suitable distribution of resources among individuals and that such a distribution need not in itself generate inefficiencies. Unfortunately, all programs that redistribute income in our society are economically costly. Taxes may encourage individuals to work less or cause firms to devote resources to avoiding taxes rather than to producing output. So, in effect, there is a trade-off between the goals of equity and efficiency, and hard choices must be made. Welfare economics, which builds on the first and second theorems, provides a useful framework for debating the normative issues that surround the equity–efficiency issue in public policy.

16.4 Efficiency in Production

Having described the conditions required to achieve an efficient allocation in the exchange of two goods, we now consider the efficient use of inputs in the production process. We assume there are fixed total supplies of two inputs, labor and capital, that are needed to produce the same two products, food and clothing. Instead of only two people, however, we now assume that many consumers own the inputs to production (including labor) and earn income by selling them. This income, in turn, is allocated between the two goods.

This framework links the various supply and demand elements of the economy. People supply inputs to production and then use the income they earn to demand and consume goods and services. When the price of one input increases, the individuals who supply a lot of that input earn more income and consume more of one of the two goods. In turn this increases the demand for the inputs needed to produce the good and has a feedback effect on the price of those inputs. Only a general equilibrium analysis can find the prices that equate supply and demand in every market.

Production in the Edgeworth Box

We will continue to use the Edgeworth box diagram, but rather than measure goods on each axis as we did before, we will now measure inputs to the production process. Figure 16.8 shows a box diagram in which labor input is measured along the horizontal axis and capital input on the vertical. Fifty hours of labor and 30 hours of capital are available for the production process. In our earlier analysis of exchange, each origin represented an individual; now each origin represents an output. The food origin is O_F, and O_C is the clothing origin. The only difference between our production analysis and our exchange analysis is that now we measure inputs rather than outputs in the diagram and we focus on two outputs rather than two consumers.

Each point in the diagram represents the labor and capital inputs to the production of food and clothing. For example, *A* represents the input of 35 hours of labor and 5 hours of capital in the production of food and the input of 15 hours of labor and 25 hours of capital in the production of clothing. Every way in which labor and capital can be combined to produce the two goods is represented by a point in the diagram.

In §6.2, we explain that an isoquant is a curve showing all possible combinations of inputs that yield the same output.

A series of production isoquants shows the levels of output produced with various input combinations. Each isoquant represents the total production of a good that can be obtained, without distinguishing the firm or firms that produced it. We have drawn three food isoquants representing 50, 60, and 80 units of food output. The isoquants for food look just like the isoquants we worked

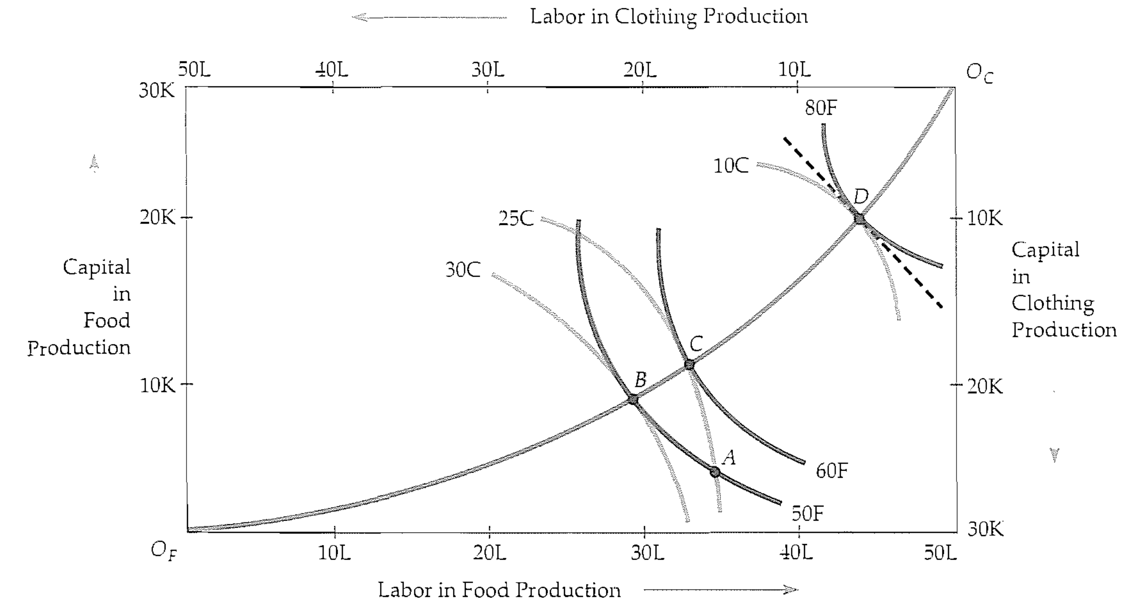

FIGURE 16.8 Efficiency in Production

In an Edgeworth production box with two fixed inputs and two goods, an efficient use of inputs occurs when the isoquants for the two goods are tangent. If production initially uses the inputs described by *A*, the shaded area shows the region in which more of both outputs can be produced by rearranging input use. Points *B*, *C*, and *D* are on the production contract curve and involve efficient input use.

with in Chapter 6, but we have rotated the clothing isoquants by 180 degrees so that they can be read from the point of view of the origin O_C. For example, the isoquant 50F represents all combinations of labor and capital that combine to produce 50 units of food, while 25C represents all combinations of labor and capital that combine to produce 25 units of clothing.

We have also drawn three isoquants representing 10, 25, and 30 units of clothing. These isoquants increase in output as we move from upper right to lower left, again because one or both inputs have increased. Now we can see that *A* simultaneously represents 50 units of food and 25 units of clothing, each associated with a different combination of production inputs.

Input Efficiency

To see how inputs can be combined efficiently, we must find the various combinations of inputs that can be used to produce each of the two outputs. A particular allocation of inputs into the production process is **technically efficient** if the output of one good cannot be increased without decreasing the output of another good. Efficiency in production is not a new concept; in Chapter 6 we saw that a production function represents the maximum output that can be achieved with a given set of inputs. Here we extend the concept to the production of two goods rather than one.

technical efficiency When firms combine inputs to produce a given output as inexpensively as possible

Figure 16.8 shows that inputs are allocated inefficiently if reallocating them generates more of one or both goods. For example, an inefficient allocation, such as that shown at point *A*, might arise if a labor union market has effectively enforced inefficient work rules. The allocation at *A* is clearly inefficient because any input combination in the shaded area generates more of both food and clothing. Thus, we can move from A to B by switching some labor from the production of food to the production of clothing and some capital from the production of clothing to the production of food. This switch generates the same amount of food (50 units), but a larger amount of clothing (from 25 to 30 units).

Production Contract Curve Points B and C in Figure 16.8 are both efficient allocations, as are all points lying on the curve that connects O_F to O_C. Each of these points is a point of tangency of two isoquants, just as every point on the exchange contract curve represents a point of tangency of two indifference curves. The **production contract curve** represents all technically efficient combinations of inputs. Every point that does not lie on this production contract curve is inefficient because the two isoquants that pass through the point intersect. When two isoquants intersect, as at point A, labor and capital can be reallocated to increase the output of at least one of the two goods. From A, we have seen that any allocation within the shaded area increases the production of both goods; therefore, A is technically inefficient.

production contract curve Curve showing all technically efficient combinations of inputs.

Producer Equilibrium in a Competitive Input Market

If input markets are competitive, a point of efficient production will be achieved. Let's see why. If the labor and capital markets are perfectly competitive, then the wage rate w will be the same in all industries. Likewise, the rental price of capital r will be the same whether capital is used in the food or clothing industry. We know from Chapter 7 that if producers of food and clothing minimize production costs, they will use combinations of labor and capital so that the ratio of the marginal products of the two inputs is equal to the ratio of the input prices:

$$MP_L/MP_K = w/r$$

But we also showed that the ratio of the marginal products of the two inputs is equal to the marginal rate of technical substitution of labor for capital $MRTS_{LK}$. As a result,

$$MRTS_{LK} = w/r \quad (16.2)$$

In §6.4, we explain that the marginal rate of technical substitution of labor for capital is the amount by which the input of capital can be reduced when one extra unit of labor is used, so that output remains constant.

Because the MRTS is the slope of the firm's isoquant, a competitive equilibrium can occur in the input market only if each producer must use labor and capital so that the slopes of the isoquants are equal to one another and to the ratio of the prices of the two inputs. As a result, *the competitive equilibrium lies on the production contract curve, and the competitive equilibrium is efficient in production.*

Where we end up on the production contract curve depends on consumers' demands for the two goods. For example, suppose consumers tend to prefer food rather than clothing. One possible competitive equilibrium occurs at D in Figure 16.8. Here, the food producer minimizes the cost of producing 80 units of food by employing 43 units of labor and 20 units of capital. The clothing producer generates 10 units of clothing with 7 units of labor and 10 units of capital. Because the wage rate is equal to the rental price of capital, the isocost lines have a slope of −1 in the diagram. At these prices neither producer will wish to purchase additional production inputs.

Recall from §7.3 that an isocost line includes all possible combinations of labor and capital that can be purchased for a given total cost.

It is easy to check that if we begin at a point off the production contract curve, both producers will find it advantageous to hire labor or rent capital so that they can reallocate their inputs to minimize costs. It is also clear from the diagram in Figure 16.8 that the input market has no unique competitive equilibrium. Efficiency in the use of inputs can involve the production of much food and little clothing, or vice versa.

The Production Possibilities Frontier

The **production possibilities frontier** shows the various combinations of food and clothing that can be produced with fixed inputs of labor and capital, holding technology constant. The frontier in Figure 16.9 is derived from the production contract curve in Figure 16.8. Each point on both the contract curve and the production possibilities frontier describes an efficiently produced level of both food and clothing.

production possibilities frontier Curve showing the combinations of two goods that can be produced with fixed quantities of inputs.

We have labeled the points on the frontier to correspond to the points on the production contract curve. Point O_F represents one extreme, in which only clothing is produced, and O_C represents the other extreme, in which only food is produced. Points B, C, and D correspond to the three other labeled points from the contract curve of Figure 16.8.

Point A, representing an inefficient allocation, lies inside the production possibilities frontier. All points within the triangle ABC involve the complete utilization of labor and capital in the production process. However, a distortion in the labor market, perhaps due to a rent-maximizing union, has caused the economy as a whole to be productively inefficient.

Recall from §14.4 that a rent-maximizing union attempts to maximize the wages that members earn in excess of their opportunity cost.

Why is the production possibilities frontier downward sloping? In order to produce more food efficiently, one must switch inputs from the production of clothing, which in turn lowers the clothing production level. Because all points lying within the frontier are inefficient, they are off the production contract curve.

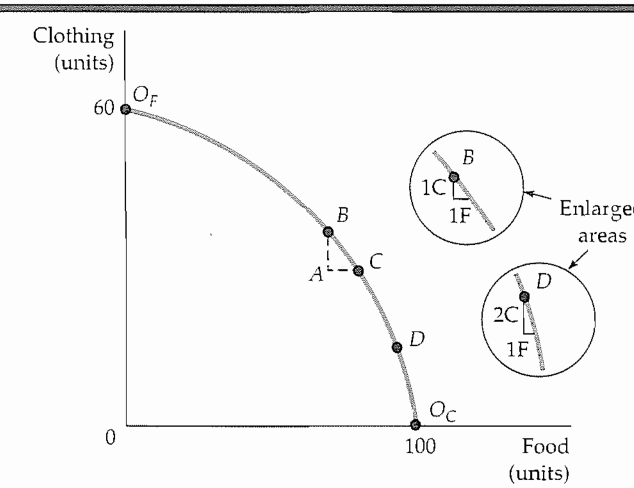

FIGURE 16.9 Production Possibilities Frontier

The production possibilities frontier shows all efficient combinations of outputs. Points B, C, and D are taken from comparable points on the production contract curve in Figure 16.8. The production possibilities frontier is concave because its slope (the marginal rate of transformation) increases as the level of production of food increases.

marginal rate of transformation Amount of one good that must be given up to produce one additional unit of a second good.

Marginal Rate of Transformation The production possibilities frontier is concave (bowed in)—i.e., its slope increases in magnitude as more food is produced. To describe this, we define the **marginal rate of transformation** of food for clothing (MRT) as the magnitude of the slope of the frontier at each point. *The MRT measures how much clothing must be given up to produce one additional unit of food.* For example, the enlarged areas of Figure 16.9 show that at *B* on the frontier, the MRT is 1 because 1 unit of clothing must be given up to obtain 1 additional unit of food. At *D*, however, the MRT is 2 because 2 units of clothing must be given up to obtain 1 more unit of food.

Note that as we increase the production of food by moving along the production possibilities frontier, the MRT increases.[5] This increase occurs because the productivity of labor and capital differs depending on whether the inputs are used to produce more food or clothing. Suppose we begin at O_F, where only clothing is produced. Now we remove some labor and capital from clothing production, where their marginal products are relatively low, and put them into food production, where their marginal products are high. Then, to obtain the first unit of food, very little clothing production is lost (the MRT is much less than 1). But as we move along the frontier and produce less clothing, the productivities of labor and capital in clothing production rise and the productivities of labor and capital in food production fall. At *B*, the productivities are equal and the MRT is 1. Continuing along the frontier, we note that because the input productivities in clothing rise more and the productivities in food decrease, the MRT becomes greater than 1.

We can also describe the shape of the production possibilities frontier in terms of the costs of production. At O_F, where very little clothing output is lost to produce additional food, the marginal cost of producing food is very low (a lot of output is produced with very little input); conversely, the marginal cost of producing clothing is very high (it takes a lot of both inputs to produce another unit of clothing). Thus, when the MRT is low, so is the ratio of the marginal cost of producing food MC_F to the marginal cost of producing clothing MC_C. In fact, *the slope of the production possibilities frontier measures the marginal cost of producing one good relative to the marginal cost of producing the other.* The curvature of the production possibilities frontier follows directly from the fact that the marginal cost of producing food relative to the marginal cost of producing clothing is increasing. At every point along the frontier, the following condition holds:

$$MRT = MC_F/MC_C \quad (16.3)$$

At *B*, for example, the MRT is equal to 1. Here, when inputs are switched from clothing to food production, 1 unit of output is lost and 1 is gained. If the input cost of producing 1 unit of either good is \$100, the ratio of the marginal costs would be \$100/\$100, or 1. Equation (16.3) also holds at *D* (and at every other point on the frontier). Suppose the inputs needed to produce 1 unit of food cost \$160. The marginal cost of food would be \$160, but the marginal cost of clothing would be only \$80 (\$160/2 units of clothing). As a result, the ratio of the marginal costs, 2, is equal to the MRT.

[5] The production possibilities frontier need not have a continually increasing MRT. Suppose, for example, that there were strongly decreasing returns to scale in the production of food. In that case, as inputs were moved from clothing to food production, the amount of clothing that must be given up to obtain one more unit of food would decline.

Output Efficiency

For an economy to be efficient, goods must not only be produced at minimum cost; *goods must also be produced in combinations that match people's willingness to pay for them.* To understand this principle, recall from Chapter 3 that the marginal rate of substitution of clothing for food (MRS) measures the consumer's willingness to pay for an additional unit of food by consuming less clothing. But the marginal rate of transformation measures the cost of an additional unit of food in terms of producing less clothing. An economy produces output efficiently only if, for each consumer,

$$MRS = MRT \quad (16.4)$$

To see why this condition is necessary for efficiency, suppose the MRT equals 1, while the MRS equals 2. In that case, consumers are willing to give up 2 units of clothing to get 1 unit of food, but the cost of getting the additional food is only 1 unit of lost clothing. Clearly, too little food is being produced. To achieve efficiency, food production must be increased, so that the MRS falls and the MRT increases until the two are equal. The outcome is efficient only when MRS = MRT for all pairs of goods.

Figure 16.10 shows this important efficiency condition graphically. Here, we have superimposed one consumer's indifference curve on the production possibilities frontier from Figure 16.9. Note that C is the only point on the production possibilities frontier that maximizes the consumer's satisfaction. Although all points on the production frontier are technically efficient, not all involve the most efficient production of goods from the consumer's perspective. At the point of tangency of the indifference curve and the production frontier, the MRS (the slope of the indifference curve) and the MRT (the slope of the production frontier) are equal.

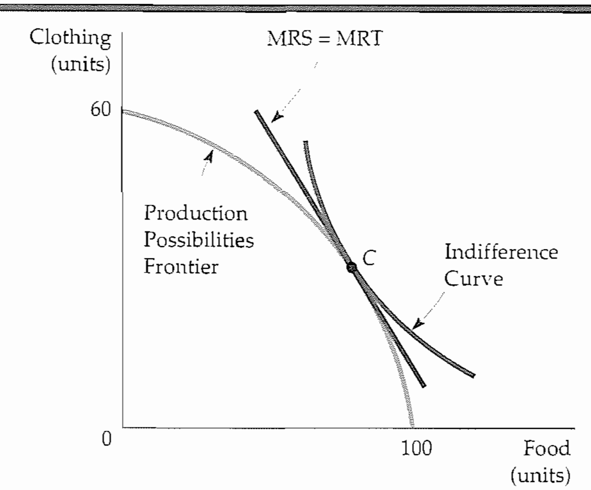

FIGURE 16.10 Output Efficiency

The efficient combination of outputs is produced when the marginal rate of transformation between the two goods (which measures the cost of producing one good relative to the other) is equal to the consumer's marginal rate of substitution (which measures the marginal benefit of consuming one good relative to the other).

If you were a planner in charge of managing an economy, you would face a difficult problem. To achieve efficiency you must equate the marginal rate of transformation with the consumer's marginal rate of substitution. But if different consumers have different preferences for food and clothing, how can you decide what levels of food and clothing to produce and what amount of each to give to every consumer, so that all consumers have the same MRS? The informational and logistical costs are enormous. That is one reason why centrally planned economies, like that of the former Soviet Union, performed so poorly. Fortunately, a well-functioning competitive market system can achieve the same efficient outcome as an ideal managed economy.

Efficiency in Output Markets

When output markets are perfectly competitive, all consumers allocate their budgets so their marginal rates of substitution between two goods are equal to the price ratio. For our two goods, food and clothing,

$$MRS = P_F/P_C$$

At the same time, each profit-maximizing firm will produce its output up to the point at which price is equal to marginal cost. Again, for our two goods,

$$P_F = MC_F \quad \text{and} \quad P_C = MC_C$$

Because the marginal rate of transformation is equal to the ratio of the marginal costs of production, it follows that

$$MRT = MC_F/MC_C = P_F/P_C = MRS \qquad (16.5)$$

In §3.3, we explain that utility maximization is achieved when the marginal benefit of consuming an additional unit of each product is equal to its marginal cost.

When output and input markets are competitive, production will be efficient in that the MRT is equal to the MRS. This condition is just another version of the marginal benefit–marginal cost rule discussed in Chapter 4. There we saw that consumers buy additional units of a good to the point at which the marginal benefit of consumption is equal to the marginal cost. Here we see that the production of food and clothing is chosen so that the marginal benefit of consuming another unit of food is equal to the marginal cost of producing food; the same is true for the consumption and production of clothing.

Figure 16.11 shows that efficient competitive output markets are achieved when production and consumption choices are separated. Suppose the market generates a price ratio of P_F^1/P_C^1. If producers are using inputs efficiently, they will produce food and clothing at A, where the price ratio is equal to the MRT, the slope of the production possibilities frontier. When faced with this budget constraint, however, consumers will consume at B, where they maximize their level of satisfaction (on indifference curve U_2). Because the producer wants to produce F_1 units of food, while consumers want to buy F_2, there will be an excess demand for food. Correspondingly, because consumers wish to buy C_2 units of clothing while producers wish to sell C_1, there will be an excess supply of clothing. Prices in the market will then adjust: The price of food will rise and that of clothing will fall. As price ratio P_F/P_C increases, the price line will move along the production frontier.

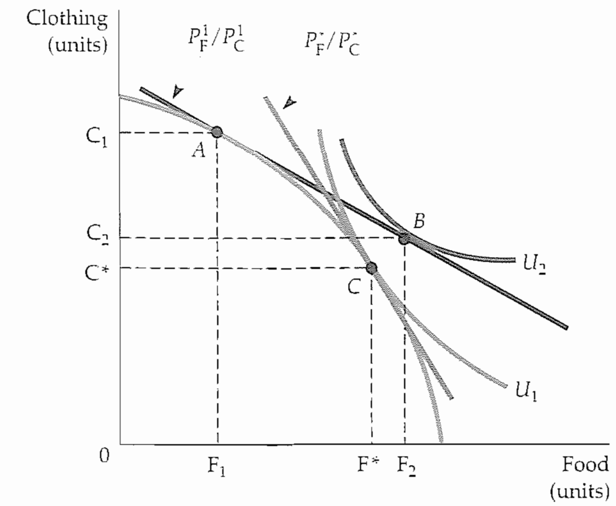

FIGURE 16.11 Competition and Output Efficiency

In a competitive ouput market, people consume to the point where their marginal rate of substitution is equal to the price ratio. Producers choose outputs so that the marginal rate of transformation is equal to the price ratio. Because the MRS equals the MRT, the competitive output market is efficient. Any other price ratio will lead to an excess demand for one good and an excess supply of the other.

An equilibrium results when the price ratio is P_F^*/P_C^* at C. Here, producers want to sell F^* units of food and C^* units of clothing; consumers want to buy the same amounts. At this equilibrium, the MRT and the MRS are equal, so again the competitive equilibrium is efficient.

16.5 The Gains from Free Trade

Clearly there are gains from international trade in an exchange economy. We have seen that two persons or two countries can benefit by trading to reach a point on the contract curve. However, there are additional gains from trade when the economies of two countries differ so that one country has a *comparative advantage* in producing one good, while a second country has a comparative advantage in producing another.

comparative advantage Situation in which country 1 has an advantage over country 2 in producing a good because the cost of producing the good in 1, relative to the cost of producing other goods in 1, is lower than the cost of producing the good in 2, relative to the cost of producing other goods in 2.

Comparative Advantage

Country 1 has a **comparative advantage** *over Country 2 in producing a good if the cost of producing that good, relative to the cost of producing other goods in 1, is lower than the cost of producing the good in 2, relative to the cost of producing other goods in 2.* Note that comparative advantage is not the same as *absolute* advantage. A country has an **absolute advantage** in producing a good if its cost is lower than the cost in another country. A comparative advantage, on the other hand, implies that a country's cost, *relative to the costs of other goods it produces*, is lower than the other country's.

absolute advantage Situation in which country 1 has an advantage over country 2 in producing a good because the cost of producing the good in 1 is lower than the cost of producing it in 2.

TABLE 16.3 Hours of Labor Required to Produce:

	CHEESE (1 LB.)	WINE (1 GAL.)
Holland	1	2
Italy	6	3

When each of two countries has a comparative advantage, they are better off producing what they are best at and purchasing the rest. To see this, suppose that the first country, Holland, has an *absolute* advantage in producing both cheese and wine. A worker there can produce a pound of cheese in 1 hour and a gallon of wine in 2 hours. In Italy, on the other hand, it takes a worker 6 hours to produce a pound of cheese and 3 hours to produce a gallon of wine. The production relationships are summarized in Table 16.3.[6]

Holland has a *comparative* advantage over Italy in producing cheese: Holland's cost of cheese production (in terms of hours of labor used) is half its cost of producing wine, whereas Italy's cost of producing cheese is twice its cost of producing wine. Likewise, Italy has a comparative advantage in producing wine, which it can produce at half the cost at which it can produce cheese.

The comparative advantage of each country determines what happens when they trade. The outcome will depend on the price of each good relative to the other when trade occurs. To see how this might work, suppose that with trade one gallon of wine sells for the same price as one pound of cheese in both Holland and Italy.

Without trade, Holland could, with 24 hours of labor input, produce 24 pounds of cheese, 12 gallons of wine, or a combination of the two, such as 18 pounds of cheese and 3 gallons of wine. But Holland can do better. For every hour of labor Holland can produce 1 pound of cheese, which it can trade for 1 gallon of wine; if the wine were produced at home, 2 hours of labor would be required. It is, therefore, in Holland's interest to specialize in the production of cheese, which it will export to Italy in exchange for wine. If, for example, Holland produced 24 pounds of cheese and traded 6, it would be able to consume 18 pounds of cheese and 6 gallons of wine—a definite improvement over the 18 pounds of cheese and 3 gallons of wine available in the absence of trade.

Italy is also better off with trade. Note that without trade, Italy can, with the same 24 hours of labor input, produce 4 pounds of cheese, 8 gallons of wine, or a combination of the two, such as 3 pounds of cheese and 2 gallons of wine. On the other hand, with every hour of labor Italy can produce one-third of a gallon of wine, which it can trade for one-third of a pound of cheese. If it produced cheese at home, twice as much time would be involved. Specialization in wine production, therefore, is advantageous for Italy. Suppose that Italy produced 8 gallons of wine and traded 6; in that case, it would be able to consume 6 pounds of cheese and 2 gallons of wine, likewise an improvement over the 3 pounds of cheese and 2 gallons of wine available without trade.

[6] This example is based on "World Trade: Jousting for Advantage," *The Economist* (September 22, 1990): 5–40.

An Expanded Production Possibilities Frontier

When there is comparative advantage, international trade has the effect of allowing a country to consume outside its production possibilities frontier. This can be seen graphically in Figure 16.12, which shows a production possibilities frontier for Holland. Suppose initially that Holland has been prevented from trading with Italy because of a protectionist trade barrier. What is the outcome of the competitive process in Holland? Production is at point A, on indifference curve U_1, where the MRT and the pre-trade relative price of wine and cheese is 2. If Holland were able to trade, it would want to export 2 pounds of cheese in exchange for 1 gallon of wine.

Suppose now that the trade barrier is dropped and Holland and Italy are both open to trade. Suppose also that as a result of differences in demand and costs in the two countries, trade occurs on a one-to-one basis. Holland will find it advantageous to produce at point B, the point of tangency of the 1/1 price line and Holland's production possibilities frontier.

That is not the end of the story, however. Point B represents the production decision in Holland (Once the trade barrier has been removed, Holland will produce less wine and more cheese domestically). With trade, however, consumption will occur at point D, at which the higher indifference curve U_2 is tangent to the trade price line. Thus trade has the effect of expanding Holland's consumption choices beyond its production possibilities frontier. Holland will import $W_D - W_B$ units of wine and export $C_B - C_D$ units of cheese.

With trade, each country will undergo a number of important adjustments. As Holland imports wine, the production of domestic wine will fall, as will

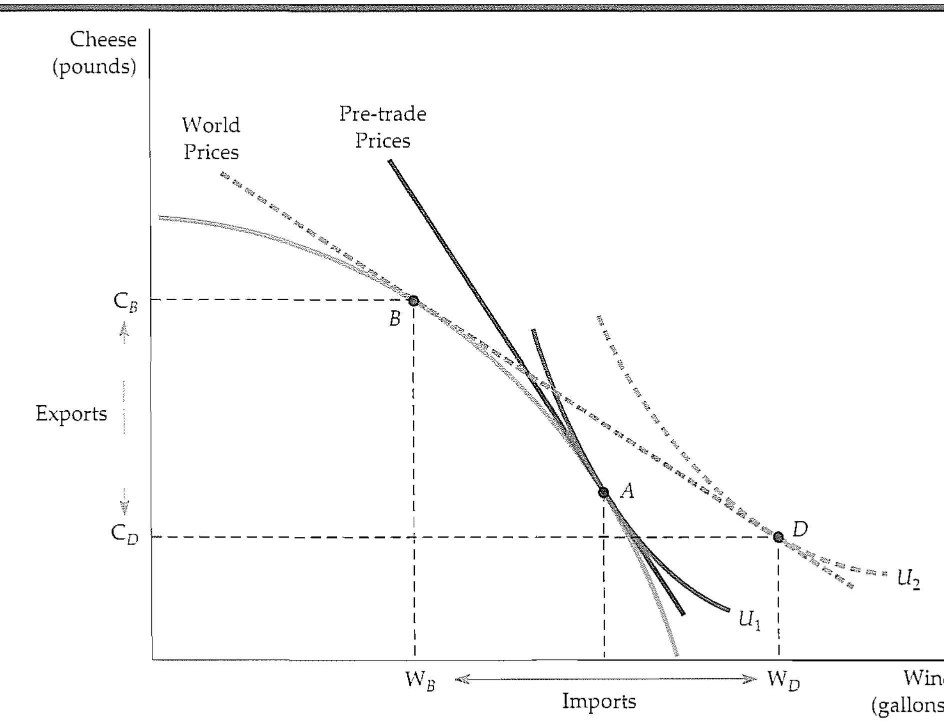

FIGURE 16.12 The Gains from Trade

Without trade, production and consumption are at point A, corresponding to a relative price of cheese to wine of 2 to 1. With trade at a relative price of 1 cheese to 1 wine, domestic production is now at B, while domestic consumption is at D. Free trade has allowed utility to increase from U_1 to U_2.

employment in the wine industry. Cheese production will increase, however, as will the number of jobs in that industry. Workers with job-specific skills may find it difficult to change employment. Thus not everyone will gain as the result of free trade. Although consumers will clearly be better off, producers of wine and workers in the wine industry are likely to be worse off, at least temporarily.

EXAMPLE 16.2 The Effects of Automobile Import Quotas

Governments can use quotas and tariffs to discourage imports and stimulate domestic production. But these devices can restrict or alter consumer choices and thereby generate substantial output inefficiencies. One example is the U.S. imposition of quotas on imports of Japanese automobiles.

During the past three decades, the U.S. automobile industry has faced increasing world competition. In 1965, for example, imports accounted for only 6.1 percent of total domestic sales. This figure increased, however, to 28.8 percent in 1980, when the industry earned a negative profit rate of −9.3 percent on its investment. Part of the industry's difficulty was due to higher-quality, lower-priced Japanese cars. To deal with these competitive problems, the automobile industry convinced the government to negotiate a voluntary export restraint (VER) agreement with the Japanese in 1981. The VER limited Japanese exports to the United States to 1.68 million cars per year, as compared with the 2.5 million cars imported in 1980. U.S. automobile manufacturers argued that the quotas would give time to retool their machines and restructure their union agreements to compete effectively in the world market.

How did these quotas affect the world market? Did they help or hurt American consumers and producers? Answers to these questions require a general equilibrium analysis of the Japanese and U.S. automobile industries, as well as analyses of the markets for labor, materials, and other inputs to the production process.

The evidence suggests that the quotas did little to help the industry retool. U.S. manufacturers had already begun to restructure their production toward smaller and more fuel-efficient cars during the late 1970s. (Real investment expenditures increased by 88 percent from 1975–1976 to 1979–1980, for example.) Although the quotas initially forced the Japanese to sell fewer cars, Japanese prices rose nearly $1000 per car in 1981–1982 and in later years, causing a $2 billion per year increase in revenues. In turn, higher Japanese prices increased the demand for U.S. cars, which allowed the U.S. auto industry to increase its prices, wages, and profits. Over the entire period of their imposition, the quotas increased U.S. auto profits by $10 billion. Finally, U.S. consumers were made worse off by about $3 billion because U.S. automobile prices were approximately $350 to $400 per car higher than they would have been without export restrictions.[7]

The quotas initially benefitted U.S. automobile workers. Without quotas, domestic sales would have been about 500,000 units lower in the early 1980s, which translates into about 26,000 jobs. But higher prices cost consumers well over $4.3 billion dollars. Each job that was retained cost approximately $160,000 ($4.3 billion/26,000). The VER was thus an extremely inefficient way to safeguard domestic employment.

[7] See Steven Berry, James Levindohn, and Ariel Pakes. "Voluntary Export Restraints on Automobiles: Evaluating a Trade Policy." *American Economic Review* (June 1999): 400–30; and Robert W. Crandall, "Import Quotas and the Automobile Industry: The Costs of Protectionism," *The Brookings Review* (Summer 1984): 8–16.

In the early 1990s, the voluntary quota program had little effect on automobile imports. In 1991, for example, Japan exported 1.8 million cars to the United States, even though the voluntary quota was 2.3 million. By March 1992 Japan opted to lower the limit voluntarily to 1.65 million, and by April 1992 the program was eliminated. Yet, despite the reduction in imported cars, Japan's share of the U.S. automobile market increased from 20.5 percent in 1981 to 30.3 percent in 1991, and has remained between 25 percent and 30 percent throughout the 1990s. The explanation for the increase in market share is simple: The production of Japanese automobiles in U.S. plants has increased substantially over the past decade. Today, Japanese automobiles are being produced in many states, including Tennessee and California.

EXAMPLE 16.3 The Costs and Benefits of Special Protection

The demands for protectionist policies increased steadily during the 1980s and into the 1990s. They remain a subject of debate, whether out of concern for trade with various Asian countries or in relation to the North American Free Trade Agreement (NAFTA). Protectionism can take many forms; they include tariffs and quotas of the kind that we analyzed in Chapter 9, regulatory hurdles, subsidies to domestic producers, and controls on the use of foreign exchange. Table 16.4 highlights the findings of one recent study of U.S.-imposed trade restrictions.[8]

TABLE 16.4 Quantifying the Costs of Protection

INDUSTRY	PRODUCER GAINS[a] ($ MILLIONS)	CONSUMER LOSSES[b] ($ MILLIONS)	EFFICIENCY LOSSES[c] ($ MILLIONS)
Book manufacturing	305	500	29
Orange juice	390	525	130
Textiles and apparel	22,000	27,000	4,850
Carbon steel	3,800	6,800	330
Color televisions	190	420	7
Sugar	550	930	130
Dairy products	5,000	5,500	1,370
Meat	1,600	1,800	145

[a] Producer gains in this tariff case are defined as the area of trapezoid *A* in Figure 9.15.
[b] Consumer losses are the sum of areas *A*, *B*, *C*, and *D* in Figure 9.15.
[c] These are given by triangles *B* and *C* in Figure 9.15.

[8] This example is based on Cletus Coughlin, K. Alec Chrystal, and Geoffrey E. Wood, "Protectionist Trade Policies: A Survey of Theory, Evidence and Rationale," *Federal Reserve Bank of St. Louis* (January/February 1988): 12–30. The data in the table are taken from Gary Clyde Hufbauer, Diane T. Berliner, and Kimberly Ann Elliott, "Trade Protection in the United States: 31 Case Studies," *Institute for International Economics* (1986).

Because one of the major purposes of protectionism is to protect jobs in particular industries, it is not surprising that these policies create gains to producers. The costs, however, involve losses to consumers and a substantial reduction in economic efficiency. These efficiency losses are the sum of the loss of producer surplus resulting from inefficient excess domestic production and the loss of consumer surplus resulting from higher domestic prices and lower consumption.

In §9.1, we explain that consumer surplus is the total benefit or value that consumers receive beyond what they pay for a good; producer surplus is the analogous measure for producers.

As the table shows, the textiles and apparel industry is the largest source of efficiency losses. Although there were substantial gains to producers, consumer losses are larger in each case. In addition, efficiency losses from excess (inefficient) domestic production of textiles and reduced domestic consumption of imported textile products were also large—an estimated $4.85 billion. The second largest source of inefficiency was the dairy industry, where losses amounted to $1.37 billion.

Finally, note that the efficiency cost of helping domestic producers varies considerably across industries. In textiles the ratio of efficiency costs to producer gains is 22 percent and in dairy products 27 percent; only orange juice is higher (33.3 percent). However, much lower ratios apply to color televisions (3.7 percent), carbon steel (8.7 percent), and book manufacturing (9.5 percent).

16.6 An Overview—The Efficiency of Competitive Markets

Our analysis of general equilibrium and economic efficiency is now complete. In the process, we have obtained two remarkable results. First, we have shown that for any initial allocation of resources, a competitive process of exchange among individuals, whether through exchange, input markets, or output markets, will lead to an economically efficient outcome. The first theorem of welfare economics tells us that a competitive system, building on the self-interested goals of consumers and producers and on the ability of market prices to convey information to both parties, will achieve an efficient allocation of resources.

Second, we have shown that with consumer preferences that are convex, any efficient allocation of resources can be achieved by a competitive process with a suitable redistribution of those resources. The second theorem of welfare economics tells us that under certain (admittedly ideal) conditions, issues of equity and efficiency can be treated distinctly from one another.

Both theorems of welfare economics depend crucially on the assumption that markets are competitive. Unfortunately, neither of these results necessarily holds when for some reason markets are no longer competitive. In the next two chapters, we will discuss ways in which markets fail and what government can do about it. Before proceeding, however, it is essential to review our understanding of the workings of the competitive process. We thus list the conditions that are required for economic efficiency in exchange, in input markets, and in output markets. These conditions are important; in each of these three cases, you should review the explanation of the conditions in this chapter and the underlying building blocks in prior chapters.

Recall from §3.3 that consumer satisfaction is maximized when the marginal rate of substitution of food for clothing is equal to the ratio of the price of food to that of clothing.

1. *Efficiency in exchange:* All allocations must lie on the exchange contract curve, so that every consumer's marginal rate of substitution of food for clothing is the same:

$$MRS^J_{FC} = MRS^K_{FC}$$

A competitive market achieves this efficient outcome because for consumers the tangency of the budget line and the highest attainable indifference curve assure that

$$MRS^J_{FC} = P_F/P_C = MRS^K_{FC}$$

2. *Efficiency in the use of inputs in production:* All input combinations must lie on the production contract curve, so that every producer's marginal rate of technical substitution of labor for capital is equal in the production of both goods:

$$MRTS^F_{LK} = MRTS^C_{LK}$$

A competitive market achieves this efficient outcome because each producer maximizes profit by choosing labor and capital inputs so that the ratio of the input prices is equal to the marginal rate of technical substitution:

$$MRTS^F_{LK} = w/r = MRTS^C_{LK}$$

Recall from §7.3 that profit maximization requires that the marginal rate of technical substitution of labor for capital be equal to the ratio of the wage rate to the cost of capital.

3. *Efficiency in the output market:* The mix of outputs must be chosen so that the marginal rate of transformation between outputs is equal to consumers' marginal rates of substitution:

$$MRT_{FC} = MRS_{FC} \text{ (for all consumers)}$$

A competitive market achieves this efficient outcome because profit-maximizing producers increase their output to the point at which marginal cost equals price:

In §8.3, we explain that because a competitive firm faces a horizontal demand curve, choosing its output so that marginal cost is equal to price is profit-maximizing.

$$P_F = MC_F, \; P_C = MC_C$$

As a result,

$$MRT_{FC} = MC_F/MC_C = P_F/P_C$$

But consumers maximize their satisfaction in competitive markets only if

$$P_F/P_C = MRS_{FC} \text{ (for all consumers)}$$

Therefore,

$$MRS_{FC} = MRT_{FC}$$

and the output efficiency conditions are satisfied. Thus efficiency requires that goods are produced in combinations and at costs that match people's willingness to pay for them.

16.7 Why Markets Fail

We can give two different interpretations of the conditions required for efficiency. The first stresses that competitive markets work. It also tells us that we ought to ensure that the prerequisites for competition hold, so that resources can be efficiently allocated. The second stresses that the prerequisites for competition are unlikely to hold. It tells us that we ought to concentrate on ways of dealing with market failures. Thus far we have focused on the first interpretation. For the remainder of the book, we concentrate on the second.

Competitive markets fail for four basic reasons: *market power, incomplete information, externalities,* and *public goods.* We will discuss each in turn.

Market Power

We have seen that inefficiency arises when a producer or supplier of a factor input has market power. Suppose, for example, that the producer of food in our Edgeworth box diagram has monopoly power. It therefore chooses the output quantity at which marginal revenue (rather than price) is equal to marginal cost and sells less output at a price higher than in a competitive market. The lower output will mean a lower marginal cost of food production. Meanwhile, the freed-up production inputs will be allocated to produce clothing, whose marginal cost will increase. As a result, the marginal rate of transformation will decrease because $MRT_{FC} = MC_F/MC_C$. We might end up, for example, at A on the production possibilities frontier in Figure 16.9. Producing too little food and too much clothing is an output inefficiency because firms with market power use a different price in their output decisions than consumers use in their consumption decisions.

In §10.2, we explain that a seller of a product has monopoly power if it can profitably charge a price greater than marginal cost; similarly, §10.5 explains that a buyer has monopsony power when its purchasing decision can affect the price of a good.

A similar argument would apply to market power in an input market. Suppose that unions gave workers market power over the supply of their labor in the production of food. Too little labor would then be supplied to the food industry at too high a wage (w_F) and too much labor to the clothing industry at too low a wage (w_C). In the clothing industry, the input efficiency conditions would be satisfied because $MRTS^C_{LK} = w_C/r$. But in the food industry, the wage paid would be greater than the wage in the clothing industry. Therefore, $MRTS^F_{LK} = w_F/r > w_C/r = MRTS^C_{LK}$. The result is input inefficiency because efficiency requires that the marginal rates of technical substitution be equal in the production of all goods.

Incomplete Information

If consumers do not have accurate information about market prices or product quality, the market system will not operate efficiently. This lack of information may give producers an incentive to supply too much of some products and too little of others. In other cases, while some consumers may not buy a product even though they would benefit from doing so, others buy products that leave them worse off. For example, consumers may buy pills that guarantee weight loss, only to find that they have no medical value. Finally, a lack of information may prevent some markets from ever developing. It may, for example, be impossible to purchase certain kinds of insurance because suppliers of insurance lack adequate information about consumers likely to be at risk.

Each of these informational problems can lead to competitive market inefficiency. We will describe informational inefficiencies in detail in Chapter 17, and see whether government intervention might cure them.

Externalities

The price system works efficiently because market prices convey information to both producers and consumers. Sometimes, however, market prices do not reflect the activities of either producers or consumers. There is an *externality* when a consumption or production activity has an indirect effect on other consumption or production activities that is not reflected directly in market prices. As we explained in Section 9.2, the word "externality" is used because the effects on others (whether benefits or costs) are external to the market.

Suppose, for example, that a steel plant dumps effluent in a river, thus making a recreation site downstream unsuitable for swimming or fishing. There is an externality because the steel producer does not bear the true cost of wastewater and so uses too much wastewater to produce its steel. This causes an input inefficiency. If this externality prevails throughout the industry, the price of steel (which is equal to the marginal cost of production) will be lower than if the cost of production reflected the effluent cost. As a result, too much steel will be produced, and there will be an output inefficiency.

We will discuss externalities and ways to deal with them in Chapter 18.

Public Goods

The last source of market failure arises when the market fails to supply goods that many consumers value. A **public good** can be made available cheaply to many consumers, but once it is provided to some consumers, it is very difficult to prevent others from consuming it. For example, suppose a firm is considering whether to undertake research on a new technology for which it cannot obtain a patent. Once the invention is made public, others can duplicate it. As long as it is difficult to exclude other firms from selling the product, the research will be unprofitable.

public good Nonexclusive, nonrival good that can be made available cheaply but which, once available, is difficult to prevent others from consuming.

Markets therefore undersupply public goods. We will see in Chapter 18 that government can sometimes resolve this problem either by supplying a good itself or by altering the incentives for private firms to produce it.

SUMMARY

1. Partial equilibrium analyses of markets assume that related markets are unaffected. General equilibrium analyses examine all markets simultaneously, taking into account feedback effects of other markets on the market being studied.
2. An allocation is efficient when no consumer can be made better off by trade without making someone else worse off. When consumers make all mutually advantageous trades, the outcome is Pareto efficient and lies on the contract curve.
3. A competitive equilibrium describes a set of prices and quantities: When each consumer chooses her most preferred allocation, the quantity demanded is equal to the quantity supplied in every market. All competitive equilibrium allocations lie on the exchange contract curve and are Pareto efficient.
4. The utility possibilities frontier measures all efficient allocations in terms of the levels of utility that each person achieves. Although both individuals prefer some allocations to an inefficient allocation, not *every* efficient allocation must be so preferred. Thus an inefficient allocation can be more equitable than an efficient one.
5. Because a competitive equilibrium need not be equitable, the government may wish to help redistribute wealth from rich to poor. Because such redistribution is costly, there is some conflict between equity and efficiency.
6. An allocation of production inputs is technically efficient if the output of one good cannot be increased without decreasing the output of another. All points of technical efficiency lie on the production contract curve and represent points of tangency of the isoquants for the two goods.
7. A competitive equilibrium in input markets occurs when the marginal rate of technical substitution between pairs of inputs is equal to the ratio of the prices of the inputs.
8. The production possibilities frontier measures all efficient allocations in terms of the levels of output that can be produced with a given combination of inputs. The marginal rate of transformation of food for clothing increases as more food and less clothing are produced. The marginal rate of transformation is equal to the ratio of the marginal cost of producing food to the marginal cost of producing clothing.
9. Efficiency in the allocation of goods to consumers is achieved only when the marginal rate of substitution of one good for another in consumption (which is the same for all consumers) is equal to the marginal rate of transformation of one good for another in production.
10. When input and output markets are perfectly competitive, the marginal rate of substitution (which equals the ratio of the prices of the goods) will equal the marginal rate of transformation (which equals the ratio of the marginal costs of producing the goods).
11. Free international trade expands a country's production possibilities frontier. As a result, consumers are better off.
12. Competitive markets may be inefficient for four reasons. First, firms or consumers may have market power in input or output markets. Second, consumers or producers may have incomplete information and may therefore err in their consumption and production decisions. Third, externalities may be present. Fourth, some socially desirable public goods may not be produced.

QUESTIONS FOR REVIEW

1. Why can feedback effects make a general equilibrium analysis substantially different from a partial equilibrium analysis?
2. In the Edgeworth box diagram, explain how one point can simultaneously represent the market baskets owned by two consumers.
3. In the analysis of exchange using the Edgeworth box diagram, explain why both consumers' marginal rates of substitution are equal at every point on the contract curve.
4. "Because all points on a contract curve are efficient, they are all equally desirable from a social point of view." Do you agree with this statement? Explain.
5. How does the utility possibilities frontier relate to the contract curve?
6. In the Edgeworth production box diagram, what conditions must hold for an allocation to be on the production contract curve? Why is a competitive equilibrium on the contract curve?
7. How is the production possibilities frontier related to the production contract curve?
8. What is the marginal rate of transformation (MRT)? Explain why the MRT of one good for another is equal to the ratio of the marginal costs of producing the two goods.
9. Explain why goods will not be distributed efficiently among consumers if the MRT is not equal to the consumers' marginal rate of substitution.
10. Why can free trade between two countries make consumers of both countries better off?
11. What are the four major sources of market failure? Explain briefly why each prevents the competitive market from operating efficiently.

EXERCISES

1. In the analysis of an exchange between two people, suppose both have identical preferences. Will the contract curve be a straight line? Explain. (Can you think of a counterexample?)
2. Give an example of conditions when the production possibilities frontier might not be concave.
3. A monopsonist buys labor for less than the competitive wage. What type of inefficiency will this use of monopsony power cause? How would your answer change if the monopsonist in the labor market were also a monopolist in the output market?
4. Jane has 8 liters of soft drinks and 2 sandwiches. Bob, on the other hand, has 2 liters of soft drinks and 4 sandwiches. With these endowments, Jane's marginal rate of substitution (MRS) of soft drinks for sandwiches is three; Bob's MRS is equal to one. Draw an Edgeworth box diagram to show whether this allocation of resources is efficient. If it is, explain why. If it is not, explain what exchanges will make both parties better off.
5. The Acme Corporation produces x and y units of goods Alpha and Beta, respectively.
   a. Use a production possibility frontier to explain how the willingness to produce more or less Alpha depends on the marginal rate of transformation of Alpha for Beta.
   b. Consider two cases of production extremes: Acme produces (i) zero units of Alpha initially or (ii) zero units of Beta initially. If Acme always tries to stay on its production possibility frontier, describe the initial positions of cases (i) and (ii). What happens as the Acme Corporation begins to produce *both* goods?
6. In the context of our analysis of the Edgeworth production box, suppose a new invention causes a constant-returns-to-scale production process for food to become a sharply increasing-returns process. How does this change affect the production contract curve?
7. Suppose gold (G) and silver (S) are substitutes for each other because both serve as hedges against inflation. Suppose also that the supplies of both are fixed in the short run ($Q_G = 50$, and $Q_S = 200$) and that the demands for gold and silver are given by the following equations:

$$P_G = 850 - Q_G + 0.5P_S$$

$$P_S = 540 - Q_S + 0.2P_G$$

   a. What are the equilibrium prices of gold and silver?
   b. Suppose a new discovery of gold increases the quantity supplied by 85 units. How will this discovery affect the prices of both gold and silver?

CHAPTER 17

Markets with Asymmetric Information

Chapter Outline

List of Examples

For most of this book, we have assumed that consumers and producers have complete information about the economic variables that are relevant for the choices they face. Now we will see what happens when some parties know more than others—i.e., when there is **asymmetric information.**

Asymmetric information is characteristic of many business situations. Frequently, a seller of a product knows more about its quality than the buyer does. Workers usually know their own skills and abilities better than employers. And business managers know more about their firm's costs, competitive position, and investment opportunities than do the firm's owners.

Asymmetric information explains many institutional arrangements in our society. It is one reason why automobile companies offer warranties on parts and service for new cars; why firms and employees sign contracts that include incentives and rewards; and why the shareholders of corporations must monitor the behavior of managers.

We begin by examining a situation in which the sellers of a product have better information about its quality than buyers have. We will see how this kind of asymmetric information can lead to market failure. In the second section, we see how sellers can avoid some of the problems associated with asymmetric information by giving potential buyers signals about the quality of their product. Product warranties provide a type of insurance that can be helpful when buyers have less information than sellers. But as the third section shows, the purchase of insurance entails difficulties of its own when buyers have better information than sellers.

In the fourth section, we show that managers may pursue goals other than profit maximization when it is costly for owners of private corporations to monitor their behavior. In other words, managers have better information than owners. We also show how firms can give managers an incentive to maximize profits even when monitoring their behavior is costly. Finally, we show that labor markets may operate inefficiently when employees have better information about their productivity than employers have.

17.1 Quality Uncertainty and the Market for Lemons

asymmetric information Situation in which a buyer and a seller possess different information about a transaction.

Suppose you bought a new car for $20,000, drove it 100 miles, and then decided you really didn't want it. There was nothing wrong with the car—it performed beautifully and met all your expectations. You simply felt that you could do just as well without it and would be better off saving the money for other things. So you decide to sell the car. How much should you expect to get for it? Probably not more than $16,000—even though the car is brand new, has been driven only 100 miles, and has a warranty that is transferable to a new owner. And if you were a prospective buyer, you probably wouldn't pay much more than $16,000 yourself.

Why does the mere fact that the car is second-hand reduce its value so much? To answer this question, think about your own concerns as a prospective buyer. Why, you would wonder, is this car for sale? Did the owner really change his or her mind about the car just like that, or is there something wrong with it? Is this car a "lemon"?

Used cars sell for much less than new cars because *there is asymmetric information about their quality:* The seller of a used car knows much more about the car than the prospective buyer does. The buyer can hire a mechanic to check the car, but the seller has had experience with it and will know more about it. Furthermore, the very fact that the car is for sale indicates that it may be a "lemon"—why sell a reliable car? As a result, the prospective buyer of a used car will always be suspicious of its quality—and with good reason.

The implications of asymmetric information about product quality were first analyzed by George Akerlof.[1] Akerlof's analysis goes far beyond the market for used cars. The markets for insurance, financial credit, and even employment are also characterized by asymmetric quality information. To understand the implications of asymmetric information, we will start with the market for used cars and then see how the same principles apply to other markets.

The Market for Used Cars

Suppose two kinds of used cars are available—high-quality cars and low-quality cars. Also, *suppose that both sellers and buyers can tell which kind of car is which.* There will then be two markets, as illustrated in Figure 17.1. In part (a), S_H is the supply curve for high-quality cars, and D_H is the demand curve. Similarly, S_L and D_L in part (b) are the supply and demand curves for low-quality cars. For any given price, S_H lies to the left of S_L because owners of high-quality cars are more relucant to part with them and must receive a higher price to do so. Similarly, D_H is higher than D_L because buyers are willing to pay more to get a high-quality car. As the figure shows, the market price for high-quality cars is $10,000, for low-quality cars $5000, and 50,000 cars of each type are sold.

In reality, the seller of a used car knows much more about its quality than a buyer does. (Buyers discover the quality only after they buy a car and drive it for a while.) Consider what happens, then, if sellers know the quality of cars, but buyers do not. Initially, buyers might think that the odds are 50-50 that a car they buy will be high quality. Why? Because when both sellers *and* buyers knew the

[1] George A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics* (August 1970): 488–500.

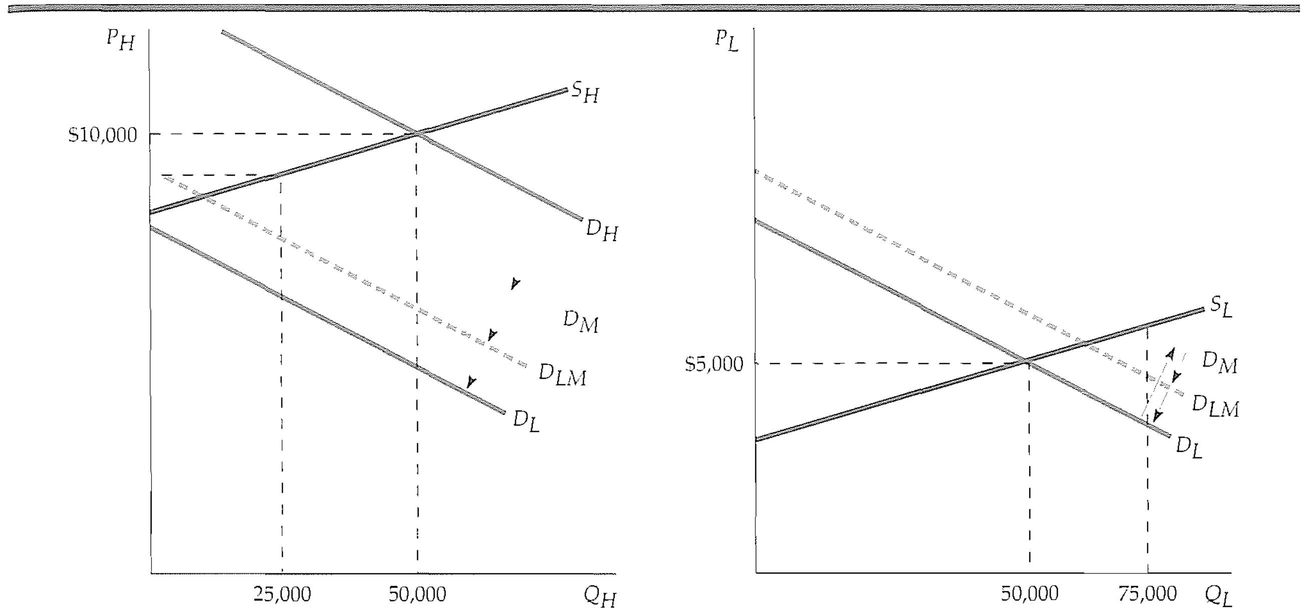

FIGURE 17.1 The Market for Used Cars

When sellers of products have better information about product quality than buyers, a "lemons problem" may arise in which low-quality goods drive out high-quality goods. In (a) the demand curve for high-quality cars is D_H. However, as buyers lower their expectations about the average quality of cars on the market, their perceived demand shifts to D_M. Likewise, in (b) the perceived demand curve for low-quality cars shifts from D_L to D_M. As a result, the quantity of high-quality cars sold falls from 50,000 to 25,000, and the quantity of low-quality cars sold increases from 50,000 to 75,000. Eventually, only low-quality cars are sold.

quality, 50,000 cars of each type were sold. When making a purchase, buyers would therefore view all cars as "medium" quality. (Of course, after buying the car, they will learn its true quality.) The demand for cars perceived to be medium-quality, denoted by D_M in Figure 17.1, is below D_H but above D_L. As the figure shows, *fewer high-quality cars (25,000) and more low-quality cars (75,000) will now be sold.*

As consumers begin to realize that most cars sold (about three-fourths of the total) are low quality, their perceived demand shifts. As Figure 17.1 shows, the new perceived demand curve might be D_{LM}, which means that, on average, cars are thought to be of low to medium quality. However, the mix of cars then shifts even more heavily to low quality. As a result, the perceived demand curve shifts further to the left, pushing the mix of cars even further toward low quality. *This shifting continues until only low-quality cars are sold.* At that point, the market price would be too low to bring forth any high-quality cars for sale, so consumers correctly assume that any car they buy will be low quality, and the only relevant demand curve will be D_L.

The situation in Figure 17.1 is extreme. The market may come into equilibrium at a price that brings forth at least some high-quality cars. *But the fraction of high-quality cars will be smaller than it would be if consumers could identify quality before making the purchase.* That is why I should expect to sell my brand new car, which *I know* is in perfect condition, for much less than I paid for it. Because of asymmetric information, low-quality goods drive high-quality goods out of the market. This phenomenon is sometimes referred to as the *lemons problem.*

Implications of Asymmetric Information

Our used cars example shows how asymmetric information can result in market failure. In an ideal world of fully functioning markets, consumers would be able to choose between low-quality and high-quality cars. While some will choose low-quality cars because they cost less, others will prefer to pay more for high-quality cars. Unfortunately, consumers cannot in fact easily determine the quality of a used car until after they purchase it. As a result, the price of used cars falls, and high-quality cars are driven out of the market.

Market failure arises, therefore, because there are owners of high-quality cars who value their cars less than potential buyers of high-quality cars. As a result, both parties can enjoy gains from trade. Unfortunately, buyers' lack of information prevents this mutually beneficial trade from occurring.

Adverse Selection Used cars are just a stylized example to illustrate an important problem that affects many markets—the problem of adverse selection. **Adverse selection** arises when products of different qualities are sold at a single price because buyers or sellers are not sufficiently informed to determine the true quality at the time of purchase. As a result, too much of the low-quality product and too little of the high-quality product are sold in the marketplace. Let's look at some other examples of asymmetric information and adverse selection. In doing so, we will also see how the government or private firms might respond to the problem.

adverse selection Form of market failure resulting from asymmetric information: if insurance companies must charge a single premium because they cannot distinguish between high-risk and low-risk individuals, more high-risk individuals will insure, making it unprofitable to sell insurance.

The Market for Insurance Why do people over age 65 have difficulty buying medical insurance at almost any price? Older people do have a much higher risk of serious illness, but why doesn't the price of insurance rise to reflect that higher risk? Again, the reason is asymmetric information. People who buy insurance know much more about their general health than any insurance company can hope to know, even if it insists on a medical examination. As a result, adverse selection arises, much as it does in the market for used cars. Because unhealthy people are more likely to want insurance, the proportion of unhealthy people in the pool of insured people increases. This forces the price of insurance to rise, so that more healthy people, aware of their low risks, elect not to be insured. This further increases the proportion of unhealthy people among the insured, which forces the price of insurance up more. This process continues until nearly all people who want to buy insurance are unhealthy. At that point, selling insurance becomes unprofitable.

Adverse selection can make the operation of insurance markets problematic in other ways. Suppose an insurance company wants to offer a policy for a particular event, such as an auto accident that results in property damage. It selects a target population—say, men under age 25—to whom it plans to market this policy, and it estimates the frequency of accidents within this group. For some of these people, the probability of being in an accident is low, much less than .01; for others it is high, much more than .01. If the insurance company cannot distinguish between high- and low-risk men, it will base the premium for all men on the average experience—i.e., an accident probability of .01. With better information, some people (those with low probabilities of an accident) will choose not to insure, while others (those with high probabilities of an accident) will purchase the insurance. This in turn raises the accident probability of those who choose to be insured above .01, forcing the insurance company to raise its premium. In the extreme, only those who are likely to be in an accident will choose to insure, making it impractical to sell insurance.

These kinds of market failure create a role for government. For health insurance, it provides an argument in favor of Medicare or related forms of government health insurance for the elderly. By providing insurance for *all* people over age 65, the government eliminates the problem of adverse selection.[2]

The Market for Credit By using a credit card, many of us borrow money without providing any collateral. Most credit cards allow the holder to run a debit of several thousand dollars, and many people hold several credit cards. Credit card companies earn money by charging interest on the debit balance. But how can a credit card company or bank distinguish high-quality borrowers (who pay their debts) from low-quality borrowers (who don't)? Clearly, borrowers have better information—i.e., they know more about whether they will pay than the lender does. Again, the lemons problem arises. Credit card companies and banks must charge the same interest rate to *all* borrowers. This attracts more low-quality borrowers, which forces the interest rate up, which increases the number of low-quality borrowers, which forces the interest rate up further, and so on.

In fact, credit card companies and banks *can*, to some extent, use computerized credit histories, which they often share with one another, to distinguish "low-quality" from "high-quality" borrowers. Many people think that computerized credit histories invade their privacy. Should companies be allowed to keep these credit histories and share them with other lenders? We can't answer this question for you, but we can point out that credit histories perform an important function: They eliminate, or at least greatly reduce, the problem of asymmetric information and adverse selection, which might otherwise prevent credit markets from operating. Without these histories, even the creditworthy would find it extremely costly to borrow money.

The Importance of Reputation and Standardization

Asymmetric information is also present in many other markets. Here are just a few examples:

- *Retail stores:* Will the store repair or allow you to return a defective product? The store knows more about its policy than you do.
- *Dealers of rare stamps, coins, books, and paintings:* Are the items real or counterfeit? The dealer knows much more about their authenticity than you do.
- *Roofers, plumbers, and electricians:* When a roofer repairs or renovates the roof of your house, do you climb up to check the quality of the work?
- *Restaurants:* How often do you go into the kitchen to check if the chef is using fresh ingredients and obeying the health laws?

In all these cases, the seller knows much more about the quality of the product than the buyer does. Unless sellers can provide information about quality to buyers, low-quality goods and services will drive out high-quality ones, and there will be market failure. Sellers of high-quality goods and services, therefore, have a big incentive to convince consumers that their quality is indeed high. In the examples cited above, this task is performed largely by *reputation*. You shop at a particular store because it has a reputation for servicing its products; you hire a particular roofer and plumber because they have reputations for doing

[2] The same general argument applies to all age groups. That is one reason that insurance companies avoid adverse selection by offering group health insurance policies at places of employment

good work; you go to a particular restaurant because it has a reputation for using fresh ingredients and nobody you know has become sick after eating there.

Sometimes, however, it is impossible for a business to develop a reputation. For example, because most of the customers of highway diners or motels go there only once or infrequently, the businesses have no opportunity to develop reputations. How, then, can they deal with the "lemons" problem? One way is *standardization*. In your hometown, you may not prefer to eat regularly at McDonald's. But a McDonald's may look more attractive when you are driving along a highway and want to stop for lunch. The reason is that McDonald's provides a standardized product: The same ingredients are used and the same food is served in every McDonald's anywhere in the country. Who knows? Joe's Diner might serve better food, but at least you *know* exactly what you will be buying at McDonald's.

EXAMPLE 17.1 Lemons in Major League Baseball

How can we test for the presence of a lemons market? One way is to compare the performance of products that are resold with similar products that are seldom put up for resale. In a lemons market, because purchasers of second-hand products will have limited information, resold products should be lower in quality than products that rarely appear on the market. One such "second-hand" market was created some time ago by a change in the rules governing contracts in major league baseball.[3]

Before 1976, major league baseball teams had the exclusive right to renew a player's contract. After a 1976 ruling declared this system illegal, a new contracting arrangement was created. After six years of major league service, players can now sign new contracts with their original teams or become free agents and sign with new teams. The availability of many free agents creates a second-hand market in baseball players.

Asymmetric information is prominent in the free-agent market. One potential purchaser, the player's original team, has better information about the player's abilities than other teams have. If we were looking at used cars, we could test for the existence of asymmetric information by comparing their repair records. In baseball we can compare player disability records. If players are working hard and following rigorous conditioning programs, we would expect a low probability of injury and a high probability that they will be able to perform if injured. In other words, more motivated players will spend less time on the bench owing to disabilities. If a lemons market exists, we would expect free agents to have higher disability rates than players who are renewed. Players may also have preexisting physical conditions which their original teams know about and which make them less desirable candidates for contract renewal. Because more such players would become free agents, free agents would experience higher disability rates for health reasons.

Table 17.1, which lists the post-contract performance of all players who have signed multiyear contracts, makes two points. First, both free agents and renewed players have increased disability rates after signing contracts. The disabled days per season increase from an average of 4.73 to an average of 12.55. Second, the postcontract disability rates of renewed and not-renewed players

[3] This example is based on Kenneth Lehn's study of the free-agent market. See "Information Asymmetries in Baseball's Free-Agent Market," *Economic Inquiry* (1984): 37–44.

TABLE 17.1 Player Disability

	DAYS SPENT ON DISABLED LIST PER SEASON		
	Precontract	Postcontract	Percentage Change
All players	4.73	12.55	165.4
Renewed players	4.76	9.68	103.4
Free agents	4.67	17.23	268.9

are significantly different. On average, renewed players are disabled 9.68 days, free agents 17.23 days.

These two findings suggest a lemons market in free agents that exists because baseball teams know their own players better than the teams with which they compete.

17.2 Market Signaling

We have seen that asymmetric information can sometimes lead to a "lemons problem": Because sellers know more about the quality of a good than buyers do, buyers may assume that quality is low, so that price falls and only low-quality goods are sold. We also saw how government intervention (in the market for health insurance, for example) or the development of a reputation (in service industries, for example) can alleviate this problem. Now we will examine another important mechanism through which sellers and buyers deal with the problem of asymmetric information: **market signaling.** The concept of market signaling was first developed by Michael Spence, who showed that in some markets, sellers send buyers *signals* that convey information about a product's quality.[4]

market signaling Process by which sellers send signals to buyers conveying information about product quality.

To see how market signaling works, let's look at a *labor market*, which is a good example of a market with asymmetric information. Suppose a firm is thinking about hiring some new people. The new workers (the "sellers" of labor) know much more about the quality of the labor they can provide than does the firm (the buyer of labor). For example, they know how hard they tend to work, how responsible they are, what their skills are, and so forth. The firm will learn these things only after workers have been hired and have been working for some time. At the time they are hired, the firm knows little about how productive they will turn out to be.

Why don't firms simply hire workers, see how well they work, and then fire those with low productivity? Because this policy is often very costly. In many countries, and in many firms in the United States, it is difficult to fire someone who has been working more than a few months. (The firm may have to show just cause or pay severance pay.) Moreover, in many jobs, workers do not become fully productive for at least six months. Before that time, considerable on-the-job training may be required, for which the firm must invest substantial resources. Thus the firm might not learn how good workers are for six months to a year. Clearly, firms would be much better off if they knew how productive potential employees were *before* they hired them.

[4] Michael Spence, *Market Signaling* (Cambridge, MA: Harvard University Press, 1974).

What characteristics can a firm examine to obtain information about people's productivity before it hires them? Can potential employees convey information about their productivity? Dressing well for the job interview might convey some information, but even unproductive people sometimes dress well to get a job. Dressing well is thus a *weak signal*—it doesn't do much to distinguish high-productivity from low-productivity people. *To be strong, a signal must be easier for high-productivity people to give than for low-productivity people to give, so that high-productivity people are more likely to give it.*

For example, *education* is a strong signal in labor markets. A person's educational level can be measured by several things—the number of years of schooling, degrees obtained, the reputation of the university or college that granted the degrees, the person's grade point average, and so on. Of course, education can directly and indirectly improve a person's productivity by providing information, skills, and general knowledge that are helpful in work. But even if education did *not* improve productivity, it would still be a useful *signal* of productivity because more productive people find it easier to attain high levels of education. Not surprisingly, productive people tend to be more intelligent, more motivated, more disciplined, and more energetic and hard-working—characteristics that are also helpful in school. More productive people are therefore more likely to attain high levels of education *in order to signal their productivity to firms and thereby obtain better-paying jobs.* Thus firms are correct in considering education a signal of productivity.

A Simple Model of Job Market Signaling

To understand how signaling works, we will discuss a simple model.[5] Let's assume there are only low-productivity workers (Group I), whose average and marginal product is 1, and high-productivity workers (Group II), whose average marginal product is 2. Workers will be employed by competitive firms whose products sell for $10,000, and who expect an average of 10 years of work from each employee. We also assume that half the workers in the population are in Group I and the other half in Group II, so that the *average* productivity of all workers is 1.5. Note that the revenue expected to be generated from Group I workers is $100,000 ($10,000/year × 10 years) and from Group II workers is $200,000 ($20,000/year × 10 years).

If firms could identify people by their productivity, they would offer them a wage equal to their marginal revenue product. Group I people would be paid $10,000 per year, Group II people $20,000. On the other hand, if firms could not identify productivity before they hired people, they would pay all workers an annual wage equal to the average productivity, $15,000. Group I people would then earn more ($15,000 instead of $10,000), at the expense of Group II people (who would earn $15,000 instead of $20,000).

Now let's consider what can happen with signaling via education. Suppose all the attributes of an education (degrees earned, grade point average, etc.) can be summarized by a single index y that represents years of higher education. All education involves a cost, and the higher the educational level y, the higher the cost. This cost includes tuition and books, the opportunity cost of foregone wages, and the psychic cost of having to work hard to obtain high grades. What is important is that *the cost of education is greater for the low-productivity group than for the high-productivity group.* We might expect this for two reasons. First, low-productivity workers may simply be less studious. Second, low-productivity workers may

[5] This is essentially the model developed in Spence, *Marketing Signaling.*

progress more slowly through degree programs in which they enroll. In particular, suppose that for Group I people, the cost of attaining educational level y is given by

$$C_I(y) = \$40{,}000y$$

and for Group II people it is

$$C_{II}(y) = \$20{,}000y$$

Now suppose (to keep things simple and to dramatize the importance of signaling) that *education does nothing to increase one's productivity; its only value is as a signal.* Let's see if we can find a market equilibrium in which different people obtain different levels of education, and firms look at education as a signal of productivity.

Consider the following possible equilibrium. Suppose firms use this decision rule: *Anyone with an education level of y^* or more is a Group II person and is offered a wage of $20,000, and anyone with an education level below y^* is a Group I person and is offered a wage of $10,000.* The particular level y^* that the firms choose is arbitrary, but for this decision rule to be part of an equilibrium, firms must have identified people correctly. Otherwise, the firms will want to change the rule. Will it work?

To answer this question, we must determine how much education the people in each group will obtain, *given that firms are using this decision rule.* To do this, remember that education allows one to get a better-paying job. The benefit of education $B(y)$ is the *increase* in the wage associated with each level of education, as shown in Figure 17.2. Observe that $B(y)$ is 0 initially, which represents the $100,000 base 10-year earnings that are earned without any college education.

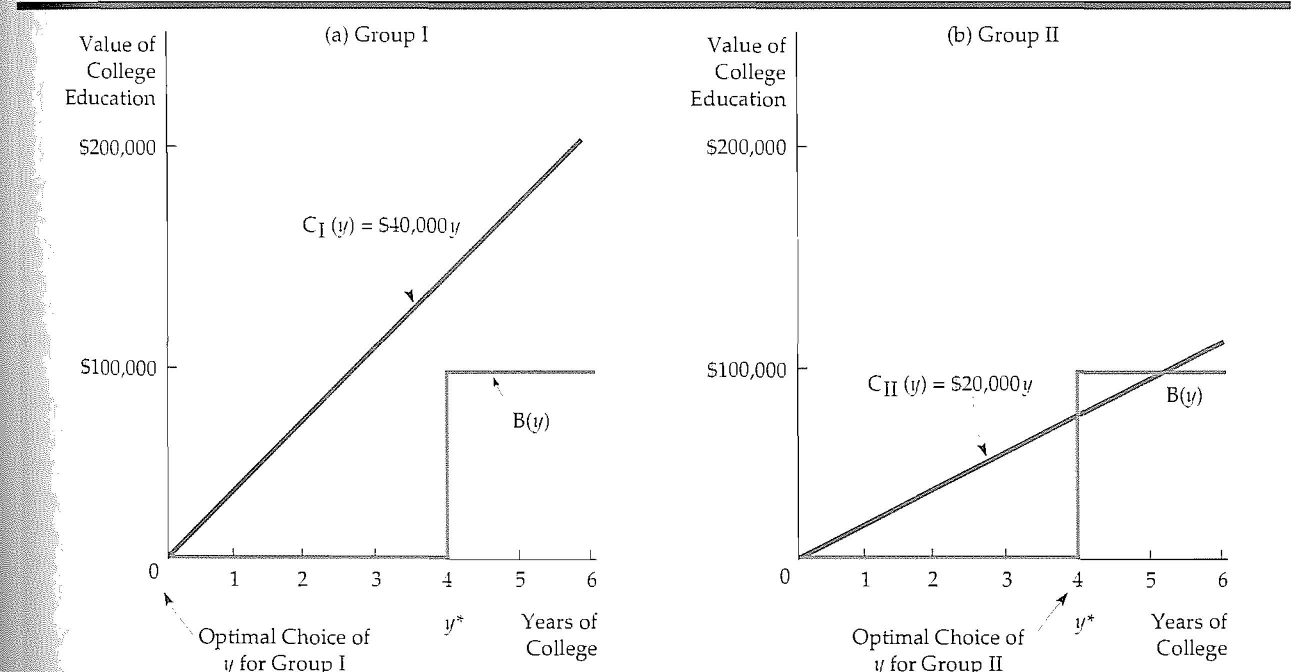

FIGURE 17.2 Signaling

Education can be a useful signal of the high productivity of a group of workers if education is easier to obtain for this group than for a low-productivity group. In (a), the low-productivity group will choose an education level of $y = 0$ because the cost of education is greater than the increased earnings resulting from education. However, in (b), the high-productivity group will choose an education level of $y^* = 4$ because the gain in earnings is greater than the cost.

For an education level less than y^*, $B(y)$ remains 0, because 10-year earnings remain at the \$100,000 base level. But when the education level reaches y^* or greater, 10-year earnings increase to \$200,000, so $B(y)$ becomes \$100,000.

How much education should a person obtain? Clearly the choice is between *no* education (i.e., $y = 0$) and an education level of y^*. The reason is that any level of education less than y^* results in the same base earnings of \$100,000, so there is no benefit from obtaining an education at a level above 0 but below y^*. Similarly, there is no benefit from obtaining an educational level above y^* because y^* is sufficient to allow one to enjoy the higher total earnings of \$200,000.

In deciding how much education to obtain, people compare the benefit of education with the cost. People in each group make the following cost-benefit calculation: *Obtain the education level y^* if the benefit (i.e., the increase in earnings) is at least as large as the cost of this education.* For both groups, the benefit (the increase in earnings) is \$100,000. The costs, however, differ. For Group I, the cost is \$40,000y, but for Group II it is only \$20,000y. Therefore, Group I will obtain *no* education as long as

$$\$100{,}000 < \$40{,}000y^* \text{ or } y^* > 2.5$$

and Group II will obtain an education level y^* as long as

$$\$100{,}000 > \$20{,}000y^* \text{ or } y^* < 5$$

These results give us an equilibrium *as long as y^* is between 2.5 and 5.* Suppose, for example, that y^* is 4.0, as in Figure 17.2. Then people in Group I will find that education does not pay, and they will not obtain any, whereas people in Group II will find that education does pay, and they will obtain the level $y = 4.0$. Now, when a firm interviews job candidates who have no college education, it correctly assumes they have low productivity and offers them a wage of \$10,000. Similarly, when the firm interviews people who have four years of college, it correctly assumes their productivity is high, warranting a wage of \$20,000. We therefore have an equilibrium. High-productivity people will obtain a college education to signal their productivity; firms will read this signal and offer them a high wage.

This is a highly simplified model, but it illustrates a significant point: Education can be an important signal that allows firms to sort workers according to productivity. Some workers (those with high productivity) will want to obtain a college education *even if that education does nothing to increase their productivity.* These workers simply want to identify themselves as highly productive, so they obtain the education needed to send a signal.

In the real world, of course, education *does* provide useful knowledge and does increase one's ultimate productivity. (We wouldn't have written this book if we didn't believe that.) But education also serves a signaling function. For example, many firms insist that a prospective manager have an MBA. One reason is that MBAs learn economics, finance, and other useful subjects. But there is a second reason: To complete an MBA program takes intelligence, discipline, and hard work, and people with those qualities tend to be very productive.

Guarantees and Warranties

We have stressed the role of signaling in labor markets, but it can also play an important role in many other markets in which there is asymmetric information. Consider the markets for such durable goods as televisions, stereos, cameras, and refrigerators. Many firms produce these items, but some brands are more dependable than others. If consumers could not tell which brands tend to be more dependable, the better brands could not be sold for higher prices. Firms that produce a higher-quality, more dependable product must therefore make consumers aware of this difference. But how can they do it in a convincing way? The answer is *guarantees and warranties*.

Guarantees and warranties effectively signal product quality because an extensive warranty is more costly for the producer of a low-quality item than for the producer of a high-quality item. The low-quality item is more likely to require servicing under the warranty, which the producer will have to pay for. As a result, in their own self-interest, producers of low-quality items will not offer extensive warranties. Thus consumers can correctly view an extensive warranty as a signal of high quality, and they will pay more for products that offer one.

EXAMPLE 17.2 Working into the Night

Job market signaling does not end when one is hired. Even after a few years of employment, a worker will still know more about his abilities than will the employer. This is especially true for workers in knowledge-based fields such as engineering, computer programming, finance, law, management, and consulting. Although an unusually talented computer programmer, for example, will be more skilled than his co-workers at writing programs that are efficient and bug-free, it may take several years before the firm fully recognizes this talent. Given this asymmetric information, what policy should employers use to decide promotions and salary increases? Can workers who are unusually talented and productive signal this fact, and thereby receive earlier promotions and larger salary increases?

Workers can often signal talent and productivity *by working harder and longer hours.* Because more talented and productive workers tend to get more enjoyment and satisfaction from their jobs, it is less costly for them to send this signal than it is for other workers. The signal is therefore strong: it conveys information. As a result, employers can—and do—rely on this signal when making promotion and salary decisions.

This signalling process has affected the way many people work. Rather than an hourly wage, knowledge-based workers are typically paid a fixed salary for a 35- or 40-hour week and do not receive overtime pay if they work additional hours. Yet such workers increasingly work well beyond their weekly schedules. Surveys by the U.S. Labor Department, for example, have found that the percentage of all workers who toil 49 hours or more a week has risen from 13 percent in 1976 to 19 percent in 1998.[6] Many young lawyers, accountants, consultants, investment bankers, and computer programmers regularly work into the night and on weekends, putting in 60- or 70-hour weeks. Is it surprising that these people are working so hard? Not at all. They are trying to send signals that can greatly affect their careers.

Employers rely increasingly on the signalling value of long hours as rapid technological change makes it harder for them to find other ways of assessing workers' skills and productivity. A study of software engineers at the Xerox Corporation, for example, found that many people work into the night because they fear that otherwise their bosses will conclude that they are shirkers

[6] "At the Desk, Off the Clock and Below Statistical Radar," *New York Times*, July 18, 1999.

For an education level less than y^*, $B(y)$ remains 0, because 10-year earnings remain at the \$100,000 base level. But when the education level reaches y^* or greater, 10-year earnings increase to \$200,000, so $B(y)$ becomes \$100,000.

How much education should a person obtain? Clearly the choice is between *no* education (i.e., $y = 0$) and an education level of y^*. The reason is that any level of education less than y^* results in the same base earnings of \$100,000, so there is no benefit from obtaining an education at a level above 0 but below y^*. Similarly, there is no benefit from obtaining an educational level above y^* because y^* is sufficient to allow one to enjoy the higher total earnings of \$200,000.

In deciding how much education to obtain, people compare the benefit of education with the cost. People in each group make the following cost-benefit calculation: *Obtain the education level y^* if the benefit (i.e., the increase in earnings) is at least as large as the cost of this education.* For both groups, the benefit (the increase in earnings) is \$100,000. The costs, however, differ. For Group I, the cost is \$40,000y, but for Group II it is only \$20,000y. Therefore, Group I will obtain *no* education as long as

$$\$100{,}000 < \$40{,}000y^* \text{ or } y^* > 2.5$$

and Group II will obtain an education level y^* as long as

$$\$100{,}000 > \$20{,}000y^* \text{ or } y^* < 5$$

These results give us an equilibrium *as long as y^* is between 2.5 and 5.* Suppose, for example, that y^* is 4.0, as in Figure 17.2. Then people in Group I will find that education does not pay, and they will not obtain any, whereas people in Group II will find that education does pay, and they will obtain the level $y = 4.0$. Now, when a firm interviews job candidates who have no college education, it correctly assumes they have low productivity and offers them a wage of \$10,000. Similarly, when the firm interviews people who have four years of college, it correctly assumes their productivity is high, warranting a wage of \$20,000. We therefore have an equilibrium. High-productivity people will obtain a college education to signal their productivity; firms will read this signal and offer them a high wage.

This is a highly simplified model, but it illustrates a significant point: Education can be an important signal that allows firms to sort workers according to productivity. Some workers (those with high productivity) will want to obtain a college education *even if that education does nothing to increase their productivity.* These workers simply want to identify themselves as highly productive, so they obtain the education needed to send a signal.

In the real world, of course, education *does* provide useful knowledge and does increase one's ultimate productivity. (We wouldn't have written this book if we didn't believe that.) But education also serves a signaling function. For example, many firms insist that a prospective manager have an MBA. One reason is that MBAs learn economics, finance, and other useful subjects. But there is a second reason: To complete an MBA program takes intelligence, discipline, and hard work, and people with those qualities tend to be very productive.

Guarantees and Warranties

We have stressed the role of signaling in labor markets, but it can also play an important role in many other markets in which there is asymmetric information. Consider the markets for such durable goods as televisions, stereos, cameras, and refrigerators. Many firms produce these items, but some brands are more dependable than others. If consumers could not tell which brands tend to be more dependable, the better brands could not be sold for higher prices. Firms that produce a higher-quality, more dependable product must therefore make consumers aware of this difference. But how can they do it in a convincing way? The answer is *guarantees and warranties.*

Guarantees and warranties effectively signal product quality because an extensive warranty is more costly for the producer of a low-quality item than for the producer of a high-quality item. The low-quality item is more likely to require servicing under the warranty, which the producer will have to pay for. As a result, in their own self-interest, producers of low-quality items will not offer extensive warranties. Thus consumers can correctly view an extensive warranty as a signal of high quality, and they will pay more for products that offer one.

EXAMPLE 17.2 Working into the Night

Job market signaling does not end when one is hired. Even after a few years of employment, a worker will still know more about his abilities than will the employer. This is especially true for workers in knowledge-based fields such as engineering, computer programming, finance, law, management, and consulting. Although an unusually talented computer programmer, for example, will be more skilled than his co-workers at writing programs that are efficient and bug-free, it may take several years before the firm fully recognizes this talent. Given this asymmetric information, what policy should employers use to decide promotions and salary increases? Can workers who are unusually talented and productive signal this fact, and thereby receive earlier promotions and larger salary increases?

Workers can often signal talent and productivity *by working harder and longer hours.* Because more talented and productive workers tend to get more enjoyment and satisfaction from their jobs, it is less costly for them to send this signal than it is for other workers. The signal is therefore strong: it conveys information. As a result, employers can—and do—rely on this signal when making promotion and salary decisions.

This signalling process has affected the way many people work. Rather than an hourly wage, knowledge-based workers are typically paid a fixed salary for a 35- or 40-hour week and do not receive overtime pay if they work additional hours. Yet such workers increasingly work well beyond their weekly schedules. Surveys by the U.S. Labor Department, for example, have found that the percentage of all workers who toil 49 hours or more a week has risen from 13 percent in 1976 to 19 percent in 1998.[6] Many young lawyers, accountants, consultants, investment bankers, and computer programmers regularly work into the night and on weekends, putting in 60- or 70-hour weeks. Is it surprising that these people are working so hard? Not at all. They are trying to send signals that can greatly affect their careers.

Employers rely increasingly on the signalling value of long hours as rapid technological change makes it harder for them to find other ways of assessing workers' skills and productivity. A study of software engineers at the Xerox Corporation, for example, found that many people work into the night because they fear that otherwise their bosses will conclude that they are shirkers

[6] "At the Desk, Off the Clock and Below Statistical Radar," *New York Times*, July 18, 1999.

who choose the easiest assignments. As the bosses make clear, this fear is warranted: "We don't know how to assess the value of a knowledge worker in these new technologies," says one Xerox manager, "so we value those who work into the night."[7]

As corporations become more reluctant to offer lifetime job security, and as competition for promotion intensifies, salaried workers feel more and more pressure to work long hours. If you find yourself working 60- or 70-hour weeks, look at the bright side—the signal you're sending is a strong one.

17.3 Moral Hazard

When one party is fully insured and cannot be accurately monitored by an insurance company with limited information, the insured party may take an action that increases the likelihood that an accident or an injury will occur. For example, if my home is fully insured against theft, I may be less diligent about locking doors when I leave, and I may choose not to install an alarm system. The possibility that an individual's behavior may change because she has insurance is an example of a problem known as **moral hazard**.

moral hazard When an insured party whose actions are unobserved can affect the probability or magnitude of a payment associated with an event.

The concept of moral hazard applies not only to problems of insurance but also to problems of workers who perform below their capabilities when employers cannot monitor their behavior ("job shirking"). In general, *moral hazard occurs when a party whose actions are unobserved affects the probability or magnitude of a payment*. For example, if I have complete medical insurance coverage, I may visit the doctor more often than I would if my coverage were limited. If the insurance provider can monitor its insurees' behavior, it can charge higher fees for those who make more claims. But if the company cannot monitor behavior, it may find its payments to be larger than expected. Under conditions of moral hazard, insurance companies may be forced to increase premiums for everyone, or even to refuse to sell insurance at all.

Consider, for example, the decisions faced by the owners of a warehouse valued at $100,000 by their insurance company. Suppose that if the owners run a $50 fire-prevention program for their employees, the probability of a fire is .005. Without this program, the probability increases to .01. Knowing this, the insurance company faces a dilemma if it cannot monitor the company's decision to conduct a fire-prevention program. The policy that the insurance company offers cannot include a clause stating that payments will be made only if there is a fire-prevention program. If the program were in place, the company could insure the warehouse for a premium equal to the expected loss from a fire—an expected loss equal to .005 × $100,000 = $500. Once the insurance policy is purchased, however, the owners no longer have an incentive to run the program. If there is a fire, they will be fully compensated for their financial loss. Thus, if the insurance company sells a policy for $500, it will incur losses because the expected loss from the fire will be $1000 (.01 × $100,000).

Moral hazard is not only a problem for insurance companies. It also alters the ability of markets to allocate resources efficiently. In Figure 17.3, for example, D

[7] Ibid.

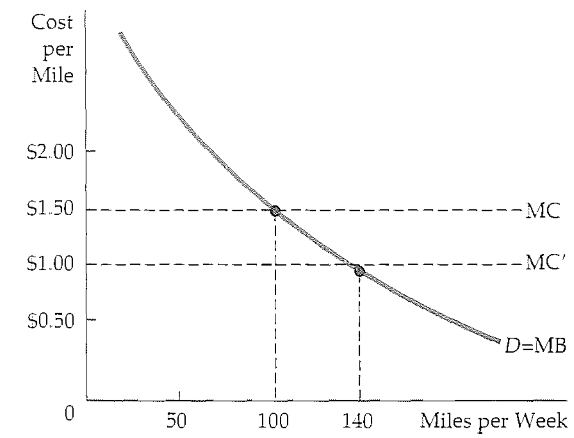

FIGURE 17.3 The Effects of Moral Hazard

Moral hazard alters the ability of markets to allocate resources efficienctly. *D* gives the demand for automobile driving. With no moral hazard, the marginal cost of transportation MC is $1.50 per mile; the driver drives 100 miles, which is the efficient amount. With moral hazard, the driver perceives the cost per mile to be MC' = $1.00 and drives 140 miles.

gives the demand for automobile driving in miles per week. The demand curve, which measures the marginal benefits of driving, is downward sloping because some people switch to alternative transportation as the cost of driving increases. Suppose that initially the cost of driving includes the insurance cost and that insurance companies can accurately measure miles driven. In this case, there is no moral hazard and the marginal cost of driving is given by MC. Drivers know that more driving will increase their insurance premium and so increase their total cost of driving (the cost per mile is assumed to be constant). For example, if the cost of driving is $1.50 per mile (50 cents of which is insurance cost), the driver will go 100 miles per week.

A moral hazard problem arises when insurance companies cannot monitor individual driving habits, so that the insurance premium does not depend on miles driven. In that case, drivers assume that any additional accident costs that they incur will be spread over a large group, with only a negligible portion accruing to each of them individually. Because their insurance premium does not vary with the number of miles that they drive, an additional mile of transportation will cost $1.00, as shown by the marginal cost curve MC', rather than $1.50. The number of miles driven will increase from 100 to the socially inefficient level of 140.

Moral hazard not only alters behavior; it also creates economic inefficiency. The inefficiency arises because the insured individual perceives either the cost or the benefit of the activity differently from the true social cost or benefit. In the driving example of Figure 17.3, the efficient level of driving is given by the intersection of the marginal benefit (MB) and marginal cost (MC) curves. With moral hazard, however, the individual's perceived marginal cost (MC') is less than actual cost, and the number of miles driven per week (140) is higher than the efficient level at which marginal benefit is equal to marginal cost (100).

EXAMPLE 17.3 Reducing Moral Hazard—Warranties of Animal Health

For buyers of livestock, information about the animals' health is very important.[8] Unhealthy animals gain weight more slowly and are less likely to reproduce. Because of asymmetric information in the livestock market (sellers know the health of an animal better than buyers do), most states require warranties on the sale of livestock. Under these laws, sellers not only promise (warrant) that animals are free from hidden diseases but are responsible for all costs arising from any diseased animals.

Although warranties solve the problem of the seller's having better information than the buyer, they also create a form of moral hazard. Guaranteeing reimbursement to the buyer for all costs associated with diseased animals means that insurance rates are not tied to the level of care that buyers or their agents take to protect their livestock against disease. As a result of these warranties, livestock buyers avoid paying for early diagnoses of diseased livestock, and losses increase.

In response to the moral hazard problem, many states have modified their animal warranty laws by requiring sellers to tell buyers whether livestock are diseased at the time of sale. Some states also require sellers to comply with state and federal animal health regulations, thereby reducing disease. Beyond these measures, however, warranties that animals are free from hidden disease must be in the form of explicit written or oral guarantees to buyers.

EXAMPLE 17.4 Crisis in the Savings and Loan Industry

In 1934, during the Great Depression, the U.S. government introduced a broad-based system of financial insurance. The Federal Deposit Insurance Corporation (FDIC) provided insurance for deposits at commercial banks, and the Federal Savings and Loan Insurance Corporation did the same (up to $100,000 per account) for deposits at savings and loans (S&Ls). These insurance programs created the seeds of moral hazard on the part of depositors: A depositor could lend money to any financial institution, no matter how risky that institution's loans, without bearing any risk.

Later, depositor moral hazard was coupled with moral hazard by owners of savings and loans. Beginning in 1982, some S&Ls found that they could attract large sums of government-insured capital and invest the money virtually without restriction in highly speculative investments. Because the deposits were insured, S&L managers had little incentive to evaluate the risks involved.

Essentially, deposit insurance enabled S&Ls to make riskier loans on a larger scale than they would otherwise. The adverse incentives created by moral hazard coupled with the collapse of the real estate boom in the Sun Belt and energy-producing states led to the failure of many savings and loans.

[8] This example is based on Terence J. Centner and Michael E. Wetzstein, "Reducing Moral Hazard Associated with Implied Warranties of Animal Health," *American Journal of Agricultural Economics* 69 (1987): 143–50.

In 1990, the cost of bailing out depositors whose money was lost when over 1000 S&Ls failed was estimated conservatively to be over $200 billion.[9] The biggest losses were in Texas, where over $42 billion had been spent by October 1990. Total outlays by the agencies responsible for deposit insurance were nearly $100 billion just through 1990.

Aware of the adverse incentives that were created by moral hazard, the government has modified its insurance system. Today, the FDIC regulates the S&L and banking industries, and S&Ls now face stiff capital requirements that force managers to bear a stake in the outcome of their investment policies. With a good deal of their own money at risk, managers are less inclined to invest speculatively.

A number of additional reforms could help to remove the moral hazard problem on the part of depositors and S&L owners. Proposals that would affect depositors include (1) lowering the amount of insurance coverage; (2) making the maximum coverage apply to each individual, no matter how many accounts he has; and (3) allowing for coinsurance, whereby deposit insurance reimburses losses on less than a dollar-for-dollar basis. Proposals directed toward owners include (1) charging an S&L insurance premiums that are based on the riskness of its portfolio—the greater the risk, the higher the premium; and (2) restricting the investment opportunities available to S&L owners.

17.4 The Principal–Agent Problem

If monitoring the productivity of workers were costless, the owners of a business could ensure that their managers and workers were working effectively. In most firms, however, owners can't monitor everything that employees do—employees are better informed than owners. This information asymmetry creates a **principal–agent problem.**

principal–agent problem Problem arising when managers (agents) pursue their own goals even when doing so entails lower profits for a firm's owners (the principals).

An *agency relationship* exists whenever there is an arrangement in which one person's welfare depends on what another person does. The **agent** is the person who acts, and the **principal** is the party whom the action affects. In our example, the manager and the workers are agents and the owner is the principal. *The principal–agent problem is that managers may pursue their own goals, even at the cost of obtaining lower profits for owners.*

agent Individual employed by a principal to achieve the principal's objective.

principal Individual who employs one or more agents to achieve an objective.

Agency relationships are widespread in our society. For example, doctors serve as agents for hospitals and, as such, may select patients and do procedures which, though consistent with their personal preferences, are not necessarily consistent with the objectives of the hospital. Similarly, managers of housing properties may not maintain the property the way that the owners would like. And sometimes insured parties may be seen as agents and insurance companies as principals.

How does incomplete information and costly monitoring affect the way agents act? And what mechanisms can give managers the incentives to operate in the owner's interest? These questions are central to any principal–agent analysis. In this section, we study the principal–agent problem from several perspectives. First, we look at the owner–manager problem within private and public enterprises. Second, we discuss ways in which owners can use contractual relationships with their employees to deal with the principal–agent problems.

[9] *American Banker*, October 9, 1990

The Principal–Agent Problem in Private Enterprises

An individual family or financial institution owns more than 10 percent of the shares of only 16 of the 100 largest industrial corporations.[10] Clearly, most large firms are controlled by management. The fact that most individual stockholders have only a small percentage of a firm's total equity makes it difficult for them to obtain information about how well the firm's managers are performing. One function of owners (or their representatives) is to monitor the behavior of managers. But monitoring is costly, and information is expensive to gather and use, especially for an individual.[11]

Managers of private enterprises can thus pursue their own objectives. But what are these objectives? One view is that managers are more concerned with growth than with profit per se: More rapid growth and larger market share provide more cash flow, which in turn allows managers to enjoy more perks. Another view emphasizes the utility that managers get from their jobs, not only from profit but also from the respect of their peers, the power to control the corporation, the fringe benefits and other perks, and long job tenure.

However, there are important limitations to managers' ability to deviate from the objectives of owners. First, stockholders can complain loudly when they feel that managers are behaving improperly. In exceptional cases they can oust the current management (perhaps with the help of the board of directors, whose job it is to monitor managerial behavior). Second, a vigorous market for corporate control can develop. If a takeover bid becomes more likely when the firm is poorly managed, managers will have a strong incentive to pursue the goal of profit maximization. Third, there can be a highly developed market for managers. If managers who maximize profit are in great demand, they will earn high wages and so give other managers an incentive to pursue the same goal.

Unfortunately, the means by which stockholders control managers' behavior are limited and imperfect. Corporate takeovers may be motivated by personal and economic power, for example, instead of economic efficiency. The managerial labor market may also work imperfectly, given that top managers are frequently near retirement and have long-term contracts. As a result, it is important to look for solutions to the principal–agent problem in which owners alter the incentives that managers face, without resorting to government intervention. We consider some of these solutions in the next section.

The Principal–Agent Problem in Public Enterprises

The principal–agent framework can also help us understand the behavior of the managers of public organizations. These managers may also be interested in power and perks, both of which can be obtained by expanding their organization beyond its "efficient" level. Because it is also costly to monitor the behavior of public managers, there are no guarantees that they will produce the efficient output. Legislative checks on a government agency are not likely to be effective as long as the agency has better information about its costs than the legislature has.

Although the public sector lacks some of the market forces that keep private managers in line, government agencies can still be effectively monitored. First, managers of government agencies care about more than just the size of their agencies. Indeed, many choose lower-paying public jobs because they are concerned about the "public interest." Second, public managers are subject to the rigors of the managerial job market, much the way private managers are. If public managers are perceived to be pursuing improper objectives, their ability to obtain high salaries in the future might be impaired. Third, the legislature and other government agencies perform an oversight function. For example, the Government Accounting Office and the Office of Management and Budget spend much of their energy monitoring other agencies.

[10] See Merritt B. Fox, *Finance and Industrial Performance in a Dynamic Economy* (New York: Columbia University Press, 1987).

[11] There are economies of scale in gathering information but there is no obvious way in which the information can be sold.

At the local rather than the federal level, public managers are subject to even more checks. Suppose, for example, that a city transit agency has expanded bus service beyond the efficient level. Citizens can vote the transit managers out of office, or, if all else fails, use alternative transportation or move. Competition among agencies can be as effective as competition among private firms in constraining the non–profit-maximizing behavior of managers.

EXAMPLE 17.5 Managers of Nonprofit Hospitals as Agents

Do the managers of nonprofit organizations have the same goals as those of for-profit organizations? Are nonprofit organizations more or less efficient than for-profit firms? We can get some insight into these issues by looking at the provision of health care. In a study of 725 hospitals, from 14 major hospital chains, the return on investment and average costs of nonprofit and for-profit hospitals were compared to determine if they performed differently.[12]

The study found that for 1977 and 1981 the rate of returns between the two types of hospitals did indeed differ. In 1977, for example, for-profits earned an 11.6 percent return, while nonprofits earned 8.8 percent. In 1981, for-profits earned 12.7 percent and nonprofits only 7.4 percent. A straight comparison of returns and costs is not appropriate, however, because the hospitals perform different functions. For example, 24 percent of the nonprofit hospitals provide medical residency programs, as compared with only 6 percent of the for-profit hospitals. Similar differences can be found in the provision of speciality care, where 10 percent of the nonprofits have open-heart units as compared with only 5 percent of the for-profits. In addition, while 43 percent of nonprofits have premature infant units, only 29 percent of the for-profits have the equivalent units.

Using a statistical regression analysis, which controls for differences in the services performed, one can determine whether differences in services account for the higher costs. The study found that after adjusting for services performed, the average cost of a patient day in nonprofit hospitals was 8 percent higher than in for-profit hospitals. This implies that the profit status of the hospital affects its performance in the way principal–agent theory predicts: Without the competitive forces faced by for-profit hospitals, nonprofit hospitals may be less cost-conscious and therefore less likely to serve appropriately as agents for their principals—namely, society at large.

Of course, nonprofit hospitals provide services that society may well wish to subsidize. But the added cost of running a nonprofit hospital should be considered when determining whether it should be granted tax-exempt status.

[12] Regina E. Herzlinger and William S. Krasker, "Who Profits from Nonprofits?" *Harvard Business Review* 65 (January–February 1987): 93–106.

Incentives in the Principal–Agent Framework

We have seen why managers' and owners' objectives are likely to differ within the principal–agent framework. How, therefore, can owners design reward systems so that managers and workers come as close as possible to meeting owners' goals? To answer this question, let's study a specific problem.[13]

A small manufacturer uses labor and machinery to produce watches. The owners want to maximize profit. They must rely on a machine repairperson whose effort will influence the likelihood that machines break down and thus affect the firm's profit level. Revenue also depends on other random factors, such as the quality of parts and the reliability of other labor. As a result of high monitoring costs, the owners can neither measure the effort of the repairperson directly nor be sure that the same effort will always generate the same profit level. Table 17.2 describes these circumstances.

The table shows that the repairperson can work with either a low or high amount of effort. Low effort ($a = 0$) generates either \$10,000 or \$20,000 in revenue (with equal probability), depending on the random factors that we mentioned. We've labeled the lower of the two revenue levels "poor luck" and the higher level "good luck." When the repairperson makes a high effort ($a = 1$), revenue will be either \$20,000 (poor luck) or \$40,000 (good luck). These numbers highlight the problem of incomplete information: When the firm's revenue is \$20,000, the owners cannot know whether the repairperson has made a low or high effort.

Suppose the repairperson's goal is to maximize the wage payment that he receives, net of the cost (in terms of lost leisure and unpleasant work time) of the effort that he makes. To simplify, we'll suppose that the cost of effort is 0 for low effort and \$10,000 for high effort. (Formally, $c = \$10{,}000a$.)

Now we can state the principal–agent problem from the owners' perspective. The owners' goal is to maximize expected profit, given the uncertainty of outcomes and given the fact that the repairperson's behavior cannot be monitored. The owners can contract to pay the repairperson for his work, but the payment scheme must be based entirely on the measurable output of the manufacturing process, not on the repairperson's effort. To signify this link, we describe the payment scheme as $w(R)$, stressing that payments can depend only on measured revenue.

What is the best payment scheme? And can that scheme be as effective as one based on effort rather than output? We can only begin to study the answers here. The best payment scheme depends on the nature of production, the degree of uncertainty, and the objectives of both owners and managers. The arrangement will not always be as effective as an ideal scheme directly tied to effort. A lack of information can lower economic efficiency because both the owners' revenue and the repairperson's payment may fall at the same time.

TABLE 17.2 Revenue from Making Watches

	POOR LUCK	GOOD LUCK
Low effort ($a = 0$)	\$10,000	\$20,000
High effort ($a = 1$)	\$20,000	\$40,000

[13] This discussion was motivated in part by Bengt Holmstrom, "Moral Hazard and Observability," *Bell Journal of Economics* 10 (1979): 74–91.

Let's see how to design a payment scheme when the repairperson wishes to maximize his payment received net of the cost of effort made.[14] Suppose first that the owners offer a fixed wage payment. Any wage will do, but we can see things most clearly if we assume that the wage is 0. (Here, 0 could represent a wage no higher than the wage rate paid in other comparable jobs.) Facing a wage of 0, the repairperson has no incentive to make a high level of effort. The reason is simple: The repairperson does not share in any of the gains that the owners enjoy from the increased effort. It follows, therefore, that a fixed payment will lead to an inefficient outcome. When $a = 0$ and $w = 0$, the owner will earn an expected revenue of \$15,000 and the repairperson a net wage of 0.

Both the owners and the repairperson will be better off if the repairperson is rewarded for his productive effort. Suppose, for example, that the owners offer the repairperson the following payment scheme:

$$\begin{aligned} &\text{If } R = \$10{,}000 \text{ or } \$20{,}000,\ w = 0 \\ &\text{If } R = \$40{,}000,\ w = \$24{,}000 \end{aligned} \qquad (17.1)$$

Under this bonus arrangement, a low effort generates no payment. A high effort, however, generates an expected payment of \$12,000, and an expected payment net of the cost of effort of \$12,000 − \$10,000 = \$2000. Now, the repairperson will choose to make a high level of effort. This arrangement makes the owners better off than before because they get an expected revenue of \$30,000 and a profit of \$18,000.

This is not the only payment scheme that will work for the owners, however. Suppose they contract to have the worker participate in the following revenue-sharing arrangement. When revenues are greater than \$18,000,

$$w = R - \$18{,}000 \qquad (17.2)$$

(Otherwise the wage is zero.) In this case, if the repairperson makes a low effort, he receives an expected payment of \$1000. But if he makes a high level of effort, his expected payment is \$12,000 and his expected payment net of the \$10,000 cost of effort is \$2000. (The owners' profit is \$18,000, as before.)

Thus, in our example, a revenue-sharing arrangement achieves the same outcome as a bonus-payment system. In more complex situations, the incentive effects of the two types of arrangements will differ. However, the basic idea illustrated here applies to all principal–agent problems: When it is impossible to measure effort directly, an incentive structure that rewards the outcome of high levels of effort can induce agents to aim for the goals that the owners set.

*17.5 Managerial Incentives in an Integrated Firm

We have seen that owners and managers of firms can have asymmetric information about demand, cost, and other variables. We've also seen how owners can design reward structures to encourage managers to make appropriate efforts. Now we focus our attention on firms that are *integrated*—that consist of several divisions, each with its own managers. Some firms are **horizontally integrated:** Several plants produce the same or related products. Others are also **vertically integrated:**

horizontal integration Organizational form in which several plants produce the same or related products for a firm.

vertical integration Organizational form in which a firm contains several divisions, with some producing parts and components that others use to produce finished products.

[14] We assume that the repairperson is risk neutral, so that no efficiency is lost. If, however, the repairperson were risk averse, there would be an efficiency loss

Upstream divisions produce materials, parts, and components that downstream divisions use to produce final products. Integration creates organizational problems. We addressed some of these problems in the appendix to Chapter 11, where we discussed *transfer pricing* in the vertically integrated firm—that is, how the firm sets prices for parts and components that upstream divisions supply to downstream divisions. Here we will examine problems that stem from asymmetric information.

Asymmetric Information and Incentive Design in the Integrated Firm

In an integrated firm, division managers are likely to have better information about their different operating costs and production potential than central management has. This asymmetric information causes two problems.

1. How can central management elicit accurate information about divisional operating costs and production potential from divisional managers? This is important because the inputs to some divisions may be the outputs of other divisions, because deliveries must be scheduled to customers, and because prices cannot be set without knowing overall production capacity and costs.
2. What reward or incentive structure should central management use to encourage divisional managers to produce as efficiently as possible? Should they be given bonuses based on how much they produce, and if so, how should they be structured?

To understand these problems, consider a firm with several plants that all produce the same product. Each plant's manager has much better information about its production capacity than central management has. In order to avoid bottlenecks and schedule deliveries reliably, central management wants to learn more about how much each plant can produce. It also wants each plant to produce as much as possible. Let's examine how central management can obtain the information it wants while also encouraging the plant managers to run the plants as efficiently as possible.

One way is to give the plant managers bonuses based on either the total output of their plant or its operating profit. While this approach would encourage managers to maximize output, it would penalize managers whose plants have higher costs and lower capacity. Even if these plants produced efficiently, their output and operating profit—and thus their bonuses—would be lower than those of plants with lower costs and higher capacities. Plant managers would also have no incentive to obtain and reveal accurate information about cost and capacity.

A second way is to ask the managers about their costs and capacities and *then* base bonuses on how well they do relative to their answers. For example, each manager might be asked how much his or her plant can produce each year. Then at the end of the year, the manager receives a bonus based on how close the plant's output was to this target. For example, if the manager's estimate of the feasible production level is Q_f, the annual bonus in dollars, B, might be

$$B = 10{,}000 - .5(Q_f - Q) \qquad (17.3)$$

where Q is the plant's actual output, 10,000 is the bonus when output is at capacity, and .5 is a factor chosen to reduce the bonus if Q is below Q_f.

With this scheme, however, managers would have an incentive to *underestimate* capacity. By claiming capacities below what they know to be true, they can more easily earn large bonuses even if they do not operate efficiently. For example,

if a manager estimates capacity to be 18,000 rather than 20,000, and the plant actually produces only 16,000, her bonus increases from $8000 to $9000. Thus this scheme fails to elicit accurate information about capacity and does not ensure that plants will be run as efficiently as possible.

Now let's modify this scheme. We will still ask managers how much their plants can feasibly produce and tie their bonuses to this estimate. However, we will use a slightly more complicated formula than the one in (17.3) to calculate the bonus:

$$\begin{aligned} &\text{If } Q > Q_f, \quad B = .3Q_f + .2(Q - Q_f) \\ &\text{If } Q \le Q_f, \quad B = .3Q_f - .5(Q_f - Q) \end{aligned} \qquad (17.4)$$

The parameters (.3, .2, and .5) have been chosen so that each manager has the incentive to reveal the *true* feasible production level *and* to make Q, the actual output of the plant, as large as possible.

To see that this scheme does the job, look at Figure 17.4. Assume that the true production limit is $Q^* = 20{,}000$ units per year. The bonus that the manager will receive if she states feasible capacity to be the true production limit is given by the line labeled $Q_f = 20{,}000$. This line is continued for outputs beyond 20,000 to illustrate the bonus scheme but dashed to signify the infeasibility of such production. Note that the manager's bonus is maximized when the firm produces at its limit of 20,000 units; the bonus is then $6000.

Suppose, however, that the manager reports a feasible capacity of only 10,000. Then the bonus she receives is given by the line labeled $Q_f = 10{,}000$. The maximum bonus is now $5000, which is obtained by producing an output of 20,000. But note that this is less than the bonus the manager would receive if she correctly stated the feasible capacity to be 20,000.

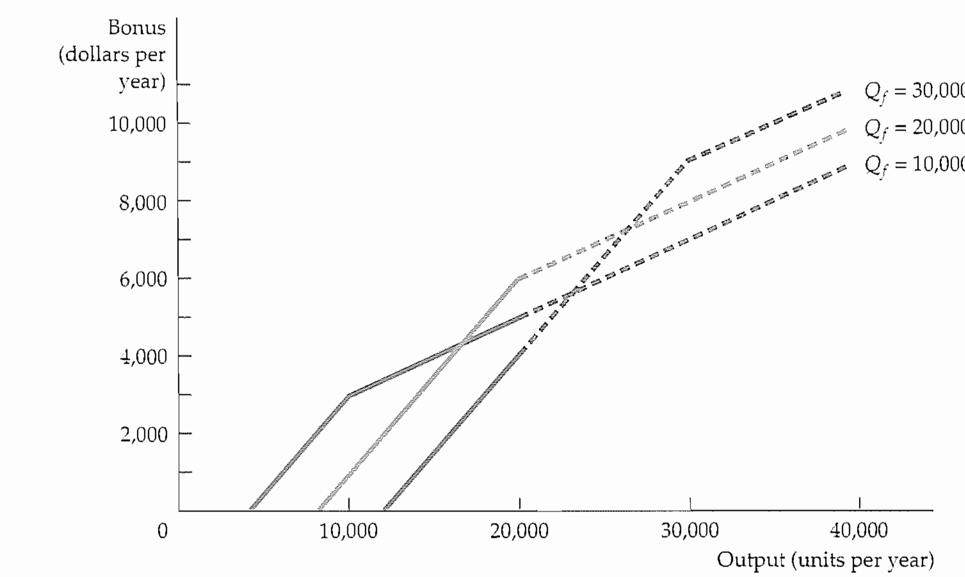

FIGURE 17.4 Incentive Design in an Integrated Firm

A bonus scheme can be designed that gives a manager the incentive to estimate accurately the size of the plant. If the manager reports a feasible capacity of 20,000 units per year, equal to the actual capacity, then the bonus received is maximized (at $6000).

The same line of argument applies when the manager exaggerates available capacity. If the manager states feasible capacity to be 30,000 units per year, the bonus is given by the line $Q_f = 30{,}000$. The maximum bonus of $4000, which is achieved at an output of 20,000, is less than the bonus she could have received had she reported feasible capacity correctly.[15]

Applications

Because the problem of asymmetric information and incentive design comes up often in managerial settings, incentive schemes like the one described above arise in many contexts. How, for example, can managers encourage salespeople to set and reveal realistic sales targets and then work as hard as possible to meet them?

Most salespeople cover specific territories. A salesperson assigned to a densely populated urban territory can usually sell more product than a salesperson assigned to a sparsely populated area. The company, however, wants to reward all salespeople equitably. It also wants to give them the incentive to work as hard as possible and to report realistic sales targets, so that it can plan production and delivery schedules. Companies have always used bonuses and commissions to reward salespeople, but incentive schemes have often been poorly designed. Typically, salespeople's commissions were proportional to their sales. This approach elicited neither accurate information about feasible sales targets nor maximum performance.

Today, companies are learning that bonus schemes like the one given by equation (17.4) provide better results. The salesperson can be given an array of numbers showing the bonus as a function of both the sales target (chosen by the salesperson) and the actual level of sales. (The numbers would be calculated from equation 17.4 or some similar formula.) Salespeople will quickly figure out that they do best by reporting feasible sales targets and then working as hard as possible to meet them.[16]

17.6 Asymmetric Information in Labor Markets: Efficiency Wage Theory

Recall from §14.1 that in a perfectly competitive labor market, firms hire labor to the point at which the real wage (the wage divided by the price of the product) is equal to the marginal product of labor.

When the labor market is competitive, all who wish to work will find jobs for wages equal to their marginal products. Yet most countries have substantial unemployment even though many people are aggressively seeking work. Many of the unemployed would presumably work even for a lower wage rate than that being received by employed people. Why don't we see firms cutting wage rates, increasing employment levels, and thereby increasing profit? Can our models of competitive equilibrum explain persistent unemployment?

efficiency wage theory Explanation for the presence of unemployment and wage discrimination which recognizes that labor productivity may be affected by the wage rate.

In this section, we show how the **efficiency wage theory** can explain the presence of unemployment and wage discrimination.[17] We have thus far determined labor productivity according to workers' abilities and firms' investment in capital. Efficiency wage models recognize that labor productivity also depends on the wage rate. There are various explanations for this relationship. Economists have suggested that the productivity of workers in developing countries depends on the wage rate for nutritional reasons: Better-paid workers can afford to buy more and better food and are therefore healthier and can work harder.

shirking model Principle that workers still have an incentive to shirk if a firm pays them a market-clearing wage, because fired workers can be hired somewhere else for the same wage.

A better explanation for the United States is found in the **shirking model.** Because monitoring workers is costly or impossible, firms have imperfect information about worker productivity, and there is a principal–agent problem. In its simplest form, the shirking model assumes perfectly competitive markets in which all workers are equally productive and earn the same wage. Once hired, workers can either work productively or slack off (shirk). But because information about their performance is limited, workers may not get fired for shirking.

The model works as follows. If a firm pays its workers the market-clearing wage w^*, they have an incentive to shirk. Even if they get caught and are fired (and they might not be), they can immediately get hired somewhere else for the same wage. In this situation, because the threat of being fired does not impose a cost on workers, they have no incentive to be productive. As an incentive not to shirk, a firm must offer workers a higher wage. At this higher wage, workers who are fired for shirking will face a decrease in wages when hired by another firm at w^*. If the difference in wages is large enough, workers will be induced to be productive, and this firm will not have a problem with shirking. The wage at which no shirking occurs is the **efficiency wage.**

efficiency wage Wage that a firm will pay to an employee as an incentive not to shirk.

Up to this point, we have looked at only one firm. But all firms face the problem of shirking. All firms, therefore, will offer wages greater than the market-clearing wage w^*—say, w_e (efficiency wage). Does this remove the incentive for workers not to shirk because they will be hired at the higher wage by other firms if they get fired? No. Because all firms are offering wages greater than w^*, the demand for labor is less than the market-clearing quantity, and there is unemployment. Consequently, workers fired for shirking will face spells of unemployment before earning w_e at another firm.

In §14.2, we explain that the equilibrium wage is given by the intersection of the demand for labor curve and the supply of labor curve.

Figure 17.5 shows shirking in the labor market. The demand for labor D_L is downward-sloping for the traditional reasons. If there were no shirking, the intersection of D_L with the supply of labor (S_L) would set the market wage at w^*, and full employment would result (L^*). With shirking, however, individual firms are unwilling to pay w^*. Rather, for every level of unemployment in the labor market, firms must pay some wage greater than w^* to induce workers to be productive. This wage is shown as the no-shirking constraint (NSC) curve. This curve shows the minimum wage, for each level of unemployment, that workers need earn in order not to shirk. Note that the greater the level of unemployment, the smaller the difference between the efficiency wage and w^*. Why is this so? Because with high levels of unemployment, people who shirk risk long periods of unemployment and therefore don't need much inducement to be productive.

In Figure 17.5, the equilibrium wage will be at the intersection of the NSC curve and D_L curves, with L_e workers earning w_e. This equilibrium occurs because the NSC curve gives the lowest wage that firms can pay and still discourage shirking. Firms need not pay more than this wage to get the number of workers they need, and they will not pay less because a lower wage will encourage shirking. Note that the NSC curve never crosses the labor supply curve. This means that there will always be some unemployment in equilibrium.

[15] Any bonus of the form $B = \beta Q_f + \alpha(Q - Q_f)$ *for* $Q > Q_f$, and $B = \beta Q_f - \gamma(Q_f - Q)$ for $Q \leq Q_f$, with $\gamma > \beta > \alpha > 0$ will work. See Martin L. Weitzman, "The New Soviet Incentive Model," *Bell Journal of Economics* VII (Spring 1976): 251–56. There is a dynamic problem with this scheme that we have ignored: Managers must weigh a large bonus for good performance this year against being assigned more ambitious targets in the future. This is discussed in Martin Weitzman, "The 'Ratchet Principle' and Performance Incentives," *Bell Journal of Economics* 11 (Spring 1980): 302–08.

[16] See Jacob Gonik, "Tie Salesmen's Bonuses to Their Forecasts," *Harvard Business Review* (May–June 1978): 116–23.

[17] See Janet L. Yellen, "Efficiency Wage Models of Unemployment," *American Economic Review* 74 (May 1984): 200–05. The analysis relies on Joseph E. Stiglitz, "The Causes and Consequences of the Dependence of Quality on Price," *Journal of Economic Literature* 25 (March 1987): 1–48.

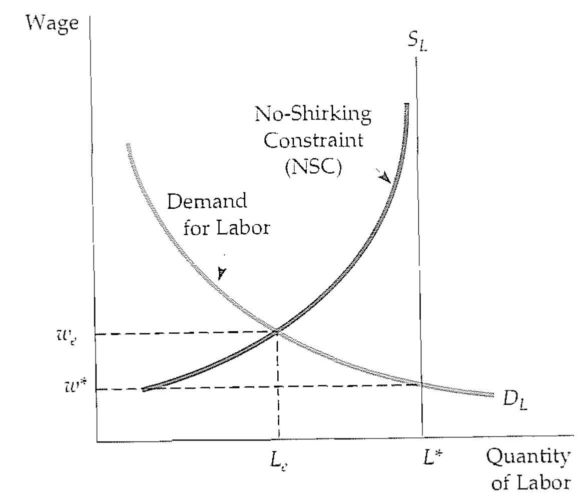

FIGURE 17.5 Unemployment in a Shirking Model

Unemployment can arise in otherwise competitive labor markets when employers cannot accurately monitor workers. Here the "no shirking constraint" (NSC) gives the wage necessary to keep workers from shirking. The firm hires L_e workers (at a higher than competitive efficiency wage w_e), creating $L^* - L_e$ of unemployment.

EXAMPLE 17.6 Efficiency Wages at Ford Motor Company

One of the early examples of the payment of efficiency wages can be found in the history of Ford Motor Company. Before 1913, automobile production depended heavily on skilled workers. But the introduction of the assembly line drastically changed the workplace. Now jobs demanded much less skill, and production depended on maintaining assembly-line equipment. But as automobile plants changed, workers became increasingly disenchanted. In 1913, turnover at Ford was 380 percent. The following year, it rose to 1000 percent, and profit margins fell sharply.

Ford needed to maintain a stable workforce, and Henry Ford (and his business partner James Couzens) provided it. In 1914, when the going wage for a day's work in industry averaged between $2 and $3, Ford introduced a pay policy of $5 a day. Improved labor efficiency (not generosity) was behind this policy. The goal was to attract better workers who would stay with their jobs—and eventually to increase profits.

Although Henry Ford was attacked for it, this policy succeeded. His workforce did become more stable, and the publicity helped Ford's sales. In addition, because Ford had his pick of workers, he could hire a group that was on average more productive. Ford stated that the wage increase did in fact increase the loyalty and personal efficiency of his workers, and quantitative estimates support his statements. According to calculations by Ford's chief of labor relations, productivity increased by 51 percent. Another study found that absenteeism had been halved, and discharges for cause had declined sharply. Thus the productivity increase more than offset the increase in wages. As a result, Ford's profitability rose from $30 million in 1914 to $60 million in 1916.

SUMMARY

1. The seller of a product often has better information about its quality than the buyer. Asymmetric information of this type creates a market failure in which bad products tend to drive good products out of the market. Market failure can be eliminated if sellers offer standardized products, provide guarantees or warranties, or find other ways to maintain good reputations for their products.
2. Insurance markets frequently involve asymmetric information because the insuring party has better information about the risk involved than the insurance company. This can lead to adverse selection, in which poor risks choose to insure and good risks do not. Another problem for insurance markets is moral hazard, in which the insuring party takes less care to avoid losses after insuring.
3. Sellers can deal with the problem of asymmetric information by sending buyers signals about the quality of their products. For example, workers can signal their high productivity by obtaining high levels of education.
4. Asymmetric information may make it costly for the owners of firms (principals) to monitor accurately the behavior of their managers (agents). Managers may seek higher fringe benefits for themselves, or a goal of sales maximization, even though shareholders would prefer to maximize profit.
5. Owners can avoid some principal–agent problems by designing contracts that give their agents the incentive to perform productively.
6. Asymmetric information can explain why labor markets have unemployment even though some workers are actively seeking work. According to efficiency wage theory, a wage higher than the competitive wage (the efficiency wage) increases worker productivity by discouraging workers from shirking on the job.

QUESTIONS FOR REVIEW

1. Why can asymmetric information between buyers and sellers lead to market failure when a market is otherwise perfectly competitive?
2. If the used car market is a "lemons" market, how would you expect the repair record of used cars that are sold to compare with the repair record of those not sold?
3. Explain the difference between adverse selection and moral hazard in insurance markets. Can one exist without the other?
4. Describe several ways in which sellers can convince buyers that their products are of high quality. Which methods apply to the following products: Maytag washing machines, Burger King hamburgers, large diamonds?
5. Why might a seller find it advantageous to signal the quality of a product? How are guarantees and warranties a form of market signaling?
6. Why might managers be able to achieve objectives other than profit maximization, the goal of the firm's shareholders?
7. How can the principal–agent model be used to explain why public enterprises, such as post offices, might pursue goals other than profit maximization?
8. Why are bonus and profit-sharing payment schemes likely to resolve principal–agent problems, whereas a fixed-wage payment will not?
9. What is an efficiency wage? Why is it profitable for the firm to pay it when workers have better information about their productivity than firms do?

EXERCISES

1. Many consumers view a well-known brand name as a signal of quality and will pay more for a brand-name product (e.g., Bayer aspirin instead of generic aspirin, or Birds Eye frozen vegetables instead of the supermarket's own brand). Can a brand name provide a useful signal of quality? Why or why not?
2. Gary is a recent college graduate. After six months at his new job, he has finally saved enough to buy his first car.
   a. Gary knows very little about the differences between makes and models. How could he use market signals, reputation, or standardization to make comparisons?

b. You are a loan officer in a bank. After selecting a car, Gary comes to you seeking a loan. Because he has only recently graduated, he does not have a long credit history. Nonetheless, the bank has a long history of financing cars for recent college graduates. Is this information useful in Gary's case? If so, how?

3. A major university bans the assignment of D or F grades. It defends its action by claiming that students tend to perform above average when they are free from the pressures of flunking out. The university states that it wants all its students to get As and Bs. If the goal is to raise overall grades to the B level or above, is this a good policy? Discuss with respect to the problem of moral hazard.

4. Professor Jones has just been hired by the economics department at a major university. The president of the board of regents has stated that the university is committed to providing top-quality education for undergraduates. Two months into the semester, Jones fails to show up for his classes. It seems he is devoting all his time to economic research rather than to teaching. Jones argues that his research will bring additional prestige to the department and the university. Should he be allowed to continue exclusively with research? Discuss with reference to the principal–agent problem.

5. Faced with a reputation for producing automobiles with poor repair records, a number of American companies have offered extensive guarantees to car purchasers (e.g., a seven-year warranty on all parts and labor associated with mechanical problems).
   a. In light of your knowledge of the lemons market, why is this a reasonable policy?
   b. Is the policy likely to create a moral hazard problem? Explain.

6. To promote competition and consumer welfare, the Federal Trade Commission requires firms to advertise truthfully. How does truth in advertising promote competition? Why would a market be less competitive if firms advertised deceptively?

7. An insurance company is considering issuing three types of fire insurance policies: (i) complete insurance coverage, (ii) complete coverage above and beyond a $10,000 deductible, and (iii) 90 percent coverage of all losses. Which policy is more likely to create moral hazard problems?

8. You have seen how asymmetric information can reduce the average quality of products sold in a market, as low-quality products drive out the high-quality ones. For those markets in which asymmetric information is prevalent, would you agree or disagree with each of the following? Explain briefly:
   a. The government should subsidize *Consumer Reports*.
   b. The government should impose quality standards—e.g., firms should not be allowed to sell low-quality items.
   c. The producer of a high-quality good will probably want to offer an extensive warranty.
   d. The government should require *all* firms to offer extensive warranties.

9. Two used car dealerships compete side by side on a main road. The first, Harry's Cars, sells high-quality cars that it carefully inspects and, if necessary, services. On average, it costs Harry's $8000 to buy and service each car that it sells. The second dealership, Lew's Motors, sells lower-quality cars. On average, it costs Lew's only $5000 for each car that it sells. If consumers knew the quality of the used cars they were buying, they would gladly pay $10,000 on average for Harry's cars and pay only $7000 on average for Lew's cars.

   Unfortunately, the dealerships are too new to have established reputations, so consumers do not know the quality of each dealership's cars. Consumers figure, however, that they have a 50–50 chance of ending up with a high-quality car, no matter which dealership they go to, and are thus willing to pay $8500 on average for a car.

   Harry's has an idea: It will offer a bumper-to-bumper warranty for all cars it sells. It knows that a warranty lasting Y years will cost $500Y$ on average, and it also knows that if Lew's tries to offer the same warranty, it will cost Lew's $1000Y$ on average.
   a. Suppose Harry's offers a one-year warranty on all cars it sells. Will this generate a credible signal of quality? Will Lew's match the offer, or will it fail to match it so that consumers can correctly assume that because of the warranty, Harry's cars are high quality and so worth $10,000 on average?
   b. What if Harry's offers a two-year warranty? Will this generate a credible signal of quality? What about a three-year warranty?
   c. If you were advising Harry's, how long a warranty would you urge it to offer? Explain why.

10. A firm's short-run revenue is given by $R = 10e - e^2$, where e is the level of effort by a typical worker (all workers are assumed to be identical). A worker chooses his level of effort to maximize his wage net of effort $w - e$ (the per-unit cost of effort is assumed to be 1). Determine the level of effort and the level of profit (revenue less wage paid) for each of the following wage arrangements. Explain why these different principal–agent relationships generate different outcomes.
   a. $w = 2$ for $e \geq 1$; otherwise $w = 0$.
   b. $w = R/2$.
   c. $w = R - 12.5$.

CHAPTER 18

Externalities and Public Goods

Chapter Outline

List of Examples

In this chapter we study *externalities*—the effects of production and consumption activities not directly reflected in the market—and *public goods*—goods that benefit all consumers, but that the market either undersupplies or does not supply at all. Externalities and public goods are important sources of market failure and thus raise serious public policy questions. For example, how much waste, if any, should firms be allowed to dump into rivers and streams? How strict should automobile emission standards be? How much money should the government spend on national defense? Education? Basic research? Public television?

When externalities are present, the price of a good need not reflect its social value. As a result, firms may produce too much or too little, so that the market outcome is inefficient. We begin by describing externalities and showing exactly how they create market inefficiencies. We then evaluate remedies. While some remedies involve government regulation, others rely primarily on bargaining among individuals or on the legal right of those adversely affected to sue those who create an externality.

Next, we analyze public goods. The marginal cost of providing a public good to an additional consumer is zero, and people cannot be prevented from consuming it. We distinguish between those goods that are difficult to provide privately and those that could have been provided by the market. We conclude by describing the problem policymakers face when trying to decide how much of a public good to provide.

18.1 Externalities

Externalities can arise between producers, between customers, or between consumers and producers. They can be *negative*—when the action of one party imposes costs on another party—or *positive*—when the action of one party benefits another party.

A *negative externality* occurs, for example, when a steel plant dumps its waste in a river that fishermen downstream depend on for their daily catch. The more waste the steel plant dumps in the river, the fewer fish will be supported. The firm, however, has no incentive to account for the external costs that it imposes on fishermen when making its production decision. Furthermore, there is no market in which these external costs

externality Action by either a producer or a consumer which affects other producers or consumers, but is not accounted for in the market price.

can be transmitted into the price of steel. A *positive externality* occurs when a home owner repaints her house and plants an attractive garden. All the neighbors benefit from this activity, yet the home owner's decision to repaint and landscape probably did not take these benefits into account.

Negative Externalities and Inefficiency

In §6.4, we explain that with a fixed-proportion production function, it is impossible to substitute among inputs because each level of output requires a specific combination of labor and capital.

In §8.3, we explain that because a competitive firm faces a horizontal demand curve, choosing its output so that marginal cost is equal to price is profit-maximizing.

Because externalities are not reflected in market prices, they can be a source of economic inefficiency. To see why, let's take our example of a steel plant dumping waste in a river. Figure 18.1(a) shows the production decision of a steel plant in a competitive market. Figure 18.1(b) shows the market demand and supply curves, assuming that all steel plants generate similar externalities. We assume that the firm has a fixed-proportions production function, so that it cannot alter its input combinations; waste and other effluent can be reduced only by lowering output. We will analyze the nature of the externality in two steps: first when only one steel plant pollutes, and then when all steel plants pollute in the same way.

The price of steel is P_1, at the intersection of the demand and supply curves in Figure 18.1(b). The MC curve in (a) gives a typical steel firm's marginal cost of production. The firm maximizes profit by producing output q_1, at which marginal cost is equal to price (which equals marginal revenue because the firm takes price as given). As the firm's output changes, however, the external cost imposed on fishermen downstream also changes. This external cost is given by the **marginal external cost** (MEC) curve in Figure 18.1(a). The curve is upward sloping for most forms of pollution: As the firm produces additional output and dumps additional effluent, the incremental harm to the fish industry increases.

marginal external cost Increase in cost imposed externally as one or more firms increase output by one unit.

From a social point of view, the firm produces too much output. The efficient level of output is the level at which the price of the product is equal to the

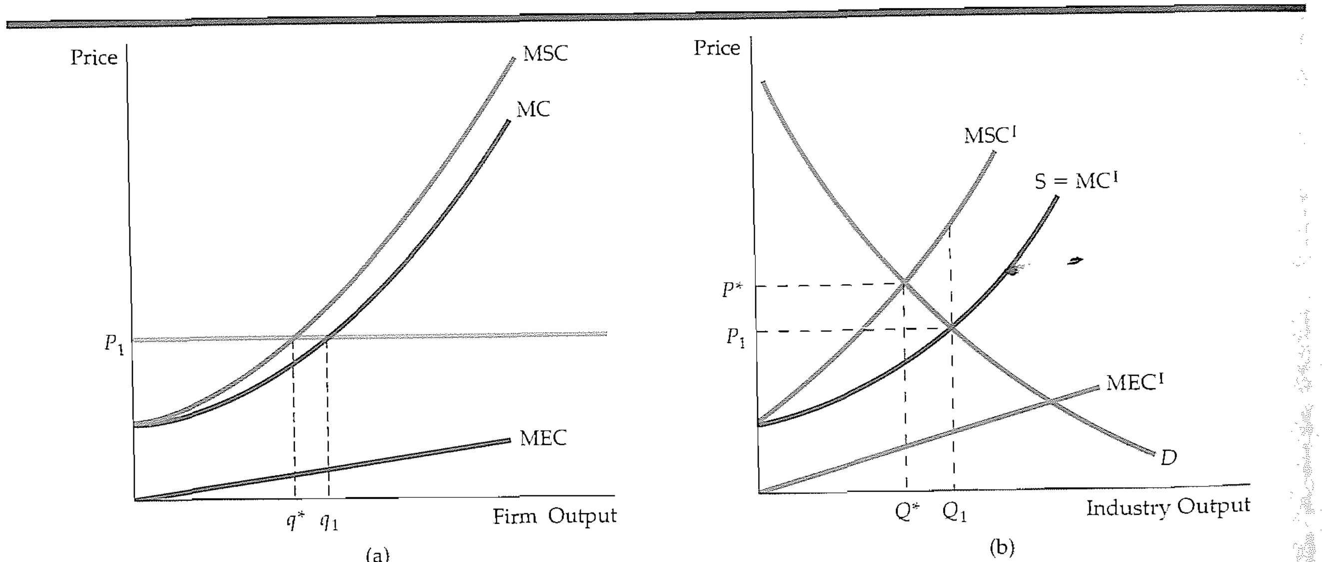

FIGURE 18.1 External Cost

When there are negative externalities, the marginal social cost MSC is higher than the marginal cost MC. The difference is the marginal external cost MEC. In (a), a profit-maximizing firm produces at q_1, where price is equal to MC. The efficient output is q^*, at which price equals MSC. In (b), the industry's competitive output is Q_1, at the intersection of industry supply MC^I and demand D. However, the efficient output Q^* is lower, at the intersection of demand and marginal social cost MSC^I.

marginal social cost of production: the marginal cost of production *plus* the marginal external cost of dumping effluent. In Figure 18.1(a), the marginal social cost curve is obtained by adding marginal cost and marginal external cost for each level of output (i.e., MSC = MC + MEC). The marginal social cost curve MSC intersects the price line at the output q^*. Because only one plant is dumping effluent into the river in this case, the market price of the product is unchanged. However, the firm is producing too much output (q_1 instead of q^*) and generating too much effluent.

marginal social cost Sum of the marginal cost of production and the marginal external cost

Now consider what happens when all steel plants dump their effluent into rivers. In Figure 18.1(b), the MC^I curve is the industry supply curve. The marginal external cost associated with the industry output, MEC^I, is obtained by summing the marginal cost of every person harmed at each level of output. The MSC^I curve represents the sum of the marginal cost of production and the marginal external cost *for all steel firms*. As a result, $MSC^I = MC^I + MEC^I$.

Is industry output efficient when there are externalities? As Figure 18.1(b) shows, the efficient industry output level is the level at which the marginal benefit of an additional unit of output is equal to the marginal social cost. Because the demand curve measures the marginal benefit to consumers, the efficient output is Q^*, at the intersection of the marginal social cost MSC^I and demand D curves. The competitive industry output, however, is at Q_1, the intersection of the demand curve and the supply curve, MC^I. Clearly, industry output is too high.

In §9.2, we explain that, absent market failure, a competitive market leads to the economically efficient output level.

In our example, each unit of output results in some effluent being dumped. Therefore, whether we are looking at one firm's pollution or the entire industry's, the economic inefficiency is the excess production that results in too much effluent being dumped in the river. The source of the inefficiency is the incorrect pricing of the product. The market price P_1 in Figure 18.1(b) is too low—it reflects the firms' marginal private cost of production, but not the marginal *social* cost. Only at the higher price P^* will steel firms produce the efficient level of output.

What is the cost to society of this inefficiency? For each unit produced above Q^*, the social cost is given by the difference between the marginal social cost and the marginal benefit (the demand curve). As a result, the aggregate social cost is shown in Figure 18.1(b) as the shaded triangle between MSC^I, D, and output Q_1.

Externalities generate both long-run and short-run inefficiencies. In Chapter 8, we saw that firms enter a competitive industry whenever the price of the product is above the *average cost* of production and exit whenever price is below average cost. In long-run equilibrium, price is equal to (long-run) average cost. When there are negative externalities, the average private cost of production is less than the average social cost. As a result, some firms remain in the industry even when it would be efficient for them to leave. Thus negative externalities encourage too many firms to remain in the industry.

Positive Externalities and Inefficiency

Externalities can also result in too little production, as the example of home repair and landscaping shows. In Figure 18.2, the horizontal axis measures the home owner's investment (in dollars) in repairs and landscaping. The marginal cost curve for home repair shows the cost of repairs as more work is done on the house; it is horizontal because this cost is unaffected by the amount of repairs. The demand curve D measures the marginal private benefit of the repairs to the home owner. The home owner will choose to invest q_1 in repairs, at the intersection of her demand and marginal cost curves. But repairs generate external

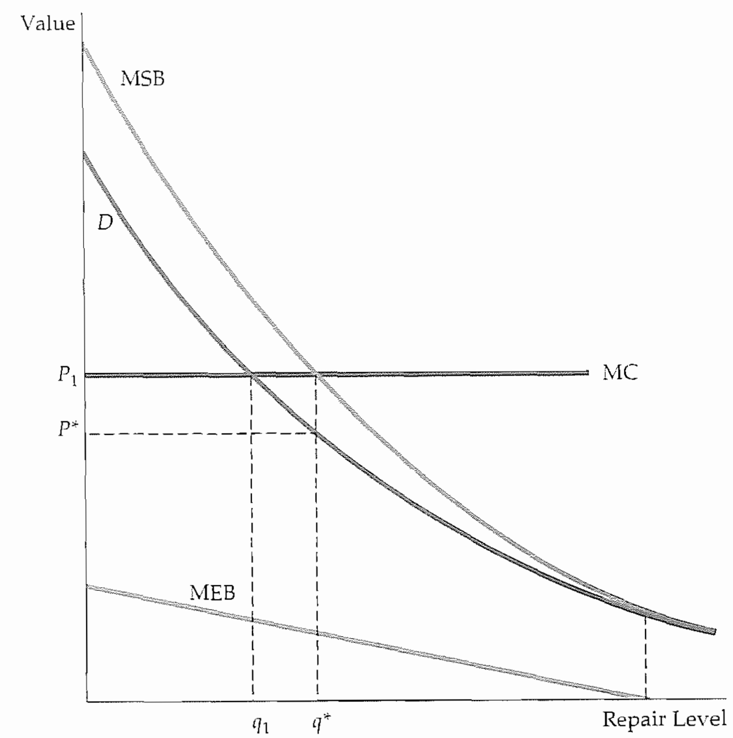

FIGURE 18.2 External Benefits

When there are positive externalities, marginal social benefits MSB are higher than marginal benefits D. The difference is the marginal external benefit MEB. A self-interested home owner invests q_1 in repairs, determined by the intersection of the marginal benefit curve D and the marginal cost curve MC. The efficient level of repair q^* is higher and is given by the intersection of the marginal social benefit and marginal cost curves.

marginal external benefit Increased benefit that accrues to other parties as a firm increases output by one unit.

marginal social benefit Sum of the marginal private benefit plus the marginal external benefit.

benefits to the neighbors, as the **marginal external benefit** curve, MEB, shows. This curve is downward sloping in this example because the marginal benefit is large for a small amount of repair but falls as the repair work becomes extensive.

The **marginal social benefit** curve MSB is calculated by adding the marginal private benefit and the marginal external benefit at every level of output. In short, MSB = D + MEB. The efficient level of output q^*, at which the marginal social benefit of additional repairs is equal to the marginal cost of those repairs, is found at the intersection of the MSB and MC curves. The inefficiency arises because the home owner doesn't receive all the benefits of her investment in repairs and landscaping. As a result, the price P_1 is too high to encourage her to invest in the socially desirable level of house repair. A lower price P^* is required to encourage the efficient level of supply, q^*.

Another example of a positive externality is the money that firms spend on research and development (R&D). Often the innovations resulting from research cannot be protected from other firms. Suppose, for example, that a firm designs a new product. If that design can be patented, the firm might earn a large profit by manufacturing and marketing the new product. But if the new design can be closely imitated by other firms, those firms can appropriate some of the developing firm's profit. Because there is then little reward for doing R&D, the market is likely to underfund it.

18.2 Ways of Correcting Market Failure

How can the inefficiency resulting from an externality be remedied? If the firm that generates the externality has a fixed-proportions production technology, the externality can be reduced only by encouraging the firm to produce less. As we saw in Chapter 8, this goal can be achieved through an output tax. Fortunately, most firms can substitute among inputs in the production process by altering their choices of technology. For example, a manufacturer can add a scrubber to its smokestack to reduce emissions.

Recall from §7.3 that a firm can substitute among inputs by changing technologies in response to an effluent fee.

Consider a firm that sells its output in a competitive market. The firm emits pollutants that damage air quality in a neighborhood. The firm can reduce its emissions, but only at a cost. Figure 18.3 illustrates this trade-off. The horizontal axis represents the level of factory emissions and the vertical axis the cost per unit of emissions. To simplify, we assume that the firm's output decision and its emissions decision are independent, and that the firm has already chosen its profit-maximizing output level. The firm is therefore ready to choose its preferred level of emissions. The curve labeled MSC represents the *marginal social cost of emissions*. This social cost curve represents the increased harm associated with the emissions of the factory and is therefore equivalent to the MEC curve described earlier. The MSC curve slopes upward because the *marginal* cost of the

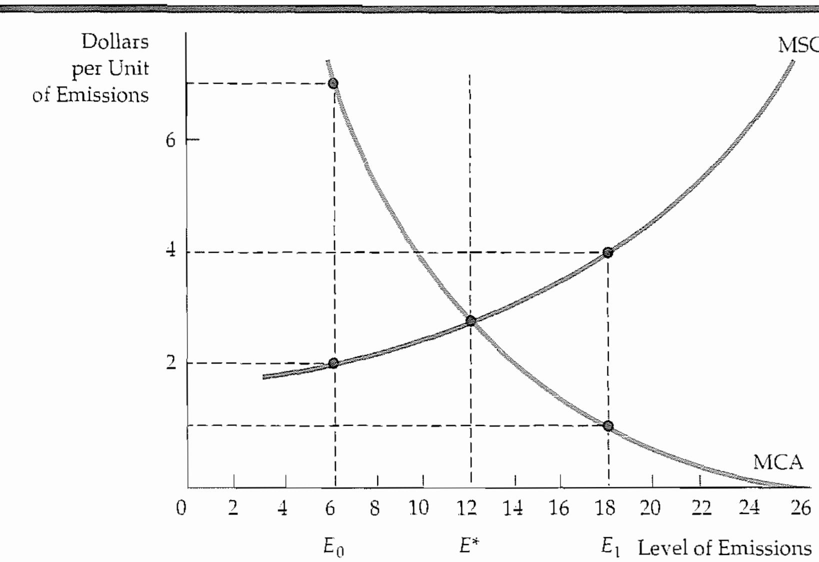

FIGURE 18.3 The Efficient Level of Emissions

The efficient level of factory emissions is the level that equates the marginal social cost of emissions MSC to the benefit associated with lower abatement costs MCA. The efficient level of 12 units is E^*.

externality is higher the more extensive it is. (Evidence from studies of the effects of air and water pollution suggests that small levels of pollutants generate little harm. However, the harm increases substantially as the level of pollutants increases.)

The curve labeled MCA is the *marginal cost of abating emissions*. It measures the additional cost to the firm of installing pollution control equipment. The MCA curve is downward sloping because the marginal cost of reducing emissions is low when the reduction has been slight, and high when it has been substantial. (A slight reduction is inexpensive—the firm can reschedule production so the greatest emissions occur at night, when few people are outside. Large reductions require costly changes in the production process.)

Because emissions reduction is costly and offers no direct benefit to the firm, the firm's profit-maximizing level of emissions is 26, the level at which the marginal cost of abatement is zero. The efficient level of emissions, 12 units, is at point E^*, where the marginal social cost of emissions, \$3, is equal to the marginal cost of abating emissions. Note that if emissions are lower than E^*—say, E_0—the marginal cost of abating emissions, \$7, is greater than the marginal social cost, \$2. Emissions, therefore, are too low relative to the social optimum. However, if the level of emissions is E_1, the marginal social cost, \$4, is greater than the marginal benefit, \$1. Emissions are then too high.

We can encourage the firm to reduce emissions to E^* in three ways: emissions standards, emissions fees, and transferable emissions permits.

An Emissions Standard

emissions standard Legal limit on the amount of pollutant that a firm can emit.

An **emissions standard** is a legal limit on how much pollutant a firm can emit. If the firm exceeds the limit, it can face monetary and even criminal penalties. In Figure 18.4, the efficient emission standard is 12 units, at point E^*. The firm will be heavily penalized for emissions greater than this level.

The standard ensures that the firm produces efficiently. The firm meets the standard by installing pollution-abatement equipment. The increased abatement expenditure will cause the firm's average cost curve to rise (by the average cost of abatement). Firms will find it profitable to enter the industry only if the price of the product is greater than the average cost of production plus abatement—the efficient condition for the industry.[1]

An Emissions Fee

emissions fee Charge levied on each unit of a firm's emissions.

An **emissions fee** is a charge levied on each unit of a firm's emissions. As Figure 18.4 shows, a \$3 emissions fee will generate efficient behavior by our factory. With this fee, the firm minimizes costs by reducing emissions from 26 to 12 units. To see why, note that the first unit of emissions can be reduced (from 26 to 25 units of emissions) at very little cost (the marginal cost of additional abatement is close to zero). For very little cost, therefore, the firm can avoid paying the \$3 per unit fee. In fact, for all levels of emission above 12 units, the marginal cost of abatement is less than the emissions fee. Thus it pays to reduce emissions. Below 12 units, however, the marginal cost of abatement is greater than the fee. In that case, the firm will prefer to pay the fee rather than further reduce emissions. The firm will therefore pay a total fee given by the gray-shaded rectangle

[1] This analysis assumes that the social costs of emissions do not change over time. If they do, the efficient standard will also change.

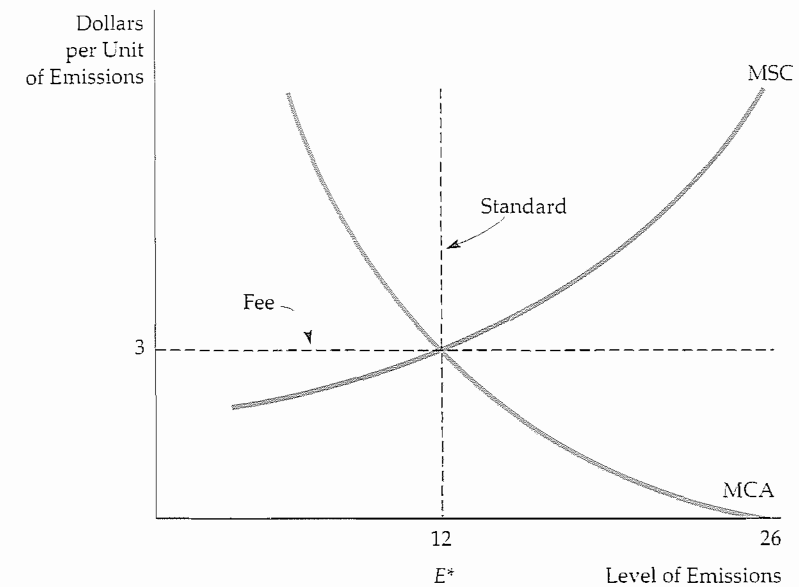

FIGURE 18.4 Standards and Fees

The efficient level of emissions at E^* can be achieved through either an emissions fee or an emissions standard. Facing a fee of \$3 per unit of emissions, a firm reduces emissions to the point at which the fee is equal to the marginal benefit. The same level of emissions reduction can be achieved with a standard that limits emissions to 12 units.

and incur a total abatement cost given by the blue-shaded triangle under the MCA curve to the right of $E = 12$. This cost is less than the fee the firm would pay if it did not reduce emissions at all.

Standards versus Fees

The United States has historically relied on standards to regulate emissions. However, other countries, such as Germany, have used fees successfully. Which method is better?

There are important differences between standards and fees when the policymaker has incomplete information and when it is costly to regulate firms' emissions. To understand these differences, let's suppose that because of administrative costs, the agency that regulates emissions must charge the same fee or set the same standard for all firms.

The Case for Fees First, let's examine the case for fees. Consider two firms that are located so that the marginal social cost of emissions is the same no matter which reduces its emissions. Because they have different abatement costs, however, their marginal cost of abatement curves are not the same. Figure 18.5 shows why emissions fees are preferable to standards in this case. MCA_1 and MCA_2 represent the marginal cost of abatement curves for the two firms. Each firm initially generates 14 units of emissions. Suppose we want to reduce total emissions by 14 units. Figure 18.5 shows that the cheapest way to do this is to have Firm 1 reduce emissions by 6 units and Firm 2 by 8. With these reductions,

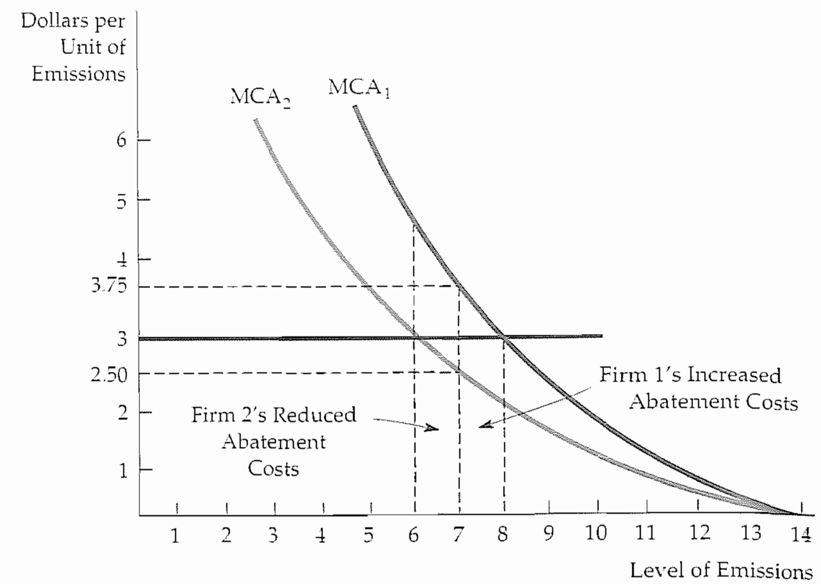

FIGURE 18.5 The Case for Fees

With limited information, a policymaker may be faced with the choice of either a single emissions fee or a single emissions standard for all firms. The fee of \$3 achieves a total emissions level of 14 units more cheaply than a 7-unit-per-firm emissions standard. With the fee, the firm with a lower abatement cost curve (Firm 2) reduces emissions more than the firm with a higher cost curve (Firm 1).

both firms have marginal costs of abatement of \$3. But consider what happens if the regulatory agency asks both firms to reduce emissions by 7 units. In that case the marginal cost of abatement of Firm 1 increases from \$3 to \$3.75, and the marginal cost of abatement of Firm 2 decreases from \$3 to \$2.50. This cannot be cost-minimizing because the second firm can reduce emissions more cheaply than the first. Only when the marginal cost of abatement is equal for both firms will emissions be reduced by 14 units at minimum cost.

Now we can see why a fee (\$3) might be preferable to a standard (7 units). With a \$3 fee, Firm 1 will reduce emissions by 6 units and Firm 2 by 8 units—the efficient outcome. By contrast, with the emissions standard, Firm 1 incurs additional abatement costs given by the green-shaded area between 7 and 8 units of emission. But Firm 2 enjoys reduced abatement costs given by the purple-shaded area between 6 and 7 units. Clearly, the added abatement costs to Firm 1 are larger than the reduced costs to Firm 2. The emissions fee thus achieves the same level of emissions at a lower cost than the equal per-firm emissions standard.

In general, fees can be preferable to standards for several reasons. First, when standards must be applied equally to all firms, fees achieve the same emissions reduction at a lower cost. Second, fees give a firm a strong incentive to install new equipment that would allow it to reduce emissions *even further*. Suppose the standard requires that each firm reduce its emission by 6 units, from 14 to 8. Firm 1 is considering installing new emissions devices that would lower its marginal cost of abatement from MCA$_1$ to MCA$_2$. If the equipment is relatively inexpensive, the firm will install it because it will lower the cost of meeting the standard. However, a \$3 emissions fee would provide a greater incentive for the firm

to reduce emissions. With the fee, not only will the firm's cost of abatement be lower on the first 6 units of reduction, but it will also be cheaper to reduce emissions by 2 more units: The emissions fee is greater than the marginal abatement cost for emissions levels between 6 and 8.

The Case for Standards Now let's examine the case for standards by looking at Figure 18.6. While the marginal social cost curve is very steep, the marginal cost of abatement is relatively flat. The efficient emissions fee is \$8. But suppose that because of limited information a lower fee of \$7 is charged (this fee amounts to a 1/8 or 12.5 percent reduction). Because the MCA curve is flat, the firm's emissions will be increased from 8 to 11 units. This lowers the firm's abatement costs somewhat, but because the MSC curve is steep, there will be substantial additional social costs. The increase in social costs, less the savings in abatement costs, is given by the entire shaded (light and dark) triangle *ABC*.

What happens if a comparable error is made in setting the standard? The efficient standard is 8 units of emissions. But suppose the standard is relaxed by 12.5 percent, from 8 to 9 units. As before, this will lead to an increase in social costs and a decrease in abatement costs. But the net increase in social costs, given by the small triangle *ADE*, is much smaller than before.

This example illustrates the difference between standards and fees. When the marginal social cost curve is relatively steep and the marginal cost of abatement curve relatively flat, the cost of not reducing emissions is high. In such cases, a standard is preferable to a fee. With incomplete information, standards offer

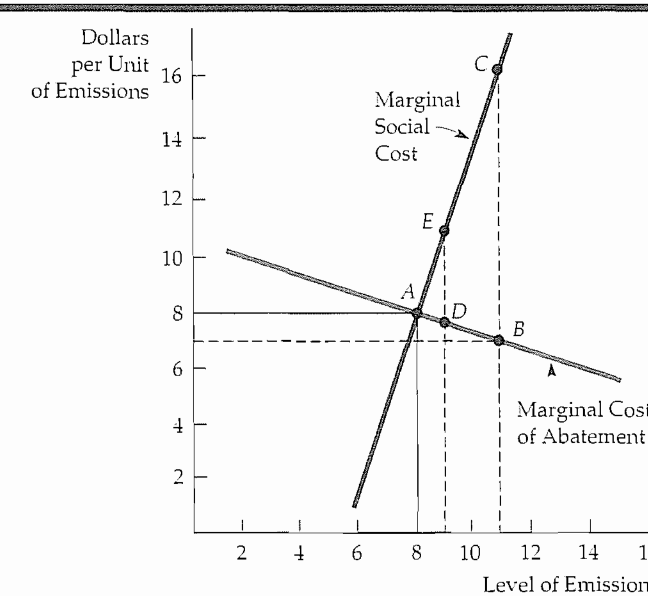

FIGURE 18.6 The Case for Standards

When the government has limited information about the costs and benefits of pollution abatement, either a standard or a fee may be preferable. The standard is preferable when the marginal social cost curve is steep and the marginal abatement cost curve is relatively flat. Here a 12.5 percent error in setting the standard leads to extra social costs of triangle *ADE*. The same percentage error in setting a fee would result in excess costs of *ABC*.

more certainty about emissions levels but leave the costs of abatement uncertain. Fees, on the other hand, offer certainty about the costs of abatement but leave the reduction of emissions levels uncertain. The preferable policy depends, therefore, on the nature of uncertainty and on the shapes of the cost curves.[2]

Transferable Emissions Permits

transferable emissions permits System of marketable permits, allocated among firms, specifying the maximum level of emissions that can be generated.

Suppose we want to reduce emissions efficiently—i.e., to reduce emissions to the point at which the marginal benefit of emissions reduction is equal to the marginal cost of abatement. However, because of uncertainty over the costs and benefits of abatement, we do not want to rely on an emissions fee. We can reach these goals by using **transferable emissions permits.** Under this system, each firm must have permits to generate emissions. Each permit specifies the number of units of emissions the firm is allowed to put out. Any firm that generates emissions that are not allowed by permit is subject to substantial monetary sanctions. Permits are allocated among firms, with the total number of permits chosen to achieve the desired maximum level of emissions. The permits are marketable; they can be bought and sold.

Under the permit system, the firms least able to reduce emissions are those that purchase permits. Thus, suppose the two firms in Figure 18.5 were given permits to emit up to 7 units. Firm 1, facing a relatively high marginal cost of abatement, would pay up to $3.75 to buy a permit for one unit of emissions, but the value of that permit is only $2.50 to Firm 2. Firm 2 should therefore sell its permit to Firm 1 for a price between $2.50 and $3.75.

If there are enough firms and permits, a competitive market for permits will develop. In market equilibrium, the price of a permit equals the marginal cost of abatement for all firms; otherwise a firm will find it advantageous to buy more permits. The level of emissions chosen by the government will be achieved at minimum cost. Those firms with relatively low marginal cost of abatement curves will be reducing emissions the most, and those firms with relatively high marginal cost of abatement curves will be buying more permits and reducing emissions the least.

Marketable emissions permits create a market for externalities. This market approach is appealing because it combines some of the advantageous features of a system of standards with the cost advantages of a fee system. The agency that administers the system determines the total number of permits and therefore the total amount of emissions, just as a system of standards would do. But the marketability of the permits allows pollution abatement to be achieved at minimum cost, just as a system of fees would do.[3]

[2] Our analysis presumes that the emissions fee is levied as a fixed fee per unit of emissions. If the fee is set too low because of limited information, the firm will generate a substantial amount of excess emissions. Suppose, however, that a fixed fee were replaced with a fee schedule designed so that the higher the level of emission, the higher the per-unit fee. In this case, if the fee schedule is set too low, the increasing fee will discourage the firm from generating substantial excess emissions. In general, a variable fee is preferable to a standard if the fee schedule can be designed to match the environmental harm caused by the emissions. In this case, firms know that the payment they make will be approximately equal to the harm that they cause and will *internalize* that harm in making their production decisions. See Louis Kaplow and Steven Shavell, "On the Superiority of Corrective Taxes to Quantity Regulation," NBER Working Paper No. W6251, November 1997.

[3] With limited information and costly monitoring, a marketable permit system is not always ideal. For example, if the total number of permits is chosen incorrectly and the marginal cost of abatement rises sharply for some firms, a permit system could drive those firms out of business by imposing high abatement costs. (This would also be a problem for fees.)

EXAMPLE 18.1 The Costs and Benefits of Reduced Sulfur Dioxide Emissions

The effect of sulfur dioxide pollution on the environment has concerned policymakers for years, but these concerns reached a new height in the 1990s (with a series of amendments to the Clean Air Act) because of the potential adverse effects of acid rain. Acid rain, formed by burning fossil fuels that release sulfur dioxide and nitrogen oxides, threatens property and health throughout the Midwestern and Northeastern United States.

How can we best deal with the sulfur dioxide problem? An important lesson comes from a policy that was pursued several decades ago. In 1968, Philadelphia imposed air quality regulations that limited the maximum allowable sulfur content in fuel oil to 1.0 percent or less. This regulation decreased sulfur dioxide levels in the air substantially—from 0.10 parts per million (ppm) in 1968 to below 0.030 ppm in 1973. Improved air quality led to better human health, less damage to materials, and higher property values. But these improvements had a cost: Industrial, manufacturing, commercial, and residential fuel users had to alter their fuel choices and to install pollution-control equipment to abate pollution. Was the benefit—the reduction in social cost due to abatement—worth the additional abatement cost? A cost-benefit study of reductions in sulfur dioxide emissions provided some answers.[4]

In Philadelphia, the emissions reductions necessitated increased costs of converting from coal and oil to gas to comply with the air-quality regulation. Emissions-control equipment also had to be added to manufacturing processes to ensure that fuels were used efficiently. Figure 18.7 shows the marginal social cost and the marginal cost to the firm of reduced emissions. Note that the marginal abatement cost jumps whenever new capital-intensive pollution-control equipment is needed to improve fuel efficiency.

The benefits of reduced sulfur dioxide emissions can be divided into three parts: (1) reductions in illness and death from diseases like cancer, bronchitis, pneumonia, emphysema, asthma, and the common cold; (2) reductions in materials costs caused by corrosion of metals, stone, and paint; and (3) improvements in visibility and other aesthetic values.

Because benefits are the negative of social costs, we can obtain information about the marginal social cost curve by asking how each of these three types of benefits decreases in value when sulfur dioxide concentrations are increased. For very low concentrations, evidence suggests little health, material, or aesthetic effects. But for moderate concentrations, studies of respiratory diseases, corrosion of materials, and lost visibility suggest that marginal social costs are positive and relatively constant. Thus the marginal social cost curve rises initially and then becomes horizontal.

The efficient level of reduced sulfur dioxide emissions is given by the number of ppm of sulfur dioxide at which the marginal cost of reduced emissions is equal to the marginal social cost. We can see from Figure 18.7 that this level is approximately 0.0275 ppm. The marginal social cost and marginal abatement cost curves intersect at a point where the marginal abatement cost curve is sharply decreasing because of the introduction of expensive desulfurization equipment. Because 0.0275 ppm is slightly below the emissions level achieved by

[4] The study is by Thomas R. Irvin, "A Cost-Benefit Analysis of Sulfur Dioxide Abatement Regulations in Philadelphia," *Business Economics* (September 1977): 12–20.

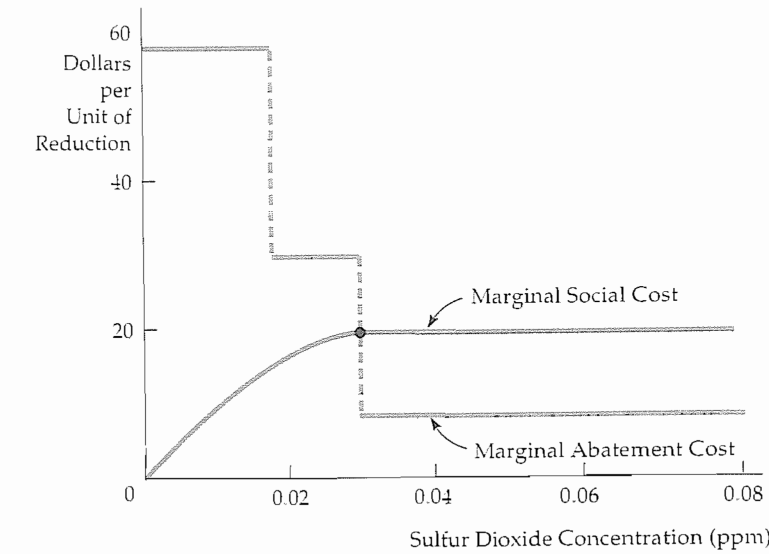

FIGURE 18.7 Sulfur Dioxide Emissions Reductions

The efficient sulfur dioxide concentration equates the marginal abatement cost to the marginal social cost. Here the marginal abatement cost curve is a series of steps, each representing the use of a different abatement technology.

the regulation in 1973, we can conclude that the regulation improved economic efficiency. In fact, given that sulfur dioxide levels were above 0.0275 ppm for most of the period, it appears that the regulations were not stringent enough to achieve the most efficient outcome.

How does the use of standards-based regulations compare to that of tradeable emissions permits? A recent study of the regulation of electric utility sulfur dioxide emissions shows that marketable permits can cut the cost of compliance with a regulatory standard in half.[5] The lower costs are achieved because firms with high abatement costs buy permits that allow emissions, while those with low abatement costs reduce emissions and sell permits.

EXAMPLE 18.2 Emissions Trading and Clean Air

The cost of air pollution control during the 1980s was approximately $18 billion per year.[6] An effective emissions trading system could reduce those costs substantially in the decades to come. The Environmental Protection

[5] Don Fullerton, Shaun P. McDermott, and Jonathan P. Caulkins, "Sulfur Dioxide Compliance of a Regulated Utility," NBER Working Paper No. 5542, April 1996. See also Dallas Burtraw, Alan J. Krupnick, Erin Mansur, David Austin, and Deirdre Farrell, "The Cost and Benefits of Reducing Acid Rain," (Washington: Resources for the Future, September 1997).

[6] See Robert W. Hahn and Gordon L. Hester, "The Market for Bads: EPA's Experience with Emissions Trading," *Regulation* (1987): 48–53; Brian J. McKean, "Evolution of Marketable Permits: The U.S. Experience with Sulfur-Dioxide Allowance Trading," Environmental Protection Agency, December, 1996.

Agency's "bubble" and "offset" programs provided a modest attempt to use a trading system to lower cleanup costs.

A bubble allows an individual firm to adjust its pollution controls for individual sources of pollutants, as long as a *total pollutant limit* for the firm is not exceeded. In theory a bubble could be used to set pollutant limits for many firms or for an entire geographic region; in practice, however, it has been applied to individual firms. The result is, in effect, that "permits" are traded within the firm: if one part of the firm could reduce its emissions, another part would be allowed to emit more. Abatement cost savings associated with the EPA's program of 42 bubbles have been approximately $300 million since 1979.

Under the offset program, new sources of emissions may be located in geographic regions in which air quality standards have not been met, but only if they offset their new emissions by reducing emissions from existing sources at least as much. Offsets can be obtained by internal trading, but external trading among firms is also allowed. Over 2000 offset transactions have occurred since 1976.

Because of their limited natures, bubble and offset programs substantially understate the potential gain from a broad-based emissions trading program. In one study, the cost of achieving an 85-percent reduction in hydrocarbon emissions in all U.S. DuPont plants was estimated under three alternative policies: (1) each source at each plant must reduce emissions by 85 percent; (2) each plant must reduce its overall emissions by 85 percent with only internal trading possible; and (3) total emissions at all plants must be reduced by 85 percent, with both internal and external trading possible.[7] When no trading was allowed, the cost of emissions reduction was $105.7 million. Internal trading reduced the cost of $42.6 million. Allowing for both external and internal trading reduced the cost further to $14.6 million.

Clearly, the potential cost savings from an effective transferable emissions program can be substantial. This may explain why Congress focused on transferable permits as a way of dealing with "acid rain" in the 1990 Clean Air Act. Acid rain can be extremely harmful to people, animals, vegetation, and buildings. The government has authorized a permit system to reduce sulfur dioxide emissions by 10 million tons and nitrogen oxide emissions by 2.5 million tons by the year 2000.

Under the plan, each tradeable permit allows a maximum of one ton of sulfur dioxide to be released into the air. Electric utilities and other polluting entities are allocated permits in proportion to their current level of emissions. Companies can make the capital investments necessary to reduce emissions, perhaps selling excess permits, or they can buy permits and avoid having to make costly emission-reducing investments.

In the early 1990s, economists expected these permits to trade for $300 or more each. In fact, during 1993, prices were below $200, and, as Figure 18.8 shows, by 1996 they fell to below $100. Why? Because reducing sulfur dioxide emissions has turned out to be less costly than anticipated (it had become cheaper to extract low-sulfur coal), and many electric utilities took advantage of this development to reduce emissions.

From a low of $70/ton early in 1996, permit prices began to move upward, reaching about $210/ton in mid-1999. Why the run-up in prices? Because the

[7] M.T. Maloney and Bruce Yandle, "Bubbles and Efficiency: Cleaner Air at Lower Cost," *Regulation* (May/June 1980): 49–52.

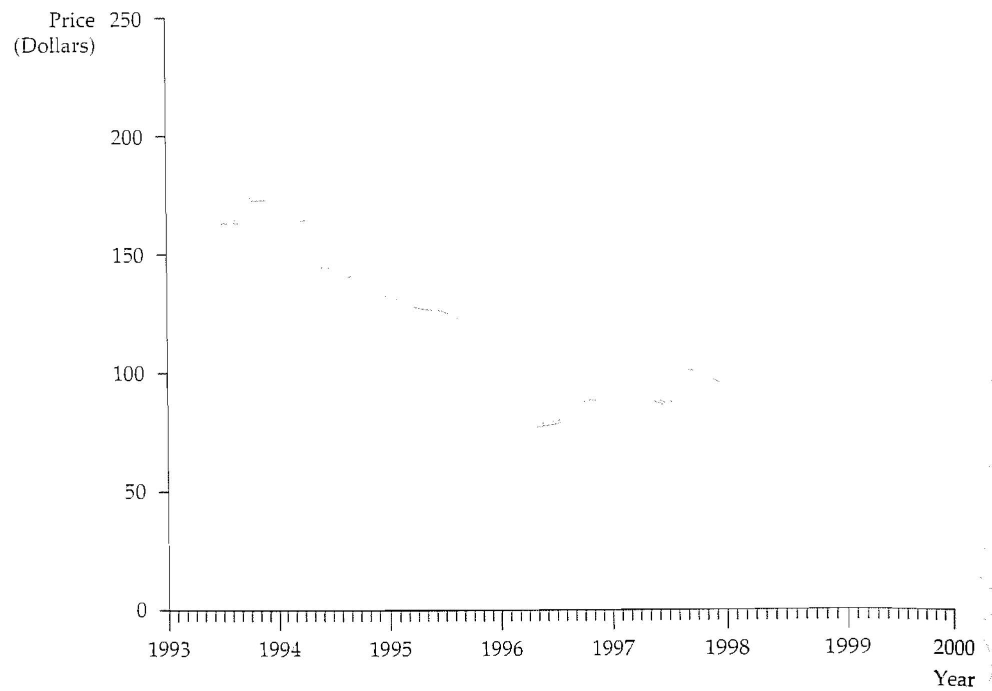

FIGURE 18.8 Price of Tradeable Emissions Permits

The price of tradeable permits for sulfur dioxide emissions declined from 1993 to 1996 due in part to the reduced cost of emissions reductions, and then increased from 1996 to 1999 in response to an increased demand for permits.

price declines in the mid-1990s were an unusual and probably one-time occurrence. Early in the 1990s, many utilities committed to making long-term investments in abatement on the assumption that the price of emissions permits would remain high—higher even than $200/ton. With hindsight, we can see that they were wrong and, as a result, overinvested in abatement. With abatement investments locked in by long-term contracts, the demand for permits fell, and with no decrease in supply of permits, the market allowed permit prices to fall as well. As we moved into the late 1990s and the market adjustment to excess abatement had run its course (through a substantial cutback in abatement), the demand for permits increased, as did permit prices.[8]

Recycling

To the extent that the disposal of waste products involves little or no private cost to either consumers or producers, society will dispose of too much waste material. The overutilization of virgin materials and the underutilization of recycled

[8] We wish to thank Elizabeth Bailey for providing the emissions permits data and for her helpful comments. For a more detailed explanation of permit prices, see A. D. Ellerman, P. L. Joskow, R. Schmalensee, J. P. Montero, and E. M. Bailey, *Markets for Clean Air: The U.S. Acid Rain Program* (M.I.T. Center for Energy and Environmental Policy Research, 1999). For more on tradeable permits generally, go to the EPA Web site at www.epa.gov to find information on the acid rain program.

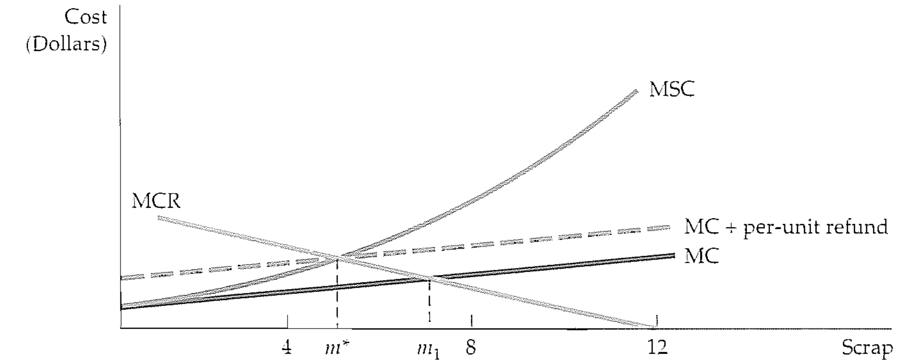

FIGURE 18.9 The Efficient Amount of Recycling

The efficient amount of recycling of scrap material is the amount that equates the marginal social cost of scrap disposal, MSC, to the marginal cost of recycling, MCR. The efficient amount of scrap for disposal m^* is less than the amount that will arise in a private market, m_1.

materials will result in a market failure that may require government intervention. Fortunately, given the appropriate incentive to recycle products, this market failure can be corrected.[9]

To see how recycling incentives can work, consider a typical household's decision with respect to the disposal of glass containers. In many communities, households are charged a fixed annual fee for trash disposal. As a result, these households can dispose of glass and other garbage at very low cost—only the time and effort to put the materials in a trash receptacle.

The low cost of disposal creates a divergence between the private and the social cost of disposal. The marginal private cost of disposal, which is the cost to the household of throwing out the glass, is likely to be constant (independent of the amount of disposal) for low to moderate levels of disposal. It will then increase for large disposal levels involving additional shipping and dump charges. In contrast, the social cost of disposal includes the harm to the environment from littering as well as the injuries caused by sharp glass objects. Marginal social cost is likely to increase, in part because the marginal private cost is increasing and in part because the environmental and aesthetic costs of littering are likely to increase sharply as the level of disposal increases.

Both cost curves are shown in Figure 18.9. The horizontal axis measures, from left to right, the amount of scrap material m that the household disposes, up to a maximum of 12 pounds per week. Consequently, the amount recycled can be read from right to left. As the amount of scrap disposal increases, the marginal private cost, MC, increases, but at a much lower rate than the marginal social cost MSC.

Recycling of containers can be accomplished by a municipality or a private firm that arranges for collection, consolidation, and processing of materials. The marginal cost of recycling is likely to increase as the amount of recycling grows, in part because collection, separation, and cleaning costs grow at an increasing rate. The marginal cost of recycling curve, MCR, in Figure 18.9 is best read from

[9] Even without market intervention, some recycling will occur if the price of virgin material is sufficiently high. For example, recall from Chapter 2 that when the price of copper is high, there is more recycling of scrap copper.

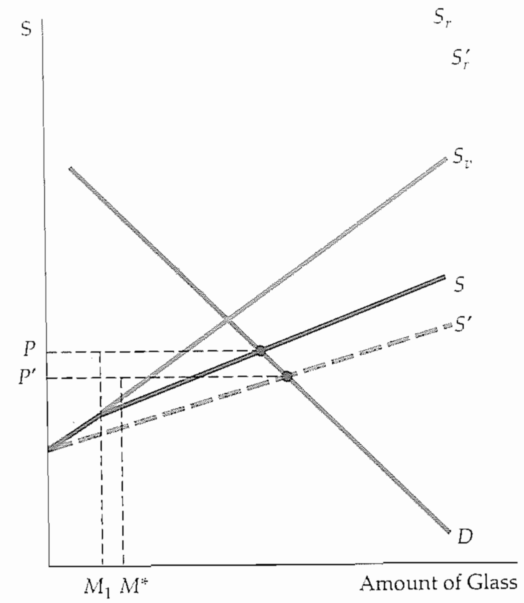

FIGURE 18.10 Refundable Deposits

Initially, equilibrium in the market for glass containers involves a price P and a supply of recycled glass M_1. By raising the relative cost of disposal and encouraging recycling, the refundable deposit increases the supply of recycled glass from S_r to S_r' and the aggregate supply of glass from S to S'. The price of glass then falls to P', the quantity of recycled glass increases to M^*, and the amount of disposed glass decreases.

right to left. Thus, when there are 12 pounds of disposed material, there is no recycling; the marginal cost is zero. As the amount of disposal decreases, the amount of recycling increases; the marginal cost of recycling increases.

The efficient amount of recycling occurs at the point at which the marginal cost of recycling, MCR, is equal to the marginal *social* cost of disposal, MSC. As Figure 18.9 shows, the efficient amount of scrap for disposal m^* is less than the amount that will arise in a private market, m_1.

Why not utilize a disposal fee, a disposal standard, or even transferable disposal permits to resolve this externality? Any of these policies can help in theory, but they are not easy to put into practice and are rarely used. For example, a disposal fee is difficult to implement because it would be very costly for a community to sort through trash to separate and then to collect glass materials. Pricing and billing for scrap disposal would also be expensive, because the weight and composition of materials would affect the social cost of the scrap and, therefore, the appropriate price to be charged.

Refundable Deposits One policy solution that has been used with some success to encourage recycling is the *refundable deposit*.[10] Under a refundable deposit system an initial deposit is paid to the store owner when the glass container product is purchased. The deposit is refunded if and when the container is returned to the

[10] See Frank Ackerman, *Why Do We Recycle: Markets, Values, and Public Policy* (Washington: Island Press, 1997), for a general discussion of recycling.

store or to a recycling center. Refundable deposits create a desirable incentive: The per-unit refund can be chosen so that households (or firms) recycle more material.

From an individual's point of view, the refundable deposit creates an additional private cost of disposal: the opportunity cost of failing to obtain a refund. As shown in Figure 18.9, with the higher cost of disposal, the individual will reduce disposal and increase recycling to the optimal social level m^*.

A similar analysis applies at the industry level. Figure 18.10 shows a downward-sloping market demand for glass containers, D. The supply of virgin glass containers is given by S_v and the supply of recycled glass by S_r. The market supply S is the horizontal sum of these two curves. As a result, the market price of glass is P and the equilibrium supply of recycled glass is M_1.

By raising the relative cost of disposal and encouraging recycling, the refundable deposit increases the supply of recycled glass from S_r to S_r', the aggregate supply increases from S to S', and the price of glass falls to P'. As a result, the quantity of recycled glass increases to M^*, which means a decrease in the amount of disposed glass.

The refundable deposit scheme has another advantage: a market for recycled products is created. In many communities, public or private firms as well as private individuals specialize in collecting and returning recyclable materials. As this market becomes larger and more efficient, the demand for recycled rather than virgin materials increases, therefore increasing the benefit to the environment.

EXAMPLE 18.3 Regulating Municipal Solid Wastes

By 1990, the average resident of Los Angeles was generating about 6.4 pounds of solid waste per day, and residents of other large American cities were not far behind. By contrast, residents of Tokyo, Paris, Hong Kong, and Rome generated 3 pounds, 2.4 pounds, 1.9 pounds, and 1.5 pounds, respectively.[11] Some of these differences are due to variations in consumption levels, but most are due to the efforts that many other countries have made to encourage recycling. In the United States, only about 25 percent of aluminum, 23 percent of paper, and 8.5 percent of glass scrap are recycled.

A number of policy proposals have been introduced to encourage recycling in the United States. The first is the refundable deposit described above. A second is a *curbside charge*, in which communities charge individuals a fee for refuse disposal that is proportional to the weight (or the volume) of the refuse. To encourage separation of recyclable materials, all separable glass materials are collected for free. Curbside charges encourage recycling, but they fail to discourage consumption of products that might require recycling.

A third alternative is to require the *mandatory separation* of recyclable materials such as glass. Random spot checks with substantial penalties for violations are required to make the system effective. Mandatory separation is perhaps the least desirable of the three alternatives, not only because it is difficult to implement, but also because individuals, if the cost of separation is sufficiently high, may be encouraged to shift to alternative containers such as plastic, which are environmentally damaging and cannot readily be recycled.

[11] This example is based on Peter S. Menell, "Beyond the Throwaway Society: An Incentive Approach to Regulating Municipal Solid Waste," *Ecology Law Quarterly* (1990): 655–739. See also Marie Lynn Miranda et al., "Unit Pricing for Residential Municipal Solid Waste: An Assessment of the Literature," US Environmental Protection Agency, March 1996.

The potential effectiveness of these three policies is illustrated by an analysis that focused on the mix between glass and plastic. Consumers were assumed to have varying preferences, with half preferring glass and half preferring plastic, for products that are otherwise identical in price, quantity, and quality. Without any incentive to recycle, a 50-50 division between glass and plastic would result. From a social perspective, however, greater use of recyclable glass would be preferred.

Mandatory separation fails as a policy in this case: The cost of separation is so high that the percentage of glass container materials purchased actually falls to 40 percent. A curbside charge does much better: It leads to a 72.5 percent use of recyclable glass. Finally, a refundable deposit system does best, with 78.9 percent of consumers purchasing recyclable glass containers.

A recent case in Perkasie, Pennsylvania, shows that recycling programs can indeed be effective. Prior to implementation of a program combining all three economic incentives just described, the total amount of unseparated solid waste was 2573 tons per year. When the program was implemented, this amount fell to within 1038 tons—a 59-percent reduction. As a result, the town saved $90,000 per year in disposal costs.

18.3 Externalities and Property Rights

We have seen how government regulation can deal with the inefficiencies that arise from externalities. Emissions fees and transferable emissions permits work because they change a firm's incentives, forcing it to take into account the external costs that it imposes. But government regulation is not the only way to deal with externalities. In this section we show that in some circumstances inefficiencies can be eliminated through private bargaining among the affected parties, or by a legal system in which parties can sue to recover the damages they suffer.

Property Rights

property rights Legal rules stating what people or firms may do with their property.

Property rights are the legal rules that describe what people or firms may do with their property. When people have property rights to land, for example, they may build on it or sell it and are protected from interference by others.

To see why property rights are important, let's return to our example of the firm that dumps effluent into the river. We assumed both that it had a property right to use the river to dispose of its waste and that the fishermen did not have a property right to "effluent-free" water. As a result, the firm had no incentive to include the cost of effluent in its production calculations. In other words, the firm *externalized* the costs generated by the effluent. But, suppose the fishermen owned the rivers, i.e., had a property right to clean water. In that case, they could demand that the firm pay them for the right to dump effluent. The firm would either cease production or pay the costs associated with the effluent. These costs would be *internalized* and an efficient allocation of resources achieved.

Bargaining and Economic Efficiency

Economic efficiency can be achieved without government intervention when the externality affects relatively few parties and when property rights are well specified. To see how this might arise, let's consider a numerical version of the

TABLE 18.1 Profits under Alternative Emissions Choices (Daily)

	FACTORY'S PROFIT ($)	FISHERMEN'S PROFIT ($)	TOTAL PROFIT ($)
No filter, no treatment plant	500	100	600
Filter, no treatment plant	300	500	800
No filter, treatment plant	500	200	700
Filter, treatment plant	300	300	600

effluent example. Suppose the steel factory's effluent reduces the fishermen's profit. As Table 18.1 shows, the factory can install a filter system to reduce its effluent, or the fishermen can pay for the installation of a water treatment plant.[12]

The efficient solution maximizes the joint profit of the factory and the fishermen. Maximization occurs when the factory installs a filter and the fishermen do not build a treatment plant. Let's see how alternative property rights lead these two parties to negotiate different solutions.

Suppose the factory has the property right to dump effluent into the river. Initially, the fishermen's profit is $100 and the factory's $500. By installing a treatment plant, the fishermen can increase their profit to $200, whereby the joint profit, without cooperation, is $700 ($500 + $200). Moreover, the fishermen are willing to pay the factory up to $300 to install a filter—the difference between the $500 profit with a filter and the $200 profit without cooperation. Because the factory loses only $200 in profit by installing a filter, it will be willing to do so because it is more than compensated for its loss. The gain to both parties by cooperating is equal to $100 in this case: the $300 gain to the fishermen less the $200 cost of a filter.

Suppose the factory and the fishermen agree to split this gain equally by having the fishermen pay the factory $250 to install the filter. As Table 18.2 shows, this bargaining solution achieves the efficient outcome. Under the column "Right to Dump," we see that without cooperation, the fishermen earn a profit of $200 and the factory $500. With cooperation, the profit of both increases by $50.

TABLE 18.2 Bargaining with Alternative Property Rights

	RIGHT TO DUMP ($)	RIGHT TO CLEAN WATER ($)
No cooperation		
Profit of factory	500	300
Profit of fishermen	200	500
Cooperation		
Profit of factory	550	300
Profit of fishermen	250	500

[12] For a more extensive discussion of a variant of this example, see Robert Cooter and Thomas Ulen, *Law and Economics* (Glenview, IL: Scott-Foresman, 1997), ch. 4.

Now suppose the fishermen are given the property right to clean water, which requires the factory to install the filter. The factory earns a profit of $300 and the fishermen $500. Because neither party can be made better off by bargaining, the original outcome is efficient.

This analysis applies to all situations in which property rights are well specified. *When parties can bargain without cost and to their mutual advantage, the resulting outcome will be efficient, regardless of how the property rights are specified.* The italicized proposition is called the **Coase theorem**, in honor of Ronald Coase, who did much to develop it.[13]

Coase theorem Principle that when parties can bargain without cost and to their mutual advantage, the resulting outcome will be efficient regardless of how property rights are specified.

Costly Bargaining—The Role of Strategic Behavior

Bargaining can be time consuming and costly, especially when property rights are not clearly specified. In that case neither party is sure how hard to bargain before the other party will agree to a settlement. In our example, both parties knew that the bargaining process had to settle on a payment between $200 and $300. If the parties were unsure of the property rights, however, the fishermen might be willing to pay only $100, and the bargaining process would break down.

Bargaining can also break down even when communication and monitoring are costless if both parties believe they can obtain larger gains. One party makes a demand for a large share and refuses to bargain, assuming incorrectly that the other party will eventually concede. This *strategic behavior* can lead to an inefficient, noncooperative outcome. Suppose the factory has the right to emit effluent and claims that it will not install a filter unless it receives $300—its final offer. The fishermen however, offer to pay $250 believing that eventually the factory will agree to the "fair" solution. In this situation, an agreement may never be reached, especially if one or both parties want to earn a reputation for tough bargaining.

In §13.8, we explain how firms can use strategic moves to change their bargaining positions.

A Legal Solution—Suing for Damages

In many situations involving externalities, a party that is harmed (the victim) by another has the legal right to sue. If successful, the victim can recover monetary damages equal to the harm it has been caused. A suit for damages is different from an emissions or effluent fee because the victim, not the government, is paid.

To see how the potential for a lawsuit can lead to an efficient outcome, let's reexamine our fishermen–factory example. Suppose first that the fishermen are given the right to clean water. The factory, in other words, is responsible for harm to the fishermen *if* it does not install a filter. The harm to the fishermen in this case is $400 (the difference between the profit that the fishermen make when there is no effluent [$500] and their profit when there is effluent [$100]). The factory has the following options:

1. Do not install filter, pay damages:	Profit = $100 ($500 − $400)
2. Install filter, avoid damages:	Profit = $300 ($500 − $200)

The factory will find it advantageous to install a filter, which is substantially cheaper than paying damages, and the efficient outcome will be achieved.

[13] Ronald Coase, "The Problem of Social Cost," *Journal of Law and Economics* 3 (1960): 1–44.

An efficient outcome (with a different division of profits) will also be achieved if the factory is given the property right to emit effluent. Under the law, the fishermen would have the legal right to require the factory to install the filter, but they would have to pay the factory for its $200 lost profit (not for the cost of the filter). This leaves the fishermen with three options:

1. Put in a treatment plant:	Profit = $200
2. Have factory put in a filter, but pay damages:	Profit = $300 ($500 − $200)
3. Do not put in treatment plant or require a filter:	Profit = $100

The fishermen earn the highest profit if they take the second option. They will thus require the factory to put in a filter but compensate it $200 for its lost profit. Just as in the situation in which the fishermen had the right to clean water, this outcome is efficient because the filter has been installed. Note, however, that the $300 profit is substantially less than the $500 profit that the fishermen get when they have a right to clean water.

This example shows that a suit for damages eliminates the need for bargaining because it specifies the consequences of the choices that the parties must make. Giving the party that is harmed the right to recover damages from the injuring party ensures an efficient outcome. (When information is imperfect, however, suing for damages may lead to inefficient outcomes.)

EXAMPLE 18.4 The Coase Theorem at Work

As a September 1987 cooperative agreement between New York City and New Jersey illustrates, the Coase theorem applies to governments as well as to people and organizations.

For many years, garbage spilling from waterfront trash facilities from New York harbor had adversely affected the quality of water along the New Jersey shore and occasionally littered the beaches. One of the worst instances occurred in August 1987, when more than 200 tons of garbage formed a 50-mile-long slick off the New Jersey shore.

New Jersey had the right to clean beaches and could have sued New York City to recover damages associated with garbage spills. New Jersey could have also asked the court to grant an injunction requiring New York City to stop using its trash facilities until the problem was removed.

But New Jersey wanted cleaner beaches, not simply the recovery of damages. And New York wanted to be able to operate its trash facility. As a result, there was room for mutually beneficial exchange. After two weeks of negotiations, New York and New Jersey reached a settlement. New Jersey agreed not to bring a lawsuit against the city. New York City agreed to use special boats and other flotation devices to contain spills that might arise from Staten Island and Brooklyn. It also agreed to form a monitoring team to survey all trash facilities and shut down those failing to comply. At the same time, New Jersey officials were allowed unlimited access to New York City trash facilities to monitor the program's effectiveness.

18.4 Common Property Resources

common property resource Resource to which anyone has free access.

Occasionally externalities arise when resources can be used without payment. **Common property resources** are those to which anyone has free access. As a result, they are likely to be overutilized. Air and water are the two most common examples. Others include fish, animal populations, mineral exploration, and extraction. Let's look at some of the inefficiencies that can occur when resources are common property rather than privately owned.

Consider a large lake with trout to which an unlimited number of fishermen have access. Each fisherman fishes up to the point at which the marginal revenue from fishing (or the marginal value, if fishing is for sport instead of profit) is equal to the cost. But the lake is a common property resource, and no fisherman has the incentive to take into account how his fishing affects the opportunities of others. As a result, the fisherman's private cost understates the true cost to society because more fishing reduces the stock of fish, making less available for others. This leads to an inefficiency—too many fish are caught.

Figure 18.11 illustrates this situation. Suppose that because the fish catch is sufficiently small relative to demand, fishermen take the price of fish as given. Suppose also that someone can control the number of fishermen with access to the lake. The efficient level of fish per month F^* is determined at the point at which the marginal benefit from fish caught is equal to the marginal social cost. The marginal benefit is the price taken from the demand curve. The marginal social cost is shown in the diagram to include not only the private operating costs but also the social cost of depleting the stock of fish.

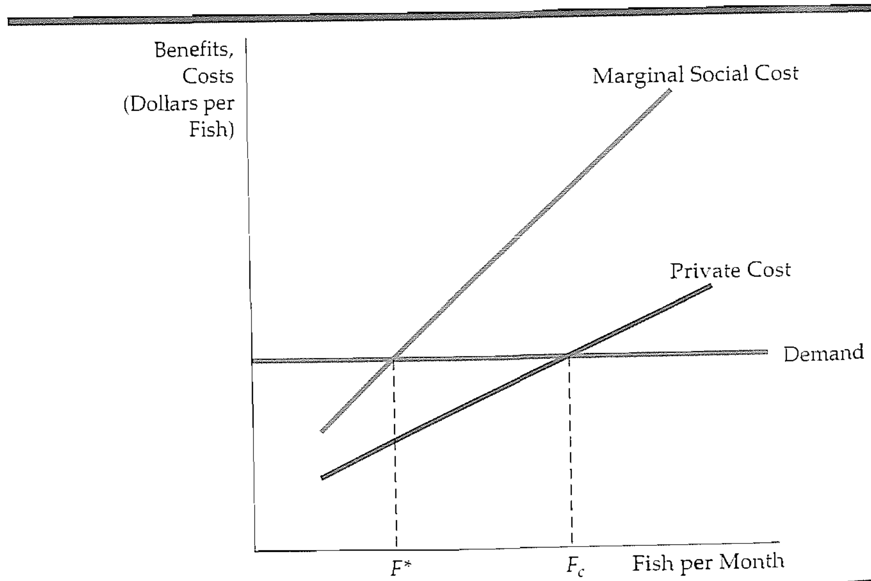

FIGURE 18.11 Common Property Resources

When a common property resource, such as a fishery, is accessible to all, the resource is used up to the point F_c at which the private cost is equal to the additional revenue generated. This usage exceeds the efficient level F^* at which the marginal social cost of using the resource is equal to the marginal benefit (as given by the demand curve).

Now compare the efficient outcome with what happens when the lake is common property. In this case the marginal external costs are not taken into account, and each fisherman fishes until there is no longer any profit to be made. When only F^* fish are caught, the revenue from fishing is greater than the cost, and there is a profit to be earned by fishing more. Entry into the fishing business occurs until the point at which the price is equal to the marginal cost, point F_c in Figure 18.11. At F_c, however, too many fish will be caught.

There is a relatively simple solution to the common property resource problem—let a single owner manage the resource. The owner will set a fee for use of the resource that is equal to the marginal cost of depleting the stock of fish. Facing the payment of this fee, fishermen in the aggregate will no longer find it profitable to catch more than F^* fish. Unfortunately, most common property resources are vast, because single ownership is not always practical. In such cases government ownership or direct government regulation may be needed.

EXAMPLE 18.5 Crawfish Fishing in Louisiana

In recent years, crawfish has become a popular restaurant item. In 1950, for example, the annual crawfish harvest in the Atchafalaya River basin in Louisiana was just over 1 million pounds. By 1995, it had grown to over 30 million pounds. Because most crawfish grow in ponds to which fishermen have unlimited access, a common property resource problem has arisen: Too many crawfish have been trapped, causing the crawfish population to fall far below the efficient level.[14]

How serious is the problem? Specifically, what is the social cost of unlimited access to fishermen? The answer can be found by estimating the private cost of trapping crawfish, the marginal social cost, and the demand for crawfish. Figure 18.12 shows portions of the relevant curves. Private cost is upward-sloping: As the catch increases, so does the additional effort that must be made to obtain it. The demand curve is downward sloping but elastic because other shellfish are close substitutes.

We can find the efficient crawfish catch graphically or algebraically. To do so, let F represent the catch of crawfish in millions of pounds per year (shown on the horizontal axis), and let C represent cost in dollars per pound (shown on the vertical axis). In the region where the various curves intersect, the three curves in the graph are as follows:

Demand:	$C = 0.401 - 0.0064F$
Marginal social cost:	$C = -5.645 + 0.6509F$
Private cost:	$C = -0.357 + 0.0573F$

The efficient crawfish catch of 9.2 million pounds, which equates demand to marginal social cost, is shown as the intersection of the two curves. The actual catch, 11.9 million pounds, is determined by equating demand to private cost and is shown as the intersection of those two curves. The yellow-shaded triangle in the figure measures the social cost of free access. This figure represents the excess of social cost above the private benefit of fishing summed from the efficient level (where demand is equal to marginal social cost) to the actual level

[14] This example is based on Frederick W. Bell, "Mitigating the Tragedy of the Commons," *Southern Economic Journal* 52 (1986): 653–64.

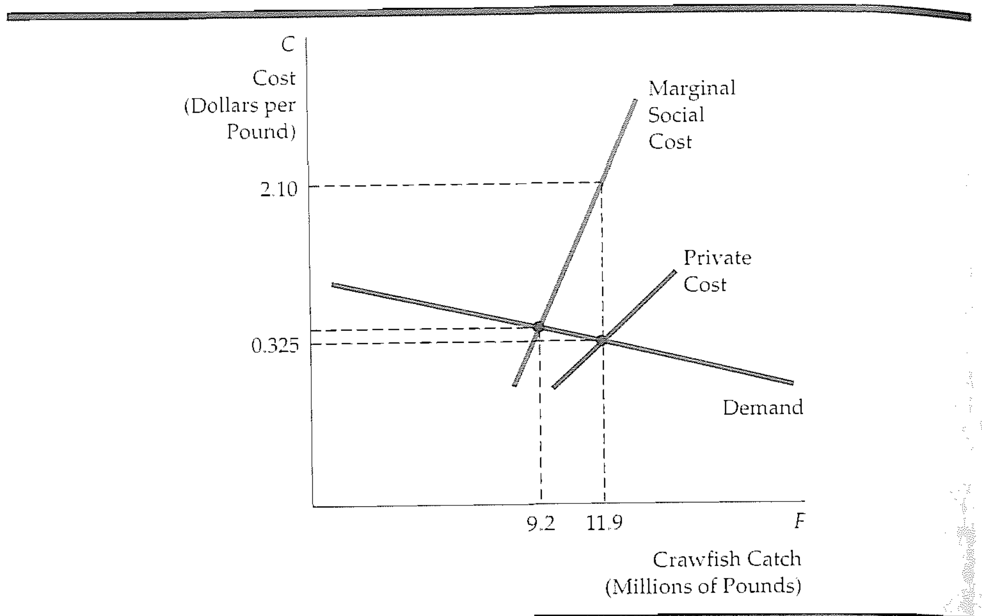

FIGURE 18.12 Crawfish as a Common Property Resource

Because crawfish are bred in ponds to which fishermen have unlimited access, they are a common property resource. The efficient level of fishing occurs when the marginal benefit is equal to the marginal social cost. However, the actual level of fishing occurs at the point at which the price for crawfish is equal to the private cost of fishing. The shaded area represents the social cost of the common property resource.

(where demand is equal to private cost). In this case, the social cost is approximated by the area of a triangle with a base of 2.7 million pounds (11.9 − 9.2) and a height of \$1.775 (\$2.10 − \$0.325), or \$2,396,000. Note that by regulating the ponds—limiting either access or the size of the catch—this social cost could be avoided.

18.5 Public Goods

We have seen that externalities, including common property resources, create market inefficiencies that sometimes warrant government regulation. When, if ever, should governments replace private firms as the producer of goods and services? In this section we describe a set of conditions under which the private market either may not provide a good at all or may not price it properly once it is available.

public good Nonexclusive and nonrival good: the marginal cost of provision to an additional consumer is zero and people cannot be excluded from consuming it.

nonrival good Good for which the marginal cost of its provision to an additional consumer is zero.

Nonrival Goods As we saw in Chapter 16, **public goods** have two characteristics: They are *nonrival* and *nonexclusive*. A good is **nonrival** if for any given level of production, the marginal cost of providing it to an additional consumer is zero. For most goods that are provided privately, the marginal cost of producing more of the good is positive. But for some goods, additional consumers do not add to cost. Consider the use of a highway during a period of low traffic volume. Because the highway already exists and there is no congestion, the additional cost of driving on it is zero. Or consider the use of a lighthouse by a ship. Once the lighthouse is built and functioning, its use by an additional ship adds nothing to its running costs. Finally, consider public television. Clearly, the cost of one more viewer is zero.

Most goods are rival in consumption. For example, when you buy furniture, you have ruled out the possibility that someone else can buy it. Goods that are rival must be allocated among individuals. Goods that are nonrival can be made available to everyone without affecting any individual's opportunity for consuming them.

nonexclusive goods Goods that people cannot be excluded from consuming, so that it is difficult or impossible to charge for their use.

Nonexclusive Goods A good is **nonexclusive** if people cannot be excluded from consuming it. As a consequence, it is difficult or impossible to charge people for using nonexclusive goods; the goods can be enjoyed without direct payment. One example of a nonexclusive good is national defense. Once a nation has provided for its national defense, all citizens enjoy its benefits. A lighthouse and public television are also examples of nonexclusive goods.

Nonexclusive goods need not be national in character. If a state or city eradicates an agricultural pest, all farmers and consumers benefit. It would be virtually impossible to exclude a particular farmer from the benefits of the program. Automobiles are exclusive (as well as rival). If a dealer sells a new car to one consumer, then the dealer has excluded other individuals from buying the car.

Some goods are exclusive but nonrival. For example, in periods of low traffic, travel on a bridge is nonrival because an additional car on the bridge does not lower the speed of other cars. But bridge travel is exclusive because bridge authorities can keep people from using it. A television signal is another example. Once a signal is broadcast, the marginal cost of making the broadcast available to another user is zero, so the good is nonrival. But broadcast signals can be made exclusive by scrambling the signals and charging for the codes that unscramble them.

Some goods are nonexclusive but rival. An ocean or large lake is nonexclusive, but fishing is rival because it imposes costs on others: the more fish caught, the fewer fish available to others. Air is nonexclusive and often nonrival; but it can be rival if the emissions of one firm adversely affect the quality of the air and the ability of others to enjoy it.

Public goods, which are both nonrival and nonexclusive, provide benefits to people at zero marginal cost, and no one can be excluded from enjoying them. The classic example of a public good is national defense. Defense is nonexclusive, as we have seen, but it is also nonrival because the marginal cost of providing defense to an additional person is zero. The lighthouse is also a public good because it is nonrival and nonexclusive; in other words, it would be difficult to charge ships for the benefits they receive from it.[15]

The list of public goods is much smaller than the list of goods that governments provide. Many publicly provided goods are either rival in consumption, exclusive, or both. For example, high school education is rival in consumption. There is a positive marginal cost of providing education to one more child

[15] Lighthouses need not be provided by the government. See Ronald Coase, "The Lighthouse in Economics," *Journal of Law and Economics* 17 (1974): 357–76, for a description of how lighthouses were privately funded in nineteenth-century England.

because other children get less attention as class sizes increase. Likewise, charging tuition can exclude some children from enjoying education. Public education is provided by local government because it entails positive externalities, not because it is a public good.

Finally, consider the management of a national park. Part of the public can be excluded from using the park by raising entrance and camping fees. Use of the park is also rival: because of crowded conditions, the entrance of an additional car into a park can reduce the benefits that others receive from it.

Efficiency and Public Goods

The efficient level of provision of a private good is determined by comparing the marginal benefit of an additional unit to the marginal cost of producing the unit. Efficiency is achieved when the marginal benefit and the marginal cost are equal. The same principle applies to public goods, but the analysis is different. With private goods, the marginal benefit is measured by the benefit the consumer receives. With a public good, we must ask how much each person values an additional unit of output. The marginal benefit is obtained by adding these values for *all* people who enjoy the good. To determine the efficient level of provision of a public good, we must then equate the sum of these marginal benefits to the marginal cost of production.

Figure 18.13 illustrates the efficient level of producing a public good. D_1 represents the demand for the public good by one consumer, and D_2 the demand by a second consumer. Each demand curve tells us the marginal benefit that the consumer

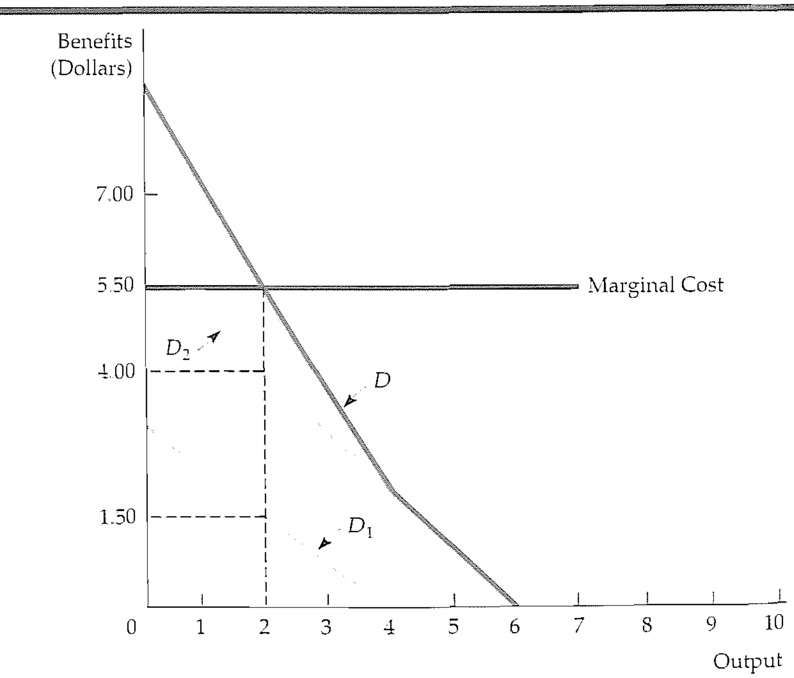

FIGURE 18.13 Efficient Public Good Provision

When a good is nonrival, the social marginal benefit of consumption, given by the demand curve D, is determined by vertically summing the individual demand curves for the good, D_1 and D_2. At the efficient level of output, the demand and the marginal cost curves intersect.

gets from consuming every level of output. For example, when there are 2 units of the public good, the first consumer is willing to pay $1.50 for the good, and $1.50 is the marginal benefit. Similarly, the second consumer has a marginal benefit of $4.00.

To calculate the sum of the marginal benefits to *both* people, we must add each of the demand curves *vertically*. For example, when the output is 2 units, we add the marginal benefit of $1.50 to the marginal benefit of $4.00 to obtain a marginal social benefit of $5.50. When this sum is calculated for every level of public output, we obtain the aggregate demand curve for the public good D.

In §4.3, we show that a market demand curve can be obtained by summing individual demand curves horizontally.

The efficient amount of output is the one at which the marginal benefit to society is equal to the marginal cost. This occurs at the intersection of the demand and the marginal cost curves. In our example, because the marginal cost of production is $5.50, 2 is the efficient output level.

To see why 2 is efficient, note what happens if only 1 unit of output is provided: Although the marginal cost remains at $5.50, the marginal benefit is approximately $7.00. Because the marginal benefit is greater than the marginal cost, too little of the good has been provided. Similarly, suppose 3 units of the public good have been produced. Now the marginal benefit of approximately $4.00 is less than the marginal cost of $5.50; too much of the good has been provided. Only when the marginal social benefit is equal to the marginal cost is the public good provided efficiently.[16]

Public Goods and Market Failure

Suppose you are considering providing a mosquito abatement program for your community. You know that the program is worth more to the community than the $50,000 it will cost. Can you make a profit by providing the program privately? You would break even if you assessed a $5.00 fee to each of the 10,000 households in your community. But you cannot force them to pay the fee, let alone devise a system in which those households that value mosquito abatement most highly pay the highest fees.

Unfortunately, mosquito abatement is nonexclusive: There is no way to provide the service without benefiting everyone. As a result, households have no incentive to pay what the program really is worth to them. People can act as **free riders**, understanding the value of the program so that they can enjoy its benefit without paying for it.

free rider Consumer or producer who does not pay for a nonexclusive good in the expectation that others will.

With public goods, the presence of free riders makes it difficult or impossible for markets to provide goods efficiently. Perhaps if few people were involved and the program were relatively inexpensive, all households might agree voluntarily to share costs. However, when many households are involved, voluntary private arrangements are usually ineffective. The public good must therefore be subsidized or provided by governments if it is to be produced efficiently.

EXAMPLE 18.6 The Demand for Clean Air

In Example 4.5, we used the demand curve for clean air to calculate the benefits of a cleaner environment. Now let's examine the public-good characteristics of clean air. Many factors, including the weather, driving patterns, and

[16] We have shown that nonexclusive, nonrival goods are inefficiently provided. A similar argument would apply to nonrival but exclusive goods.

industrial emissions determine a region's air quality. Any effort to clean up the air will generally improve air quality throughout the region. As a result, clean air is nonexclusive: it is difficult to stop any one person from enjoying it. Clean air is also nonrival: My enjoyment does not inhibit yours.

Because clean air is a public good, there is no market and no observable price at which people are willing to trade clean air for other commodities. Fortunately, we can infer people's willingness to pay for clean air from the housing market—households will pay more for a home located in an area with good air quality than for an otherwise identical home in an area with poor air quality.

Let's look at the estimates of the demand for clean air obtained from a statistical analysis of housing data for the Boston metropolitan area.[17] The analysis correlates housing prices with the quality of air and other characteristics of the houses and their neighborhoods. Figure 18.14 shows three demand curves in which the value put on clean air depends on the level of nitrogen oxides and on income. The horizontal axis measures the level of air pollution in terms of parts per hundred million (pphm) of nitrogen oxides in the air. The vertical axis measures each household's willingness to pay for a one-part-per-hundred million reduction in the nitrogen oxide level.

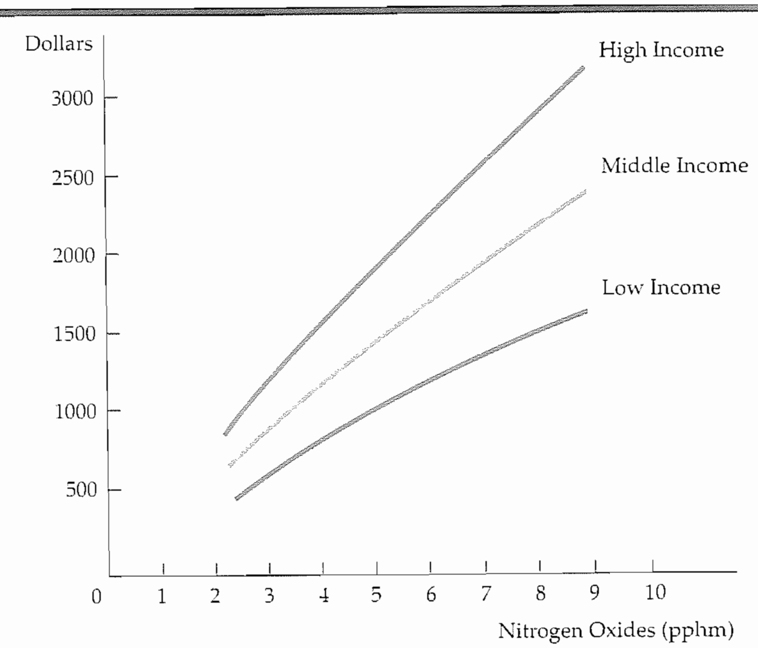

FIGURE 18.14 The Demand for Clean Air

The three curves describe the willingness to pay for clean air (a reduction in the level of nitrogen oxides) for each of three different households (low income, middle income, and high income). In general, higher-income households have greater demands for clean air than lower-income households. Moreover, each household is less willing to pay for clean air as the level of air quality increases.

[17] David Harrison, Jr., and Daniel L. Rubinfeld, "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management* 5 (1978): 81–102.

The demand curves are upward sloping because we are measuring pollution rather than clean air on the horizontal axis. As we would expect, the cleaner the air, the lower the willingness to pay for more of the good. These differences in the willingness to pay for clean air vary substantially. In Boston, for example, nitrogen oxide levels ranged from 3 to 9 pphm. A middle-income household would be willing to pay $800 for a 1 pphm reduction in nitrogen oxide levels when the level is 3 pphm, but the figure would jump to $2200 for a 1 pphm reduction when the level is 9 pphm.

Note that higher-income households are willing to pay more than lower-income households to obtain a small improvement in air quality. At low nitrogen oxide levels (3 pphm), the differential between low- and middle-income households is only $200, but it increases to about $700 at high levels (9 pphm).

With the quantitative information about the demand for clean air and separate estimates of the costs of improving air quality, we can determine whether the benefits of environmental regulations outweigh the costs. A study by the National Academy of Sciences of regulations on automobile emissions did just this. The study found that controls would lower the level of pollutants, such as nitrogen oxides, by approximately 10 percent. The benefit to all residents of the United States of this 10-percent improvement was calculated to be approximately $2 billion. The study also estimated that it would cost somewhat less than $2 billion to install pollution control equipment in automobiles to meet emissions standards. The study concluded, therefore, that the benefits of the regulations did outweigh the costs.

18.6 Private Preferences for Public Goods

Government production of a public good is advantageous because the government can assess taxes or fees to pay for it. But how can government determine how *much* of a public good to provide when the free rider problem gives people an incentive to misrepresent their preferences? In this section we discuss one mechanism for determining private preferences for government-produced goods.

Voting is commonly used to decide allocation questions. For example, people vote directly on some local budget issues and elect legislators who vote on others. Many state and local referenda are based on *majority-rule voting:* Each person has one vote, and the candidate or the issue that receives more than 50 percent of the votes wins. Let's see how majority-rule voting determines the provision of public education. Figure 18.15 describes the preferences for spending on education (on a per-pupil basis) of three citizens who are representative of three interest groups in the school district.

Curve W_1 gives the first citizen's willingness to pay for education, minus any required tax payments. The willingness to pay for each spending level is the maximum amount of money the citizen will pay to enjoy that spending level rather than no spending at all.[18] In general, the benefit from increased spending on education increases as spending increases. But the tax payments required to

[18] In other words, the willingness to pay measures the consumer surplus that the citizen enjoys when a particular level of spending is chosen

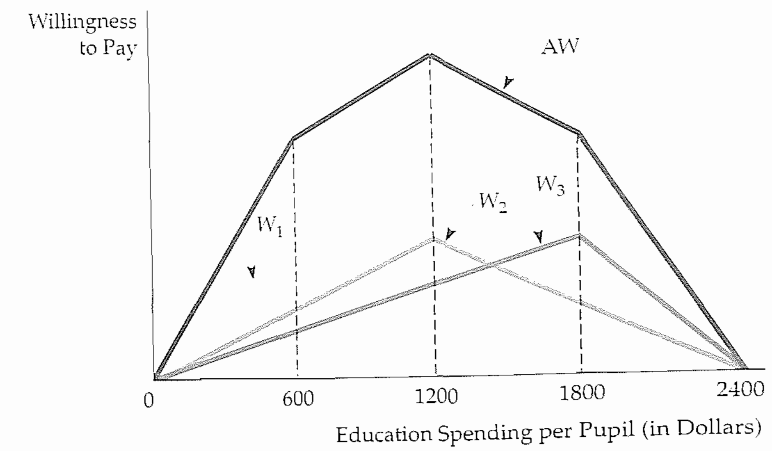

FIGURE 18.15 Determining the Level of Educational Spending

The efficient level of educational spending is determined by summing the willingness to pay for education (net of tax payments) of each of three citizens. Curves W_1, W_2 and W_3 represent their willingness to pay, and curve AW represents the aggregate willingness to pay. The efficient level of spending is $1200 per pupil. The level of spending actually provided is the level demanded by the median voter. In this particular case, the median voter's preference (given by the peak of the W_2 curve) is also the efficient level.

pay for that education increase as well. The willingness-to-pay curve, which represents the net benefit of educational spending, initially slopes upward because the citizen places great value on low spending levels. When spending increases beyond $600 per pupil, however, the value that the household puts on education increases at a diminishing rate. The net benefit, therefore, actually declines. Eventually, the spending level becomes so great (at $2400 per pupil) that the citizen is indifferent between this level of spending and no spending at all.

Curve W_2, which represents the second citizen's willingness to pay (net of taxes) is similarly shaped but reaches its maximum at a spending level of $1200 per pupil. Finally, W_3, the willingness to pay of the third citizen, peaks at $1800 per pupil.

The dark line labeled AW represents the aggregate willingness to pay for education—the vertical summation of the W_1, W_2, and W_3 curves. The AW curve measures the maximum amount that all three citizens are willing to pay to enjoy each spending level. As Figure 18.15 shows, the aggregate willingness to pay is maximized when $1200 per pupil is spent. Because the AW curve measures the benefit of spending net of the tax payments required to pay for that spending, the maximum point, $1200 per pupil, also represents the efficient level of spending.

Will majority-rule voting achieve the efficient outcome in this case? Suppose the public must vote whether to spend $1200 or $600 per pupil. The first citizen will vote for $600, but the other two citizens will vote for $1200, which will then have been chosen by majority rule. In fact, $1200 per pupil will beat any other alternative in a majority-rule vote. Thus $1200 represents the most preferred alternative of the *median voter*—the citizen with the median or middle preference. (The first citizen prefers $600 and the third $1800.) *Under majority rule voting, the preferred spending level of the median voter will always win an election against any other alternative.*

But will the preference of the median voter be the efficient level of spending? In this case yes, because $1200 is efficient. But the preference of the median voter is often *not* the efficient spending level. Suppose the third citizen's preferences were the same as the second's. In that case, although the median voter's choice would still be $1200 per pupil, the efficient level of spending would be less than $1200 (because the efficient level involves an average of the preferences of all three citizens). In this case, majority rule would lead to too much spending on education. If we reversed the example so that the first and second citizens' preferences were identical, majority rule would generate too little educational spending.

Thus although majority-rule voting allows the preferences of the median voter to determine referenda outcomes, these outcomes need not be economically efficient. Majority rule is inefficient because it weighs each citizen's preference equally: The efficient outcome weighs each citizen's vote by his or her strength of preference.

SUMMARY

1. An externality occurs when a producer or a consumer affects the production or consumption activities of others in a manner that is not directly reflected in the market. Externalities cause market inefficiencies because they inhibit the ability of market prices to convey accurate information about how much to produce and how much to buy.
2. Pollution is a common example of an externality that leads to market failure. It can be corrected by emissions standards, emissions fees, marketable emissions permits, or by encouraging recycling. When there is uncertainty about costs and benefits, any one of these mechanisms can be preferable, depending on the shapes of the marginal social cost and marginal benefit curves.
3. Inefficiencies due to market failure may be eliminated through private bargaining among the affected parties. According to the Coase theorem, the bargaining solution will be efficient when property rights are clearly specified, when transactions costs are zero, and when there is no strategic behavior. But bargaining is unlikely to generate an efficient outcome because parties frequently behave strategically.
4. Common property resources are not controlled by a single person and can be used without a price being paid. As a result of free usage, an externality is created in which the current overuse of the resource harms those who might use it in the future.
5. Goods that private markets are not likely to produce efficiently are either nonrival or nonexclusive. Public goods are both. A good is nonrival if for any given level of production, the marginal cost of providing it to an additional consumer is zero. A good is nonexclusive if it is expensive or impossible to exclude people from consuming it.
6. A public good is provided efficiently when the vertical sum of the individual demands for the public good is equal to the marginal cost of producing it.
7. Majority-rule voting is one way for citizens to voice their preference for public goods. Under majority rule, the level of spending provided will be that preferred by the median voter. This level need not be the efficient outcome.

QUESTIONS FOR REVIEW

1. Which of the following describes an externality and which does not? Explain the difference.
   a. A policy of restricted coffee exports in Brazil causes the U.S. price of coffee to rise—an increase which in turn also causes the price of tea to rise.
   b. An advertising blimp distracts a motorist who then hits a telephone pole.
2. Compare and contrast the following three mechanisms for treating pollution externalities when the costs and benefits of abatement are uncertain: (a) an emissions fee, (b) an emissions standard, and (c) a system of transferable emissions permits.
3. When do externalities require government intervention? When is such intervention unlikely to be necessary?

4. An emissions fee is paid to the government, whereas an injurer who is sued and held liable pays damages directly to the party harmed by an externality. What differences in the behavior of victims might you expect to arise under these two arrangements?
5. Why does free access to a common property resource generate an inefficient outcome?
6. Public goods are both nonrival and nonexclusive. Explain each of these terms and show clearly how they differ from each other.
7. Public television is funded in part by private donations, even though anyone with a television set can watch for free. Can you explain this phenomenon in light of the free rider problem?
8. Explain why the median voter outcome need not be efficient when majority-rule voting determines the level of public spending.

EXERCISES

1. A number of firms have located in the western portion of a town after single-family residences took up the eastern portion. Each firm produces the same product and in the process emits noxious fumes that adversely affect the residents of the community.
   a. Why is there an externality created by the firms?
   b. Do you think that private bargaining can resolve the problem? Explain.
   c. How might the community determine the efficient level of air quality?
2. A computer programmer lobbies against copyrighting software. He argues that everyone should benefit from innovative programs written for personal computers and that exposure to a wide variety of computer programs will inspire young programmers to create even more innovative programs. Considering the marginal social benefits possibly gained by his proposal, do you agree with his position?
3. Assume that scientific studies provide you with the following information concerning the benefits and costs of sulfur dioxide emissions:

Benefits of abating (reducing) emissions:	$MB = 400 - 10A$
Costs of abating emissions:	$MC = 100 + 20A$

where A is the quantity abated in millions of tons and the benefits and costs are given in dollars per ton.
   a. What is the socially efficient level of emissions abatement?
   b. What are the marginal benefit and marginal cost of abatement at the socially efficient level of abatement?
   c. What happens to net social benefits (benefits minus costs) if you abate one million more tons than the efficient level? One million fewer?
   d. Why is it socially efficient to set marginal benefits equal to marginal costs rather than abating until total benefits equal total costs?
4. Four firms located at different points on a river dump various quantities of effluent into it. The effluent adversely affects the quality of swimming for homeowners who live downstream. These people can build swimming pools to avoid swimming in the river, and the firms can purchase filters that eliminate harmful chemicals in the material dumped in the river. As a policy adviser for a regional planning organization, how would you compare and contrast the following options for dealing with the harmful effect of the effluent:
   a. An equal-rate effluent fee on firms located on the river.
   b. An equal standard per firm on the level of effluent that each can dump.
   c. A transferable effluent permit system in which the aggregate level of effluent is fixed and all firms receive identical permits.
5. Medical research has shown the negative health effects of "secondhand" smoke. Recent social trends point to growing intolerance of smoking in public areas. If you are a smoker and you wish to continue smoking despite tougher anti-smoking laws, describe the effect of the following legislative proposals on your behavior. As a result of these programs, do you, the individual smoker, benefit? Does society benefit as a whole?
   a. A bill is proposed that would lower tar and nicotine levels in all cigarettes.
   b. A tax is levied on each pack of cigarettes.
   c. A tax is levied on each pack of cigarettes sold.
   d. Smokers would be required to carry government-issued smoking permits at all times.
6. A beekeeper lives adjacent to an apple orchard. The orchard owner benefits from the bees because each hive pollinates about one acre of apple trees. The orchard owner pays nothing for this service, however, because the bees come to the orchard without his having to do anything. Because there are not enough bees to pollinate the entire orchard, the orchard owner must complete the pollination by artificial means, at a cost of $10 per acre of trees.

   Beekeeping has a marginal cost $MC = 10 + 2Q$, where Q is the number of beehives. Each hive yields $20 worth of honey.
   a. How many beehives will the beekeeper maintain?
   b. Is this the economically efficient number of hives?
   c. What changes would lead to the more efficient operation?
7. There are three groups in a community. Their demand curves for public television in hours of programming, T, are given respectively by

$$W_1 = \$150 - T$$
$$W_2 = \$200 - 2T$$
$$W_3 = \$250 - T$$

Suppose public television is a pure public good that can be produced at a constant marginal cost of $200 per hour.
   a. What is the efficient number of hours of public television?
   b. How much public television would a competitive private market provide?
8. Reconsider the common resource problem given in Example 18.5. Suppose that crawfish popularity continues to increase, and that the demand curve shifts from $C = 0.401 - 0.0064F$ to $C = 0.50 - 0.0064F$. How does this shift in demand affect the actual crawfish catch, the efficient catch, and the social cost of common access? (*Hint:* Use the marginal social cost and private cost curves given in the example.)
9. The Georges Bank, a highly productive fishing area off New England, can be divided into two zones in terms of fish population. Zone 1 has the higher population per square mile but is subject to severe diminishing returns to fishing effort. The daily fish catch (in tons) in Zone 1 is

$$F_1 = 200(X_1) - 2(X_1)^2$$

where X_1 is the number of boats fishing there. Zone 2 has fewer fish per mile but is larger, and diminishing returns are less of a problem. Its daily fish catch is

$$F_2 = 100(X_2) - (X_2)^2$$

where X_2 is the number of boats fishing in Zone 2. The marginal fish catch MFC in each zone can be represented as

$$MFC_1 = 200 - 4(X_1)$$
$$MFC_2 = 100 - 2(X_2)$$

There are 100 boats now licensed by the U.S. government to fish in these two zones. The fish are sold at $100 per ton. Total cost (capital and operating) per boat is constant at $1000 per day. Answer the following questions about this situation:
   a. If the boats are allowed to fish where they want, with no government restriction, how many will fish in each zone? What will be the gross value of the catch?
   b. If the U.S. government can restrict the boats, how many should be allocated to each zone? What will be the gross value of the catch? Assume the total number of boats remains at 100.
   c. If additional fishermen want to buy boats and join the fishing fleet, should a government wishing to maximize the net value of the catch grant them licenses? Why or why not?

APPENDIX

The Basics of Regression

This appendix explains the basics of **multiple regression analysis**, using an example to illustrate its application in economics.[1] Multiple regression is a means of fitting economic relationships to data. It lets us quantify economic relationships and test hypotheses about them.

multiple regression analysis Statistical procedure for quantifying economic relationships and testing hypotheses about them.

In a **linear regression**, the relationships that we fit to the data are of the following form:

linear regression Model specifying a linear relationship between a dependent variable and several independent (or explanatory) variables and an error term

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k + e \qquad \textbf{(A.1)}$$

Equation (A.1) relates a *dependent* variable Y to several *independent* (or *explanatory*) variables, $X_1, X_2, \ldots$. For example, in an equation with two independent variables, Y might be the demand for a good, X_1 its price, and X_2 income. The equation also includes an *error term* e that represents the collective influence of any omitted variables that may also affect Y (for example, prices of other goods, the weather, unexplainable shifts in consumers' tastes, etc.). Data are available for Y and the Xs, but the error term is assumed to be unobservable.

Note that Equation (A.1) must be linear in the *parameters*, but it need not be linear in the variables. For example, if Equation (A.1) represented a demand function, Y might be the *logarithm* of quantity (log Q), X_1 the logarithm of price (log P), and X_2 the logarithm of income (log I):

$$\log Q = b_0 + b_1 \log P + b_2 \log I + e \qquad \textbf{(A.2)}$$

Our objective is to obtain *estimates* of the parameters $b_0, b_1, \ldots, b_k$ that provide a "best fit" to the data. We explain how this is done below.

An Example

Suppose we wish to explain and then forecast quarterly automobile sales in the United States. Let's start with a simplified case in which sales S (in billions of dollars) is the dependent variable that will be explained. The only explanatory variable is the price of new automobiles P (measured by a new car price index scaled so that 1967 = 100). We could write this simple model as

$$S = b_0 + b_1P + e \qquad \textbf{(A.3)}$$

[1] For a textbook treatment of applied econometrics, it's hard to think of a better reference than R. S. Pindyck and D. L. Rubinfeld, *Econometric Models and Economic Forecasts*, 4th ed. (New York: McGraw-Hill, 1998).

In Equation (A.3), b_0 and b_1 are the parameters to be determined from the data, and e is the random error term. The parameter b_0 is the intercept, while b_1 is the slope: It measures the effect of a change in the new car price index on automobile sales.

Were no error term present, the relationship between S and P would be a straight line that describes the systematic relationship between the two variables. However, not all the actual observations fall on the line, so the error term e is required to account for omitted factors.

Estimation

Some criterion for a "best fit" is needed to choose values for the regression parameters. The criterion most often used is to *minimize the sum of squared residuals* between the actual values of Y and the *fitted* values for Y obtained after equation (A.1) has been estimated. This is called the **least-squares criterion.** If we denote the estimated parameters (or *coefficients*) for the model in (A.1) by $\hat{b}_0$, $\hat{b}_1, \ldots, \hat{b}_k$, then the *fitted* values for Y are given by

least-squares criterion Criterion of "best fit" used to choose values for regression parameters, usually by minimizing the sum of squared residuals between the actual values of the dependent variable and the fitted values.

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_k X_k \qquad \textbf{(A.4)}$$

Figure A.1 illustrates this for our example, in which there is a single independent variable. The data are shown as a scatter of points with sales on the vertical axis and price on the horizontal. The fitted regression line is drawn through the data points. The fitted value for sales associated with any particular value for the price values P_i is given by $\hat{S}_i = \hat{b}_0 + \hat{b}_1 P_i$ (at point B).

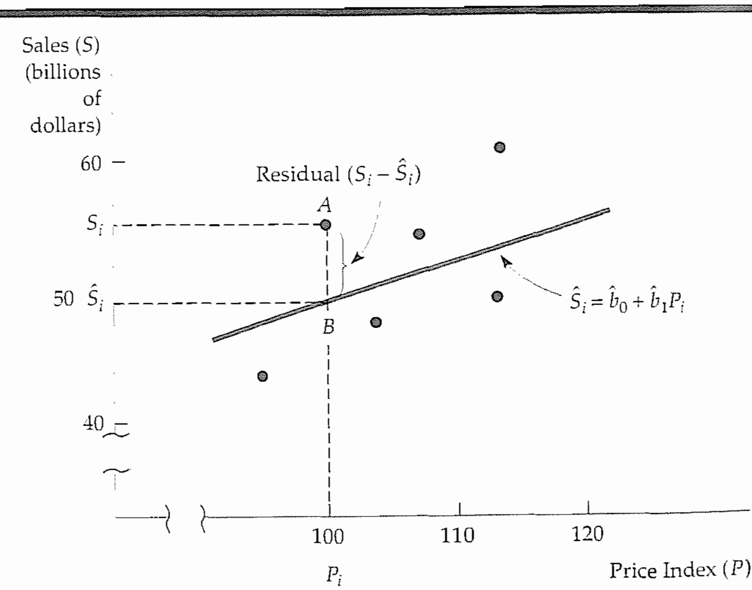

FIGURE A.1 Least Squares

The regression line is chosen to minimize the sum of squared residuals. The residual associated with price P_i is given by line AB.

For each data point, the regression *residual* is the difference between the actual and fitted value of the dependent variable. The residual $\hat{e}_i$, associated with data point A in the figure, is given by $\hat{e}_i = S_i - \hat{S}_i$. The parameter values are chosen so that when all the residuals are squared and then added, the resulting sum is minimized. In this way, positive errors and negative errors are treated symmetrically; large errors are given a more-than-proportional weight. As we will see shortly, this criterion lets us do some simple statistical tests to help interpret the regression.

As an example of estimation, let's return to the two-variable model of auto sales given by equation (A.3). The result of fitting this equation to the data using the least-squares criterion is

$$\hat{S} = -25.5 + 0.57P \qquad \textbf{(A.5)}$$

In equation (A.5), the intercept -25.5 indicates that if the price index were zero, sales would be \$−25.5 billion. The slope parameter indicates that a 1-unit increase in the price index for new cars leads to a \$0.57 billion increase in auto sales. This rather surprising result—an upward-sloping demand curve—is inconsistent with economic theory and should make us question the validity of our model.

Let's expand the model to consider the possible effects of two additional explanatory variables: personal income I (in billions of dollars) and the rate of interest R (the three-month Treasury bill rate). The estimated regression when there are three explanatory variables is

$$\hat{S} = 51.1 - 0.42P + 0.046I - 0.84R \qquad \textbf{(A.6)}$$

The importance of including all relevant variables in the model is suggested by the change in the regression results after the income and interest rate variables are added. Note that the coefficient of the P variable has changed substantially, from 0.57 to -0.42. The coefficient -0.42 measures the effect of an increase in price on sales, *with the effect of interest rates and income held constant*. The negative price coefficient is consistent with a downward-sloping demand curve. Clearly, the failure to control for interest rates and income leads to the false conclusion that sales and price are positively related.

The income coefficient, 0.046, tells us that for every \$1 billion increase in personal income in the United States, automobile sales are likely to increase by \$46 million (or \$0.046 billion). The interest rate coefficient reflects the fact that for every one percentage point increase in the rate of interest, automobile sales are likely to fall by \$840 million. Clearly, automobile sales are very sensitive to the cost of borrowing.

Statistical Tests

Our estimates of the true (but unknown) parameters are numbers that depend on the set of observations that we started with—that is, with our **sample.** With a different sample we would obtain different estimates.[2] If we continue to collect more and more samples and generate additional estimates, the estimates of

sample Set of observations for study, drawn from a larger universe.

[2] The least-squares formula that generates these estimates is called the *least-squares estimator*, and its values vary from sample to sample.

each parameter will follow a probability distribution. This distribution can be summarized by a *mean* and a measure of dispersion around that mean, a standard deviation that we refer to as the *standard error of the coefficient*.

Least-squares has several desirable properties. First, it is *unbiased*. Intuitively, this means that if we could run our regression over and over again with different samples, the average of the many estimates we obtained for each coefficient would be the true parameter. Second, least-squares is *consistent*. In other words, if our sample were very large, we would obtain estimates that came very close to the true parameters.

In econometric work, we often assume that the error term, and therefore the estimated parameters, are normally distributed. The normal distribution has the property that the area within 1.96 standard errors of its mean is equal to 95 percent of the total area. With this information, we can ask the following question: Can we construct an interval around $\hat{b}$ such that there is a 95-percent probability that the true parameter lies within that interval? The answer is yes, and this 95-percent *confidence interval* is given by

$$\hat{b} \pm 1.96 \text{ (standard error of } \hat{b}) \qquad \textbf{(A.7)}$$

Thus when working with an estimated regression equation, we must not only look at the *point* estimates but also examine the standard errors of the coefficients to determine bounds for the true parameters.[3]

If a 95-percent confidence interval contains 0, then the true parameter b may actually be zero (even if our estimate $\hat{b}$ is not). This result implies that the corresponding independent variable may *not* really affect the dependent variable, even if we thought it did. We can test the hypothesis that a true parameter is actually equal to 0 by looking at its *t-statistic*, which is defined as

$$t = \frac{\hat{b}}{\text{Standard error of } \hat{b}} \qquad \textbf{(A.8)}$$

If the t-statistic is less than 1.96 in magnitude, the 95-percent confidence interval around $\hat{b}$ must include 0. This means that we cannot reject the hypothesis that the true parameter b equals 0. We therefore say that our estimate, whatever it may be, is *not statistically significant*. Conversely, if the t-statistic is greater than 1.96 in absolute value, we reject the hypothesis that $b = 0$ and call our estimate *statistically significant*.

Equation (A.9) shows the multiple regression for the auto sales model (equation A.6) with a set of standard errors and t-statistics added:

$$\begin{array}{cccc} \hat{S} = 51.1 & -0.42P & +0.046I & -0.84R \\ (9.4) & (0.13) & (0.006) & (0.32) \\ t = 5.44 & -3.23 & 7.67 & -2.63 \end{array} \qquad \textbf{(A.9)}$$

The standard error of each estimated parameter is given in parentheses just below the estimate, and the corresponding t-statistics appear below that.

[3] When there are fewer than 100 observations, we multiply the standard error by a number somewhat larger than 1.96.

Let's begin by considering the price variable. The standard error of 0.13 is small relative to the coefficient −0.42. In fact, we can be 95 percent certain that the *true* value of the price coefficient is −0.42 plus or minus 1.96 standard deviations (i.e., −0.42 plus or minus [1.96][0.13] = −0.42 ± 0.25). This puts the true value of the coefficient between −0.17 and −0.67. Because this range does not include zero, the effect of price is both significantly different from zero and negative. We can also arrive at this result from the t-statistic. The t of −3.23 reported in equation (A.9) for the price variable is equal to −0.42 divided by 0.13. Because this t-statistic exceeds 1.96 in absolute value, we conclude that price is a significant determinant of auto sales.

Note that the income and interest rate variables are also significantly different from zero. The regression results tell us that an increase in income is likely to have a statistically significant positive effect on auto sales, whereas an increase in interest rates w'll have a significant negative effect.

Goodness of Fit

Reported regression results usually contain information that tells us how closely the regression line fits the data. One statistic, the **standard error of the regression (SER)** is an estimate of the standard deviation of the regression error term e. Whenever all the data points lie on the regression line, the SER is zero. Other things being equal, the larger the standard error of the regression, the poorer the fit of the data to the regression line. To decide whether the SER is large or small, we compare it in magnitude with the mean of the dependent variable. This comparison provides a measure of the *relative* size of the SER, a more meaningful statistic than its absolute size.

standard error of the regression Estimate of the standard deviation of the regression error.

R-squared (R^2), the percentage of the variation in the dependent variable that is accounted for by all the explanatory variables, measures the overall goodness-of-fit of the multiple regression equation.[4] Its value ranges from 0 to 1. An R^2 of 0 means that the independent variables explain none of the variation of the dependent variable; an R^2 of 1 means that the independent variables explain the variation perfectly. The R^2 for the sales equation (A.9) is 0.94. This tells us that the three independent variables explain 94 percent of the variation in sales.

R-squared Percentage of the variation in the dependent variable that is accounted for by all the explanatory variables.

Note that a high R^2 does not by itself mean that the variables actually included in the model are the appropriate ones. First, the R^2 varies with the types of data being studied. Time series data with substantial upward growth usually generate much higher R^2s than do cross-section data. Second, the underlying economic theory provides a vital check. If a regression of auto sales on the price of wheat happened to yield a high R^2, we would question the model's reliability because our theory tells us that changes in the price of wheat have little or no effect on automobile sales.

The overall reliability of a regression result depends on the formulation of the model. When studying an estimated regression, we should consider things that might make the reported results suspicious. First, have variables that should appear in the relationship been omitted? That is, is the *specification* of the equation wrong? Second, is the functional form of the equation correct? For instance, should variables be in logarithms? Third, is there another relationship that relates one of the explanatory variables (say X) to the dependent variable Y? If so, X and Y are jointly determined, and we must deal with a two-equation model,

[4] The variation in Y is the sum of the squared deviations of Y from its mean. R^2 and SER provide similar information about goodness of fit, because $R^2 = 1 - \text{SER}^2/\text{Variance}(Y)$.

not one with a single equation. Finally, does adding or removing one or two data points result in a major change in the estimated coefficients—i.e., is the equation *robust*? If not, we should be very careful not to overstate the importance or reliability of the results.

Economic Forecasting

A forecast is a prediction about the values of the dependent variable, given information about the explanatory variables. Often, we use regression models to generate *ex ante forecasts*, in which we predict values of the dependent variable beyond the time period over which the model has been estimated. (If we know the values of the explanatory variables, the forecast is *unconditional*; if they must be predicted as well, the forecast is *conditional* on these predictions.) Sometimes *ex post* forecasts, in which we predict what the value of the dependent variable would have been if the values of the independent variables had been different, can be useful. An *ex post* forecast has a forecast period such that all values of the dependent and explanatory variables are known. Thus *ex post* forecasts can be checked against existing data and provide a direct means of evaluating a model.

For example, reconsider the auto sales regression discussed above. In general, the forecasted value for auto sales is given by

$$\hat{S} = \hat{b}_0 + \hat{b}_1 P + \hat{b}_2 I + \hat{b}_3 R + \hat{e} \quad \text{(A.10)}$$

where $\hat{e}$ is our prediction for the error term. Without additional information, we usually take $\hat{e}$ to be zero.

Then, to calculate the forecast we use the estimated sales equation:

$$\hat{S} = 51.1 - 0.42P + 0.046I - 0.84R \quad \text{(A.11)}$$

We can use (A.11) to predict sales when, for example, $P = 100$, $I = \$1$ trillion, and $R = 8$ percent. Then,

$$\hat{S} = 51.1 - 0.42(100) + 0.046(1000 \text{ billion}) - 0.84(8\%) = \$48.4 \text{ billion}$$

Note that \$48.4 billion is an *ex post* forecast for a time when $P = 100$, $I = \$1$ trillion, and $R = 8$ percent.

To determine the reliability of *ex ante* and *ex post* forecasts, we use the *standard error of forecast (SEF)*. The SEF measures the standard deviation of the forecast error within a sample in which the explanatory variables are known with certainty. Two sources of error are implicit in the SEF. The first is the error term itself, because $\hat{e}$ may not equal 0 in the forecast period. The second source arises because the estimated parameters of the regression model may not be exactly equal to the true parameters.

As an application, consider the SEF of \$7.0 billion associated with equation (A.11). If the sample size is large enough, the probability is roughly 95 percent that the predicted sales will be within 1.96 standard errors of the forecasted value. In this case, the 95-percent confidence interval is \$48.4 billion ± \$14.0 billion, i.e., from \$34.4 billion to \$62.4 billion.

Now suppose we wish to forecast automobile sales for some date in the future, such as 2003. To do so, the forecast must be conditional because we need to predict the values for the independent variables before calculating the forecast for automobile sales. Assume, for example, that our predictions of these variables are as follows: $\hat{P} = 200$, $\hat{I} = \$5$ trillion, and $\hat{R} = 10$ percent. Then, the forecast is given by $\hat{P} = 51.1 - 0.42(200) + 0.046(5000 \text{ billion}) - 0.84(10) = \$188.7$ billion. Here \$190.8 billion is an *ex ante* conditional forecast.

Because we are predicting the future, and because the explanatory variables do not lie close to the means of the variables throughout our period of study, the SEF is equal to \$8.2 billion, which is somewhat greater than the SEF that we calculated previously.[5] The 95-percent confidence interval associated with our forecast is the interval from \$172.3 billion to \$205.1 billion.

EXAMPLE A.1 The Demand for Coal

Suppose we want to estimate the demand for bituminous coal (given by sales in tons per year, COAL) and then use the relationship to forecast future coal sales. We would expect the quantity demanded to depend on the price of coal (given by the Producer Price Index for coal, PCOAL) and on the price of a close substitute for coal (given by the Producer Price Index for natural gas, PGAS). Because coal is used to produce steel and electricity, we would also expect the level of steel production (given by the Federal Reserve Board Index of iron and steel production, FIS) and electricity production (given by the Federal Reserve Board Index of electric utility production, FEU) to be important demand determinants.

Our model of coal demand is therefore given by the following equation:

$$\text{COAL} = b_0 + b_1 \text{PCOAL} + b_2 \text{PGAS} + b_3 \text{FIS} + b_4 \text{FEU} + e$$

From our theory, we would expect b_1 to be negative because the demand curve for coal is downward sloping. We would also expect b_2 to be positive because a higher price of natural gas should lead industrial consumers of energy to substitute coal for natural gas. Finally, we would expect both b_3 and b_4 to be positive because the greater the production of steel and electricity, the greater the demand for coal.

This model was estimated using monthly time-series data covering eight years. The results (with t-statistics in parentheses) are:

$$\underset{(3.51)}{\text{COAL} = 12{,}262} + \underset{(6.46)}{92.34\ \text{FIS}} + \underset{(7.14)}{118.57\ \text{FEU}} - \underset{(-3.82)}{48.90\ \text{PCOAL}} + \underset{(3.18)}{118.91\ \text{PGAS}}$$

$$R^2 = 0.692 \qquad \text{SER} = 120{,}000$$

All the estimated coefficients have the signs that econometric theory would predict. Each coefficient is also statistically significantly different from zero because the t-statistics are all greater than 1.96 in absolute value. The R^2 of 0.692 says that the model explains more than two-thirds of the variation in coal sales. The standard error of the regression SER is equal to 120,000 tons of coal. Because the mean level of coal production was 3.9 million tons, SER represents approximately 3 percent of the mean value of the dependent variable. This suggests a reasonably good model fit.

[5] For more on SEF, see Pindyck and Rubinfeld, *Econometric Models and Economic Forecasts*, ch. 8.

TABLE A.1 Forecasting Coal Demand

	FORECAST	CONFIDENCE INTERVAL
1-month forecast (tons)	5.2 million	4.9–5.5 million
6-month forecast (tons)	4.7 million	4.4–5.0 million
12-month forecast (tons)	5.0 million	4.7–5.3 million

Now suppose we want to use the estimated coal demand equation to forecast coal sales up to one year into the future. To do so, we substitute values for each of the explanatory values for the 12-month forecasting period into the estimated equation. We also estimate the standard error of forecast (the estimate is 0.17 million tons) and use it to calculate 95-percent confidence intervals for the forecasted values of coal demand. Some representative forecasts and confidence intervals are given in Table A.1.

SUMMARY

1. Multiple regression is a means of fitting economic relationships to data.
2. The linear regression model, which relates one dependent variable to one or more independent variables, is usually estimated by choosing the intercept and slope parameters that minimize the sum of the squared residuals between the actual and predicted values of the dependent variable.
3. In a multiple-regression model, each slope coefficient measures the effect on the dependent variable of a change in the corresponding independent variable, holding the effects of all other independent variables constant.
4. A t-test can be used to test the hypothesis that a particular slope coefficient is different from zero.
5. The overall fit of the regression equation can be evaluated using the standard error of the regression (SER) (a value close to zero means a good fit) or R^2 (a value close to one means a good fit).
6. Regression models can be used to forecast future values of the dependent variable. The standard error of forecast (SEF) measures the accuracy of the forecast.

GLOSSARY

A

absolute advantage *(page 585)* Situation in which country 1 has an advantage over country 2 in producing a good because the cost of producing the good in 1 is lower than the cost of producing it in 2.

accounting cost *(page 204)* Actual expenses plus depreciation charges for capital equipment.

accounting profit *(page 273)* The difference between a firm's revenues and its costs, including accounting depreciation but excluding any opportunity costs.

actual return *(page 168)* Return that an asset earns.

actuarially fair *(page 163)* Situation in which an insurance premium is equal to the expected payout.

adverse selection *(page 598)* Form of market failure resulting from asymmetric information: if insurance companies must charge a single premium because they cannot distinguish between high-risk and low-risk individuals, more high-risk individuals will insure, making it unprofitable to sell insurance.

advertising elasticity of demand *(page 405)* Percentage change in quantity demanded resulting from a 1-percent increase in advertising expenditures.

advertising-to-sales ratio *(page 405)* Ratio of a firm's advertising expenditures to its sales.

agent *(page 609)* Individual employed by a principal to achieve the principal's objective.

antitrust laws *(page 360)* Rules and regulations prohibiting actions that restrain, or are likely to restrain, competition.

arbitrage *(page 8)* Practice of buying at a low price at one location and selling at a higher price in another.

arc elasticity of demand *(page 120)* Price elasticity calculated over a range of prices.

asset *(page 166)* Something that provides a flow of money or services to its owner.

asset beta *(page 547)* A constant that measures the sensitivity of an asset's return to market movements and, therefore, the asset's nondiversifiable risk.

asymmetric information *(page 596)* Situation in which a buyer and a seller possess different information about a transaction.

auction markets *(page 491)* Markets in which products are bought and sold through formal bidding processes.

average cost *(page 209)* Production cost per unit of output.

average expenditure *(page 352)* Price paid per unit of a good.

average expenditure curve *(page 510)* Supply curve representing the price per unit that a firm pays for a good.

average fixed cost (AFC) *(page 209)* Fixed cost divided by the level of output.

average product *(page 182)* Output per unit of a particular input.

average revenue *(page 329)* Revenue divided by the number of units sold, i.e., price per unit.

average total cost (ATC) *(page 209)* Firm's total cost divided by its level of output.

average variable cost (AVC) *(page 210)* Variable cost divided by the level of output.

B

backward-bending labor supply curve *(page 511)* The portion of the labor supply curve at which the wage rate increases and the hours of work supplied decreases, giving the curve a negative slope.

bad *(page 70)* Good for which less is preferred rather than more.

bandwagon effect *(page 127)* Positive network externality in which a consumer wishes to possess a good in part because others do.

barrier to entry *(page 346)* Condition that impedes entry by new competitors.

Bertrand model *(page 437)* Oligopoly model in which firms produce a homogeneous good, each firm treats the price of its competitors as fixed, and all firms decide simultaneously what price to charge.

bilateral monopoly *(page 358)* Market with only one seller and one buyer.

block pricing *(page 375)* Practice of charging different prices for different quantities or "blocks" of a good.

bond *(page 538)* Contract in which a borrower agrees to pay the bondholder (the lender) a stream of money.

budget constraint *(page 75)* Constraints that consumers face as a result of limited incomes.

budget line *(page 75)* All combinations of goods for which the total amount of money spent is equal to income.

bundling *(page 392)* Practice of selling two or more products as a package.

C

capital *(page 178)* Buildings, equipment, and inventories which can be utilized (along with labor and raw materials) to produce output.

Capital Asset Pricing Model (CAPM) *(page 547)* Model in which the risk premium for a capital investment depends on the correlation of the investment's return with the return on the entire stock market.

cardinal utility function *(page 74)* Utility function describing by how much one market basket is preferred to another.

cartel *(page 424)* Market in which some or all firms explicitly collude, coordinating prices and output levels to maximize joint profits.

cash flow *(page 204)* The actual outlays by a firm, including wages, salaries, costs of materials, and property rentals.

ceiling price *(page 54)* A maximum price that firms are allowed by the government to charge for a good.

chain-weighted price index *(page 96)* Cost-of-living index that accounts for changes in quantities of goods and services.

Clayton Act *(page 360)* As amended by the Robinson-Patman Act, a law that makes it illegal to discriminate by charging buyers of essentially the same product different prices.

Coase theorem *(page 640)* Principle that when parties can bargain without cost and to their mutual advantage, the resulting outcome will be efficient regardless of how property rights are specified.

Cobb-Douglas production function *(page 248)* Production function of the form $Q = AK^{\alpha}L^{\beta}$, where Q is the rate of output, K is the quantity of capital, and L is the quantity of labor, and where α and β are constants.

Cobb-Douglas utility function *(page 143)* Utility function $U(X,Y) = X^{a}Y^{1-a}$, where X and Y are two goods and a is a constant.

common property resource *(page 642)* Resource to which anyone has free access.

common-value auction *(page 492)* Auction in which the item has the same value to all bidders, but bidders do not know that value precisely and their estimates of it vary.

company cost of capital *(page 548)* Weighted average of the expected return on a company's stock and the interest rate that it pays for debt.

comparative advantage *(page 585)* Situation in which country 1 has an advantage over country 2 in producing a good because the cost of producing the good in 1, relative to the cost of producing other goods in 1, is lower than the cost of producing the good in 2, relative to the cost of producing other goods in 2.

competitive markets *(page 8)* Markets in which buyers and sellers individually have little or no ability to affect prices.

complements *(page 23)* Two goods for which an increase in the price of one leads to a decrease in the quantity demanded of the other.

completely inelastic demand *(page 32)* Consumers will buy a fixed quantity of a good regardless of its price.

constant returns to scale *(page 198)* Output doubles when all inputs are doubled.

constant-cost industry *(page 277)* Industry whose long-run supply curve is horizontal.

Consumer Price Index *(page 11)* Measure of the aggregate price level.

consumer surplus (individual) *(page 123)* Difference between what a consumer is willing to pay for a good and the amount actually paid.

consumer surplus (market) *(page 124)* Net benefit to all consumers purchasing a good. Equal to area under the demand curve above the price.

contract curve *(page 571)* Curve showing all efficient allocations of goods between two consumers, or of two inputs between two production functions.

cooperative game *(page 462)* Game in which participants can negotiate binding contracts that allow them to plan joint strategies.

corner solution *(page 84)* Situation in which the marginal rate of substitution for one good in a chosen market basket is not equal to the slope of the budget line.

cost function *(page 237)* Function relating cost of production to level of output and other variables that the firm can control.

cost-of-living index *(page 93)* Ratio of the present cost of a typical bundle of consumer goods and services compared with the cost during a base period.

Cournot equilibrium *(page 433)* Equilibrium in the Cournot model, in which each firm correctly assumes how much its competitor will produce and sets its own production level accordingly.

Cournot model *(page 431)* Oligopoly model in which firms produce a homogeneous good, each firm treats the output of its competitors as fixed, and all firms decide simultaneously how much to produce.

cross-price elasticity of demand *(page 32)* Percentage change in the quantity demanded of one good resulting from a 1-percent increase in the price of another.

cyclical industries *(page 38)* Industries in which sales tend to magnify cyclical changes in gross national product and national income.

D

deadweight loss *(page 292)* Net loss of total (consumer plus producer) surplus.

decreasing returns to scale *(page 198)* Output less than doubles when all inputs are doubled.

decreasing-cost industry *(page 280)* Industry whose long-run supply curve is downward sloping.

degree of economies of scope (SC) *(page 231)* Percentage of cost savings resulting when two or more products are produced jointly rather than individually.

demand curve *(page 21)* Relationship between the quantity of a good that consumers are willing to buy and the price of the good.

depletable resource *(page 552)* A natural resource, such as oil or copper, which if produced today is unavailable for future production.

depreciation *(page 204)* The decline in value of a capital asset as it is used over time.

derived demand *(page 502)* Demand for an input that depends on, and is derived from, both the firm's level of output and the cost of inputs.

deviation *(page 151)* Difference between expected payoff and actual payoff.

diminishing marginal utility *(page 90)* Principle that as more of a good is consumed, the consumption of additional amounts will yield smaller additions to utility.

discount rate *(page 542)* Rate used to compare the value of a dollar received in the future to the value of a dollar received today.

diseconomies of scale *(page 227)* A doubling of output requires more than a doubling of cost.

diseconomies of scope *(page 231)* Joint output of a single firm is less than could be achieved by separate firms when each produces a single product.

diversifiable risk *(page 546)* Risk that can be eliminated either by investing in many projects or by holding the stocks of many companies.

diversification *(page 161)* Reducing risk by allocating resources to a variety of activities whose outcomes are not closely related.

dominant firm *(page 450)* Firm with a large share of total sales that sets price to maximize profits, taking into account the supply response of smaller firms.

dominant strategy *(page 464)* Strategy that is optimal no matter what an opponent does.

duality *(page 144)* Alternative way of looking at the consumer's utility maximization decision: Rather than choosing the highest indifference curve, given a budget constraint, the consumer chooses the lowest budget line that touches a given indifference curve.

duopoly *(page 430)* Market in which two firms compete with each other.

durable good *(page 36)* A consumption or capital good bought to provide services for a long time.

Dutch auction *(page 492)* Auction in which a seller begins by offering an item at a relatively high price, then reduces it by fixed amounts until the item is sold.

E

economic cost *(page 204)* Cost to a firm of utilizing economic resources in production, including opportunity cost.

economic efficiency *(page 294)* Maximization of aggregate consumer and producer surplus.

economic profit *(page 273)* The difference between a firm's revenues and its costs, including any opportunity costs.

economic rent *(page 275)* Amount that firms are willing to pay for an input less the minimum amount necessary to obtain it.

economies of scale *(page 227)* Output can be doubled for less than a doubling of cost.

economies of scope *(page 231)* Joint output of a single firm is greater than output that could be achieved by two different firms when each produces a single product.

Edgeworth box *(page 569)* Diagram showing all possible allocations of either two goods between two people or of two inputs between two production processes.

effective yield (or rate of return) *(page 539)* Percentage return that one receives by investing in a bond.

efficiency wage *(page 617)* Wage that a firm will pay to an employee as an incentive not to shirk.

efficiency wage theory *(page 616)* Explanation for the presence of unemployment and wage discrimination which recognizes that labor productivity may be affected by the wage rate.

efficient allocation *(page 567)* Allocation of goods in which no one can be made better off unless someone else is made worse off.

elastic demand *(page 31)* When the percentage change in quantity demanded of a good in response to a 1-percent change in price is greater than 1 in magnitude.

elasticity *(page 30)* Percentage change in one variable resulting from a 1-percent increase in another.

emissions fee *(page 626)* Charge levied on each unit of a firm's emissions.

emissions standard *(page 626)* Legal limit on the amount of pollutant that a firm can emit.

Engel curve *(page 106)* Curve relating the quantity of a good consumed to income.

English (or oral) auction *(page 491)* Auction in which a seller actively solicits progressively higher bids from a group of potential buyers.

equal marginal principle *(page 91)* Principle that utility is maximized when the consumer has equalized the marginal utility per dollar of expenditure across all goods.

equilibrium (or market-clearing) price *(page 23)* Price that equates the quantity supplied to the quantity demanded.

equilibrium in dominant strategies *(page 465)* Outcome of a game in which each firm is doing the best it can regardless of what its competitors are doing.

excess demand *(page 573)* When the quantity demanded of a good exceeds quantity supplied.

excess supply *(page 573)* When the quantity supplied of a good exceeds quantity demanded.

exchange economy *(page 567)* Market in which two or more consumers trade two goods among themselves.

expansion path *(page 222)* Curve passing through points of tangency between a firm's isocost lines and its isoquants.

expected return *(page 168)* Return that an asset should earn on average.

expected utility *(page 156)* Sum of the utilities associated with all possible outcomes, weighted by the probability that each outcome will occur.

expected value *(page 150)* Probability-weighted average of the values associated with all possible outcomes.

extensive form of a game *(page 477)* Representation of possible moves in a game in the form of a decision tree.

extent of a market *(page 9)* Boundaries of a market, both geographical and in terms of range of products produced and sold within it.

externality *(page 294, 622)* Action taken by either a producer or a consumer which affects other producers or consumers but is not accounted for by the market price.

F

factors of production *(page 178)* Inputs into the production process (e.g., labor, capital, and materials).

feedback effect *(page 564)* A price or quantity adjustment in one market that is caused by price and quantity adjustments in related markets.

first-degree price discrimination *(page 371)* Practice of charging each customer her reservation price.

first-price auction *(page 492)* Auction in which the sales price is equal to the highest bid.

first theorem of welfare economics *(page 574)* If everyone trades in the competitive marketplace, all mutually beneficial trades will be completed and the resulting equilibrium allocation of resources will be economically efficient.

fixed cost (FC) *(page 206)* Cost that does not vary with the level of output.

fixed input *(page 181)* Production factor that cannot be varied.

fixed-proportions production function *(page 195)* Production function with L-shaped isoquants, so that only one combination of labor and capital can be used to produce each level of output.

fixed-weight index *(page 96)* Cost-of-living index in which the quantities of goods and services remain unchanged.

free entry (exit) *(page 253)* When there are no special costs that make it difficult for a firm to enter (or exit) an industry.

free rider *(page 647)* Consumer or producer who does not pay for a nonexclusive good in the expectation that others will.

G

game *(page 462)* Situation in which players (participants) make strategic decisions that take into account each other's actions and responses.

general equilibrium analysis *(page 564)* Simultaneous determination of the prices and quantities in all relevant markets, taking feedback effects into account.

Giffen good *(page 113)* Good whose demand curve slopes upward because the (positive) income effect is larger than the (negative) substitution effect.

H

Hicksian substitution effect *(page 147)* Alternative to the Slutsky equation for decomposing price changes without recourse to indifference curves.

horizontal integration *(page 613)* Organizational form in which several plants produce the same or related products for a firm.

I

ideal cost-of-living index *(page 94)* Cost of attaining a given level of utility at current prices relative to the cost of attaining the same utility at base-year prices.

import quota *(page 309)* Limit on the quantity of a good that can be imported.

income effect *(page 112)* Change in consumption of a good resulting from an increase in purchasing power, with relative price held constant.

income elasticity of demand *(page 32)* Percentage change in the quantity demanded resulting from a 1-percent increase in income.

income-consumption curve *(page 105)* Curve tracing the utility-maximizing combinations of two goods as a consumer's income changes.

increasing returns to scale *(page 198)* Output more than doubles when all inputs are doubled.

increasing-cost industry *(page 279)* Industry whose long-run supply curve is upward sloping.

indifference curve *(page 64)* Curve representing all combinations of market baskets that provide a consumer with the same level of satisfaction.

indifference map *(page 66)* Graph containing a set of indifference curves showing the market baskets among which a consumer is indifferent.

individual demand curve *(page 103)* Curve relating the quantity of a good that a single consumer will buy to its price.

industry *(page 8)* A collection of firms that sell the same or closely related products.

inferior good *(page 106)* A good for which consumption falls as an individual's income rises.

infinitely elastic demand *(page 32)* Consumers will buy as much of a good as they can get at a single price, but for any higher price the quantity demanded drops to zero, while for any lower price the quantity demanded increases without limit.

interest rate *(page 534)* Rate at which one can borrow or lend money.

intertemporal price discrimination *(page 382)* Practice of separating consumers with different demand functions into different groups by charging different prices at different points in time.

isocost line *(page 217)* Graph showing all possible combinations of labor and capital that can be purchased for a given total cost.

isoelastic demand curve *(page 118)* Demand curve with a constant price elasticity.

isoquant *(page 179)* Curve showing all possible combinations of inputs that yield the same output.

isoquant map *(page 180)* Graph combining several isoquants, used to describe a production function.

K

kinked demand curve model *(page 446)* Oligopoly model in which each firm faces a demand curve kinked at the currently prevailing price: at higher prices demand is very elastic, whereas at lower prices it is inelastic.

L

labor productivity *(page 188)* Average product of labor for an entire industry or for the economy as a whole.

Lagrangian *(page 140)* Function to be maximized or minimized, plus a variable (the Lagrange multiplier) multiplied by the constraint.

Laspeyres price index *(page 94)* Amount of money at current-year prices that an individual requires to purchase a bundle of goods and services chosen in a base year divided by the cost of purchasing the same bundle at base-year prices.

law of diminishing marginal returns *(page 185)* Principle that as the use of an input increases with other inputs fixed, the resulting additions to output will eventually decrease.

learning curve *(page 233)* Graph relating amount of inputs needed by a firm to produce each unit of output to its cumulative output.

least-squares criterion *(page 656)* Criterion of "best fit" used to choose values for regression parameters, usually by minimizing the sum of squared residuals between the actual values of the dependent variable and the fitted values.

Lerner Index of Monopoly Power *(page 341)* Measure of monopoly power calculated as excess of price over marginal cost as a fraction of price.

linear demand curve *(page 31)* Demand curve that is a straight line.

linear regression *(page 655)* Model specifying a linear relationship between a dependent variable and several independent (or explanatory) variables and an error term.

long run *(page 181)* Amount of time needed to make all production inputs variable.

long-run average cost curve (LAC) *(page 226)* Curve relating average cost of production to output when all inputs, including capital, are variable.

long-run competitive equilibrium *(page 274)* All firms in an industry are maximizing profit, no firm has an incentive to enter or exit, and price is such that quantity supplied equals quantity demanded.

long-run marginal cost curve (LMC) *(page 226)* Change in long-run total cost as output is increased incrementally by 1 unit.

M

macroeconomics *(page 4)* Branch of economics that deals with aggregate economic variables, such as the level and growth rate of national output, interest rates, unemployment, and inflation.

marginal benefit *(page 80)* Benefit from the consumption of one additional unit of a good.

marginal cost *(page 80)* Cost of one additional unit of a good.

marginal expenditure *(page 352)* Additional cost of buying one more unit of a good.

marginal expenditure curve *(page 510)* Curve describing the incremental cost of purchasing one additional unit of a good.

marginal external benefit *(page 624)* Increased benefit that accrues to other parties as a firm increases output by one unit.

marginal external cost *(page 622)* Increase in cost imposed externally as one or more firms increase output by one unit.

marginal product *(page 182)* Additional output produced as an input is increased by one unit.

marginal rate of substitution (MRS) *(page 68)* Amount of a good that a consumer is willing to give up in order to obtain one additional unit of another good.

marginal rate of technical substitution (MRTS) *(page 192)* Amount by which the quantity of one input can be reduced when one extra unit of another input is used, so that output remains constant.

marginal rate of transformation *(page 582)* Amount of one good that must be given up to produce one additional unit of a second good.

marginal revenue *(page 256)* Change in revenue resulting from a 1 unit increase in output.

marginal revenue product *(page 502)* Additional revenue resulting from the sale of output created by the use of one additional unit of an input.

marginal social benefit *(page 624)* Sum of the marginal private benefit plus the marginal external benefit.

marginal social cost *(page 622)* Sum of the marginal cost of production and the marginal external cost.

marginal utility (MU) *(page 90)* Additional satisfaction obtained from consuming one additional unit of a good.

marginal value *(page 352)* Additional benefit derived from purchasing one more unit of a good.

market *(page 7)* Collection of buyers and sellers that, through their actual or potential interactions, determine the price of a product or set of products.

market basket (or bundle) *(page 62)* List with specific quantities of one or more goods.

market definition *(page 8)* Determination of the buyers, sellers, and range of products that should be included in a particular market.

market demand curve *(page 116)* Curve relating the quantity of a good that all consumers in a market will buy to price.

market failure *(page 294)* Situation in which an unregulated competitive market is inefficient because prices fail to provide proper signals to consumers and producers.

market mechanism *(page 23)* Tendency in a free market for price to change until the market clears.

market power *(page 328)* Ability of a seller or buyer to affect the price of a good.

market price *(page 8)* Price prevailing in a competitive market.

market signaling *(page 601)* Process by which sellers send signals to buyers conveying information about product quality.

markup pricing *(page 341)* Increasing the production cost of a good by a fixed percentage to determine a sales price.

maximin strategy *(page 469)* Strategy that maximizes the minimum gain that can be earned.

median voter *(page 650)* The individual with the median preferred outcome among all voters.

method of Lagrange multipliers *(page 140)* Technique to maximize or minimize a function subject to one or more constraints.

microeconomics *(page 4)* Branch of economics that deals with the behavior of individual economic units—consumers, firms, workers, and investors—as well as the markets that these units comprise.

mixed bundling *(page 397)* Practice of selling two or more goods both as a package and individually.

mixed strategy *(page 470)* Strategy in which a player makes a random choice among two or more possible actions, based on a set of chosen probabilities.

monopolistic competition *(page 424)* Market in which firms can enter freely, each producing its own brand or version of a differentiated product.

monopoly *(page 328)* Market with only one seller.

monopoly power *(page 328)* The ability of a firm to profitably charge a price higher than marginal cost.

monopsony *(page 328)* Market with only one buyer.

monopsony power *(page 352)* Buyer's ability to affect the price of a good.

moral hazard *(page 606)* When an insured party whose actions are unobserved can affect the probability or magnitude of a payment associated with an event.

multiple regression analysis *(page 655)* Statistical procedure for quantifying economic relationships and testing hypotheses about them.

N

Nash equilibrium *(page 430)* Set of strategies or actions in which each firm does the best it can given its competitors' actions.

natural monopoly *(page 350)* Firm that can produce the entire output of the market at a cost lower than what it would be if there were several firms.

negatively correlated *(page 161)* Having a tendency to move in opposite directions (said of two variables).

net present value (NPV) criterion *(page 542)* Rule holding that one should invest if the present value of the expected future cash flow from an investment is larger than the cost of the investment.

network externality *(page 127)* When each individual's demand depends on the purchases of other individuals.

nominal discount rate *(page 544)* A discount rate that includes the effects of inflation.

nominal price *(page 11)* Absolute price of a good, unadjusted for inflation.

noncooperative game *(page 443)* Game in which negotiation and enforcement of binding contracts between players are not possible.

nondiversifiable risk *(page 546)* Risk that cannot be eliminated by investing in many projects or by holding the stocks of many companies.

nonexclusive goods *(page 645)* Goods that people cannot be excluded from consuming, so that it is difficult or impossible to charge for their use.

nonrival good *(page 644)* Good for which the marginal cost of its provision to an additional consumer is zero.

normal good *(page 106)* A good for which consumption increases when income rises.

normative analysis *(page 7)* Analysis examining questions of what ought to be.

O

oligopoly *(page 424)* Market in which only a few firms compete with one another, and entry by new firms is impeded.

oligopsony *(page 352)* Market with only a few buyers.

opportunity cost *(page 204)* Cost associated with opportunities that are forgone when a firm's resources are not put to their highest-value use.

opportunity cost of capital *(page 542)* Rate of return that one could earn by investing in an alternate project with similar risk.

optimal strategy *(page 462)* Strategy that maximizes player's expected payoff.

ordinal utility function *(page 74)* Utility function that generates a ranking of market baskets in order of most to least preferred.

P

Paasche index *(page 95)* Amount of money at current-year prices that an individual requires to purchase a bundle of goods and services divided by the cost of purchasing the same bundle in a base year.

parallel conduct *(page 360)* Form of implicit collusion in which one firm consistently follows actions of another.

Pareto efficiency *(page 567)* Synonymous with "efficient allocation"—an allocation of goods in which no one can be made better off without making someone else worse off.

partial equilibrium analysis *(page 564)* Determination of equilibrium prices and quantities in a market independent of effects from other markets.

payoff *(page 462)* Outcome of a game that generates rewards or benefits for the player.

payoff *(page 150)* Value associated with a possible outcome.

payoff matrix *(page 443)* Table showing profit (or payoff) to each firm given its decision and the decision of its competitor.

peak-load pricing *(page 382)* Practice of charging higher prices during peak periods when capacity constraints cause marginal costs to be high.

perfect complements *(page 70)* Two goods for which the MRS is infinite; the indifference curves are shaped as right angles.

perfect substitutes *(page 70)* Two goods for which the marginal rate of substitution of one for the other is a constant.

perpetuity *(page 538)* Bond paying out a fixed amount of money each year, forever.

point elasticity of demand *(page 119)* Price elasticity at a particular point on the demand curve.

positive analysis *(page 6)* Analysis describing relationships of cause and effect.

positively correlated *(page 162)* Having a tendency to move in the same direction (said of two variables).

predatory pricing *(page 361)* Practice of pricing to drive current competitors out of business and to discourage new entrants in a market so that a firm can enjoy higher future profits.

present discounted value *(page 535)* The current value of an expected future cash flow.

price-consumption curve *(page 102)* A curve tracing the utility-maximizing combinations of two goods as the price of one changes.

price discrimination *(page 371)* Practice of charging different prices to different consumers for similar goods.

price elasticity of demand *(page 30)* Percentage change in quantity demanded of a good resulting from a 1-percent increase in its price.

price elasticity of supply *(page 33)* Percentage change in quantity supplied resulting from a 1-percent increase in price.

price leadership *(page 447)* Pattern of pricing in which one firm regularly announces price changes that other firms then match.

price of risk *(page 170)* Extra risk that an investor must incur to enjoy a higher expected return.

price rigidity *(page 446)* Characteristic of oligopolistic markets by which firms are reluctant to change prices even if costs or demands change.

price signaling *(page 447)* Form of implicit collusion in which a firm announces a price increase in the hope that other firms will follow suit.

price support *(page 302)* Price set by government above free-market level and maintained by governmental purchases of excess supply.

price taker *(page 252)* Firm that has no influence over market price and that thus takes the price as a given.

price-consumption curve *(page 102)* Curve tracing the utility-maximizing combinations of two goods as the price of one changes.

principal–agent problem *(page 609)* Problem arising when managers (agents) pursue their own goals even when doing so entails lower profits for a firm's owners (the principals).

principal *(page 609)* Individual who employs one or more agents to achieve an objective.

prisoners' dilemma *(page 443)* Game theory example in which two prisoners must decide separately whether to confess to a crime; if a prisoner confesses, he will receive a lighter sentence and his accomplice will receive a heavier one, but if neither confesses, sentences will be lighter than if both confess.

private-value auction *(page 492)* Auction in which each bidder knows his individual valuation of the object up for bid, with valuations differing from bidder to bidder.

probability *(page 150)* Likelihood that a given outcome will occur.

producer surplus *(page 269)* Sum (over all units produced by a firm) of differences between market price of a good and marginal costs of production.

product transformation curve *(page 230)* Curve showing the various combinations of two different outputs (products) that can be produced with a given set of inputs.

production contract curve *(page 580)* Curve showing all technically efficient combinations of inputs.

production function *(page 178)* Function showing the highest output that a firm can produce for every specified combination of inputs.

production possibilities frontier *(page 581)* Curve showing the combinations of two goods that can be produced with fixed quantities of inputs.

profit *(page 255)* Difference between total revenue and total cost.

profit maximization *(page 254)* The goal of a firm; it is achieved when the marginal revenue of the firm is equal to the marginal cost of production.

property rights *(page 638)* Legal rules stating what people or firms may do with their property.

prospective sunk cost *(page 205)* A cost that has not yet been incurred but that cannot be recovered once it is incurred.

public good *(page 593)* Nonexclusive and nonrival good: the marginal cost of provision to an additional consumer is zero and people cannot be excluded from consuming it.

pure bundling *(page 397)* Practice of selling products only as a package.

pure strategy *(page 470)* Strategy in which a player makes a specific choice or takes a specific action.

R

R-squared *(page 659)* Percentage of the variation in the dependent variable that is accounted for by all the explanatory variables.

rate-of-return regulation *(page 351)* Whereby the maximum price allowed by a regulatory agency is based on the (expected) rate of return that a firm will earn.

reaction curve *(page 432)* Relationship between a firm's profit-maximizing output and the amount it thinks its competitor will produce.

real discount rate *(page 544)* The discount rate that applies when cash flows are in real terms, i.e., after netting out inflation.

real price *(page 11)* Price of a good relative to an aggregate measure of prices; price adjusted for inflation.

real return *(page 167)* Simple (or nominal) return on an asset, less the rate of inflation.

regulatory lag *(page 351)* Delays that are usually required to change a regulated price.

rent seeking *(page 348)* Spending money in socially unproductive efforts to acquire, maintain, or exercise monopoly power.

rental rate *(page 216)* Cost per year of renting one unit of capital.

repeated game *(page 472)* Game in which actions are taken and payoffs received over and over again.

reservation price *(page 371)* Maximum price that a customer is willing to pay for a good.

return *(page 167)* Total monetary flow of an asset as a fraction of its price.

returns to scale *(page 198)* Rate at which output increases as inputs are increased proportionately.

revealed preference *(page 86)* An approach to consumer theory in which preferences are determined by observing the choices consumers make.

risk *(page 150)* The possibility of several different outcomes occurring when the probability of each outcome is known.

risk averse *(page 157)* Preferring a certain income to a risky income with the same expected value.

risk-free return *(page 167)* A return which is free of risk, whether of default or interest rate fluctuations. An example is the return on U.S. Treasury bills.

risk loving *(page 157)* Preferring a risky income to a certain income with the same expected value.

risk neutral *(page 157)* Being indifferent between a certain income and an uncertain income with the same expected value.

risk premium *(page 158)* Maximum amount of money that a risk-averse person will pay to avoid taking a risk.

riskless (or risk-free) asset *(page 167)* Asset that provides a flow of money or services that is known with certainty.

risky asset *(page 166)* Asset that provides an uncertain flow of money or services.

S

sample *(page 657)* Set of observations for study, drawn from a larger universe.

sealed-bid auction *(page 492)* Auction in which all bids are made simultaneously in sealed envelopes, the winning bidder being the individual who has submitted the highest bid.

secondary supply *(page 40)* The supply from recycled scrap material.

second-degree price discrimination *(page 374)* Practice of charging different prices per unit for different quantities of the same good or service.

second-price auction *(page 492)* Auction in which the sales price is equal to the second-highest bid.

second theorem of welfare economics *(page 577)* If individual preferences are convex, then every efficient allocation is a competitive equilibrium for some initial allocation of goods.

sequential game *(page 476)* Game in which players move in turn, responding to each other's actions and reactions.

Sherman Act *(page 360)* A law that prohibits contracts, combinations, or conspiracies that restrain trade, and makes monopolizing or attempting to monopolize illegal.

shirking model *(page 617)* Principle that workers still have an incentive to shirk if a firm pays them a market-clearing wage, because fired workers can be hired somewhere else for the same wage.

short run *(page 181)* Period of time in which quantities of one or more production factors cannot be changed.

short-run average cost curve (SAC) *(page 226)* Curve relating average cost of production to output when level of capital is fixed.

shortage *(page 24)* Situation in which the quantity demanded exceeds the quantity supplied.

Slutsky equation *(page 146)* Formula for decomposing the effects of a price change into substitution and income effects.

snob effect *(page 129)* Negative network externality in which a consumer wishes to own an exclusive or unique good.

social welfare function *(page 576)* Weights applied to each individual's utility in determining what is socially desirable.

specific tax *(page 314)* Tax of a certain amount of money per unit sold.

Stackelberg model *(page 436)* Oligopoly model in which one firm sets its output before other firms do.

standard deviation *(page 152)* Square root of the average of the squares of the deviations of the payoffs associated with each outcome from their expected values.

standard error of the regression *(page 659)* Estimate of the standard deviation of the regression error.

stock of capital *(page 189)* Total amount of capital available for use in production.

strategic move *(page 479)* Action that gives a player an advantage by constraining his behavior.

strategy *(page 462)* Rule or plan of action for playing a game.

subsidy *(page 317)* Payment reducing the buyer's price below the seller's price; i.e., a negative tax.

substitutes *(page 22)* Two goods for which an increase in the price of one leads to an increase in the quantity demanded of the other.

substitution effect *(page 111)* Change in consumption of a good associated with a change in its price, with the level of utility held constant.

sunk cost *(page 205)* Expenditure that has been made and cannot be recovered.

supply curve *(page 20)* Relationship between the quantity of a good that producers are willing to sell and the price of the good.

surplus *(page 24)* Situation in which the quantity supplied exceeds the quantity demanded.

T

tariff *(page 309)* Tax on an imported good.

technical efficiency *(page 579)* When firms combine inputs to produce a given output as inexpensively as possible.

technological change *(page 189)* Development of new technologies allowing factors of production to be used more effectively.

theory of consumer behavior *(page 62)* Description of how consumers allocate incomes among different goods and services to maximize their well-being.

theory of the firm *(page 178)* Explanation of how a firm makes cost-minimizing production decisions and how its cost varies with its output.

third-degree price discrimination *(page 376)* Practice of dividing consumers into two or more groups with separate demand curves and charging different prices to each group.

tit-for-tat strategy *(page 473)* Repeated-game strategy in which a player responds in kind to an opponent's previous play, cooperating with cooperative opponents and retaliating against uncooperative ones.

total cost (TC) *(page 206)* Total economic cost of production, consisting of fixed and variable costs.

transfer prices *(page 413)* Internal prices at which parts and components from upstream divisions are "sold" to downstream divisions within a firm.

transferable emissions permits *(page 630)* System of marketable permits, allocated among firms, specifying the maximum level of emissions that can be generated.

transitivity of preferences *(page 63)* If a consumer prefers basket *A* to basket *B*, and also prefers basket *B* to basket *C*, then he will prefer basket *A* to basket *C*.

two-part tariff *(page 385)* Form of pricing in which consumers are charged both an entry and a usage fee.

tying *(page 402)* Practice of requiring a customer to purchase one good in order to purchase another.

U

user cost of capital *(page 215)* Sum of the annual cost of owning and using a capital asset, equal to economic depreciation plus forgone interest.

user cost of production *(page 554)* Opportunity cost of producing and selling a unit today and so making it unavailable for production and sale in the future.

utility *(page 73)* Numerical score representing the satisfaction that a consumer gets from a given market basket.

utility function *(page 73)* Formula that assigns a level of utility to individual market baskets.

utility possibilities frontier *(page 575)* Curve showing all efficient allocations of resources measured in terms of the utility levels of two individuals.

V

value of complete information *(page 164)* Difference between the expected value of a choice when there is complete information and the expected value when information is incomplete.

variability *(page 151)* Extent to which possible outcomes of an uncertain event may differ.

variable cost (VC) *(page 206)* Cost that varies as output varies.

variable profit *(page 372)* Sum of profits on each incremental unit produced by a firm; i.e., profit ignoring fixed costs.

vertical integration *(page 613)* Organizational form in which a firm contains several divisions, with some producing parts and components that others use to produce finished products.

W

welfare economics *(page 574)* Normative evaluation of markets and economic policy.

welfare effects *(page 289)* Gains and losses caused by government intervention in the market.

winner's curse *(page 494)* Situation in which the winner of a common-value auction is worse off as a consequence of overestimating the value of the item and thereby overbidding.

Z

zero economic profit *(page 273)* A firm is earning a normal return on its investment—i.e., that it is doing as well as it could by investing its money elsewhere.

ANSWERS TO SELECTED EXERCISES

Chapter 1

1. a. *False*. There is little or no substitutability across geographical regions of the United States. A consumer in Los Angeles, for example, will not travel to Houston, Atlanta, or New York for lunch just because hamburger prices are lower in those cities. Likewise, a McDonald's or Burger King in New York cannot supply hamburgers in Los Angeles, even if prices were higher in Los Angeles. In other words, a fast-food price increase in New York will affect neither the quantity demanded nor the quantity supplied in Los Angeles or other parts of the country.
   b. *False*. Although consumers are unlikely to travel across the country to buy clothing, suppliers can easily move clothing from one part of the country to another. Thus if clothing prices were substantially higher in Atlanta than Los Angeles, clothing companies could shift supplies to Atlanta, which would reduce the price there.
   c. *False*. Although some consumers might be die-hard Coke or Pepsi loyalists, there are many consumers who will substitute one for the other based on price differences. Thus there is a single market for colas.

Chapter 2

2. In 1998, $Q_D = 3244 - 283P$ and $Q_S = 1944 + 207P$. With new markets, $Q'_D = Q_D + 200 = 3444 - 283P$, $Q'_S = Q_S$, $3444 - 283P = 1944 + 207P$, $1500 = 490P$, and $P^* = \$3.06$. At P^*, $Q^* = 3244 - 283(3.06) = 2378.02$. So $P = \$3.06$, $Q = 2378$.

*5. a. Total demand is $Q = 3244 - 283P$; domestic demand is $Q_D = 1700 - 107P$; subtracting domestic demand from total demand gives export demand $Q_E = 1544 - 176P$. The initial market equilibrium price (as given in example) is $P^* = \$2.65$. With a 40-percent decrease in export demand, total demand becomes $Q = Q_D + 0.6Q_E = 1700 - 107P + 0.6(1544 - 176P) = 2626.4 - 212.6P$. Demand is equal to supply. Therefore:

$$2626.4 - 212.6P = 1944 + 207P$$
$$682.4 = 419.6P$$

So $P = \frac{682.4}{419.6} = \$1.626$ or \$1.63. At this price, $Q = 2281$. Yes, farmers should be worried. With this drop in quantity and price, revenue goes from \$6609 million to \$3718 million.

   b. If the U.S. government supports a price of \$3.50, the market is not in equilibrium. At this support price, demand is equal to $1700 - 107(3.50) = 1325.5$ and supply is $1944 + 207(3.50) = 2668.5$. There is excess supply (1343) which the government must buy, costing $\$3.50(1343) = \$4700.5$ million.

8. a. First, considering non-OPEC supply: $S_C = Q^* = 13$. With $E_S = 0.10$ and $P^* = \$18$, $E_S = d(P^*/Q^*)$ implies $d = 0.07$. Substituting for d, S_C, and P in the supply equation, $c = 11.74$, and $S_C = 11.74 + 0.07P$. Similarly, since $Q_D = 23$, $E_D = -b(P^*/Q^*) = -0.05$, and $b = 0.06$. Substituting for b, $Q_D = 23$, and $P = 18$ in the demand equation gives $23 = a - 0.06(18)$, so that $a = 24.08$. Hence $Q_D = 24.08 - 0.06P$.
   b. As above, $E_S = 0.4$ and $E_D = -0.4$: $E_S = d(P^*/Q^*)$ and $E_D = -b(P^*/Q^*)$, implying $0.4 = d(18/13)$ and $-0.4 = -b(18/23)$. So $d = 0.29$ and $b = 0.51$. Next solve for c and a: $S_C = c + dP$ and $Q_D = a - bP$, implying $13 = c + (0.29)(18)$ and $23 = a - (0.51)(18)$. So $c = 7.78$ and $a = 32.18$.

Chapter 3

5. a. See Figure 3(a) where B represents the number of packages of butter and M the number of packages of margarine.
   b. Convexity means that the curve is "bowed inward." Here, the indifference curves are not "strictly convex," since they are straight lines.

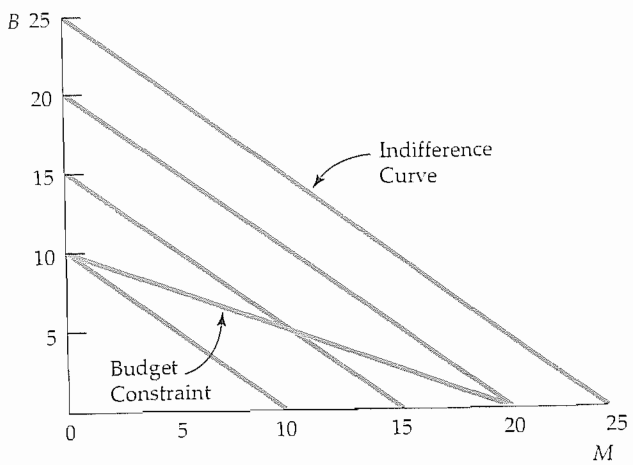

Figure 3(a)

c. The budget constraint is $Y = P_B B + P_M M$, $20 = 2B + M$, $B = 10 - 0.5M$. Given that Bill is indifferent between butter and margarine, and that the price of butter is greater than the price of margarine, Bill will buy only margarine.

6. a. See Figure 3(b), where A is the quantity of alcoholic drinks and N is the quantity of nonalcoholic drinks.

b. At any combination of A and N, Jones is willing to give up less of A to get some N than Smith is. Thus Jones has a *lower* MRS of A for N than Smith has. Jones' indifference curves are less steep than Smith's at any point on the graph.

c. To maximize satisfaction, each consumer must consume quantities such that the MRS between any two commodities equals the ratio of the prices. Their MRS's must be equal because they face the same prices. But because they have different preferences, they will consume different amounts of the two goods, A and N.

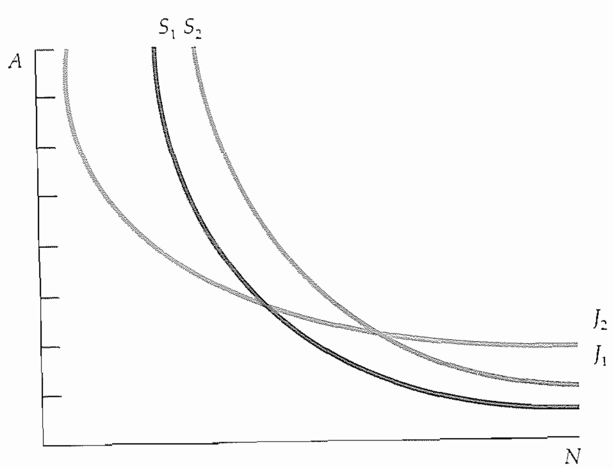

Figure 3(b)

8. In Figure 3(c) we plot miles flown, M, against all other goods, G, in dollars. The slope of the budget line is $-P_M/P_G$. The price of miles flown changes as miles flown changes, so the budget curve is kinked at 25,000 and 50,000 miles. Suppose P_M is \$1 per mile for $\leq$ 25,000 miles, then P_M = \$0.75 for $25{,}000 < M \leq 50{,}000$, and P_M = \$.50 for $M > 50{,}000$. Also, let P_G = \$1. Then the slope of the first segment is -1, the slope of the second segment is -0.75, and the slope of the last segment is -0.5.

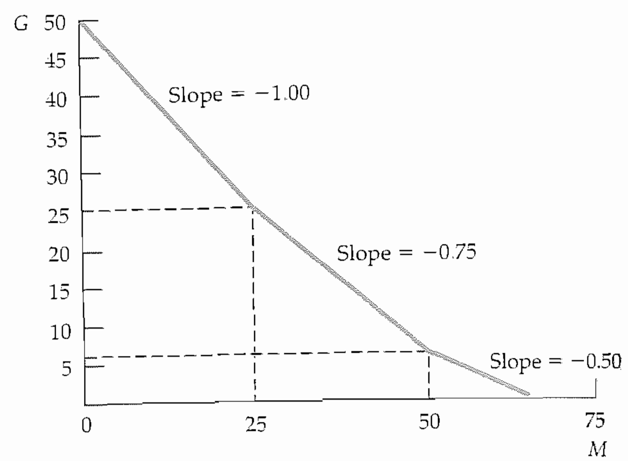

Figure 3(c)

Chapter 4

1. a. With a small change in price, the following point elasticity formula is appropriate: $E_p = \%\Delta Q/\%\Delta P$. For computer chips, $E_p = -2$, and for disk drives, $E_p = -1$. Let $TR_1 = P_1Q_1$ be revenue before the price exchange and $TR_2 = P_2Q_2$ be revenue after the price change. Then ΔTR would be $TR_2 - TR_1$. For computer chips, $\Delta TR = -12\%TR_1$. For disk drives, $\Delta TR = -1\%TR_1$.

b. Although we know the responsiveness of demand to changes in price, we need to know the quantities and the prices of the products to determine total sales revenues.

5. a. The demand curve is a straight line with a vertical intercept of $P = 12$ and a horizontal intercept of $Q = 6$ (since $Q = 6 - P/2$).

b. If there were no toll, the price P would be 0, so that $Q = 6$.

c. If the toll is \$6, $Q = 3$. The consumer surplus lost is the difference between consumer surplus when $P = 0$ (\$36) and consumer surplus when $P = 3$ (\$9), or \$27.

10. a. With small changes in price, the point elasticity formula would be appropriate. But here, the price of food doubles from \$2 to \$4, so arc elasticity should be used: $E_p = (\Delta Q/\Delta P)(\overline{P}/\overline{Q})$. We know that $E_p = -1$, $P = 2$, $\Delta P = 2$, and $Q = 5000$. So, if there is no change in income, we can solve for ΔQ: $-1 = (\Delta Q/2)[(2 + 1)/(5000 + \Delta Q/2)] = (\Delta Q \cdot 3)/(10{,}000 + \Delta Q)$. We find that $\Delta Q = -2500$: she decreases her consumption of food from 5000 to 2500 units.

b. A tax rebate of \$5000 implies an income increase of \$5000. To calculate the response of demand to the tax rebate, we use the definition of the arc income elasticity: $E_I = (\Delta Q/\Delta I)(\overline{I}/\overline{Q})$. We know that $E_I = 0.5$, $I = 25{,}000$, $\Delta I = 5000$, and $Q = 2500$. We solve for ΔQ: $0.5 = (\Delta Q/5000)[(25{,}000 + 2500)/(2500 + \Delta Q/2)]$. Since $\Delta Q = 238$, she increases her consumption of food from 2500 to 2738 units.

c. On her final indifference curve, she chooses to consume 2738 units of food (for \$10,952) and \$19,048 worth of all other goods. At the original food price of \$2, this combination would have cost her 2738 · \$2 + \$19,048 = \$24,524. So, she would have had an extra \$476 to spend on either food or other consumption, and would have been better off.

Chapter 4—Appendix

1. The first utility function can be represented as a series of straight lines; the second as a series of hyperbolas in the positive quadrant; and the third as a series of "L"s. Only the second utility function meets the definition of a strictly convex shape.

3. The Slutsky equation is $dX/dP_X = \partial X/\partial P^*|_{U=U^*} - X(\Delta X/\Delta I)$, where the first term represents the substitution effect and the second term represents the income effect. With this type of utility function the consumer does not substitute one good for the other when the price changes, so the substitution effect is zero.

Chapter 5

2. The four mutually exclusive states are given in Table 5 below.

TABLE 5

	Congress Passes Tariff	Congress Does Not Pass Tariff
Slow growth rate	State 1: Slow growth with tariff	State 2: Slow growth without tariff
Fast growth rate	State 3: Fast growth with tariff	State 4: Fast growth without tariff

7. Consumers with income X^* will simultaneously purchase unfair insurance and take unfair gambles. They are risk averse against large losses, but risk loving for large income gains.

Chapter 6

1. a. The average product of labor, AP, is equal to Q/L. The marginal product of labor, MP, is equal to $\Delta Q/\Delta L$. The relevant calculations are given in the following table.

L	Q	AP	MP
0	0	—	—
1	10	10	10
2	17	$8\frac{1}{2}$	7
3	22	$7\frac{1}{3}$	5
4	25	$6\frac{1}{4}$	3
5	26	$5\frac{1}{5}$	1
6	25	$4\frac{1}{6}$	−1
7	23	$3\frac{2}{7}$	−2

b. This production process exhibits diminishing returns to labor, which is characteristic of all production functions with one fixed input. Each additional unit of labor yields a smaller increase in output than the last unit of labor.

c. Labor's negative marginal product can arise from congestion in the chair manufacturer's factory. As more laborers are using a fixed amount of capital, they get in each other's way, decreasing output.

5. If the marginal product (MP) of labor is greater than the average product (AP) of labor, then each additional unit of labor is more productive than the average of all previous units. By adding the last unit, the average of all units increases. The AP is at a maximum when the productivity of the last unit is equal to the average of all previous units.

8. a. Let Q_1 be the output of DISK, Inc., Q_2 be the output of FLOPPY, Inc., and X be equal amounts of capital and labor for the two firms. Then, $Q_1 = 10X^{0.5}X^{0.5} = 10X^{(0.5+0.5)} = 10X$ and $Q_2 = 10X^{0.6}X^{0.4} = 10X^{(0.6+0.4)} = 10X$. Because $Q_1 = Q_2$, they both generate the same output with the same inputs.

b. With capital fixed at 9 machine units, the production functions become $Q_1 = 30L^{0.5}$ and $Q_2 = 37.37L^{0.4}$. Consider the following table:

L	Q Firm 1	MP Firm 1	Q Firm 2	MP Firm 2
0	0	—	0	—
1	30.00	30.00	37.37	37.37
2	42.43	12.43	49.31	11.94
3	51.96	9.53	57.99	8.68
4	60.00	8.04	65.06	7.07

For each unit of labor above 1 unit, the marginal product of labor is greater for DISK, Inc.

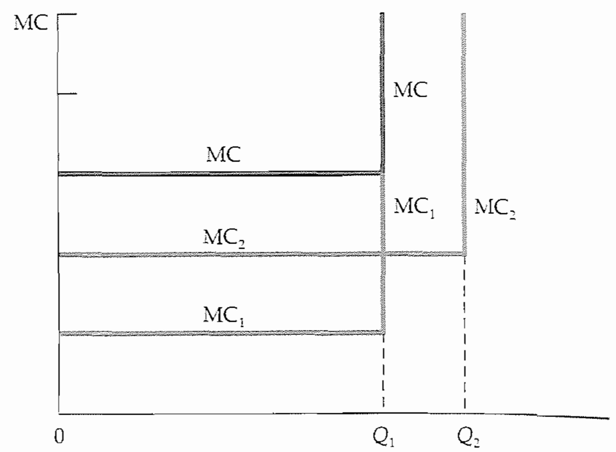

Figure 7

Chapter 7

3. a. Total cost, TC, is equal to fixed cost, FC, plus variable cost, VC. Since the franchise fee, FF, is a fixed sum, the firm's fixed costs increase by the fee. Then average cost, equal to $(FC + VC)/Q$, and average fixed cost, equal to (FC/Q), increase by the average franchise fee (FF/Q). Average variable cost is unaffected by the fee, as is marginal cost.
   b. When a tax t is imposed, variable costs increase by tQ. Average variable cost increases by t (fixed cost is constant), as does average (total) cost. Because total cost increases by t with each additional unit, marginal cost increases by t.

4. It is probably referring to accounting profit; this is the standard concept used in most discussions of how firms are doing financially. In this case, the article points to a substantial difference between accounting and economic profits. It claims that, under the current labor contract, automakers must pay many workers even if they are not working. This implies that their wages are *sunk* for the life of the contract. Accounting profits would subtract wages paid; economic profits would not, since they are sunk costs. Therefore automakers may be earning economic profits on these sales, even if they have accounting losses.

5. If the firm can produce one chair with either 4 hours of labor or 4 hours of machinery or any combination, then the isoquant is a straight line with a slope of -1 and intercepts at $K = 4$ and $L = 4$. The isocost line, $TC = 22L + 110K$, has a slope of $-1/5$ and intercepts at $K = TC/110$ and $L = TC/22$. The cost-minimizing point is a corner solution, where $L = 4$ and $K = 0$, and $TC = 88$.

8. The production of gasoline involves distilling crude oil and refining the distillate into gasoline. Given that the marginal cost of production is constant up to the capacity constraint for both processes, the marginal cost curves are "mirror" L-shapes. Total marginal cost $MC = MC_1 + MC_2$, where MC_1 is the marginal cost of distilling crude oil up to the capacity constraint Q_1, and MC_2 is the marginal cost of refining distillate up to the capacity constraint Q_2. If the capacity constraint of the distilling unit is lower than that of the hydrocracking unit, total MC is vertical at Q_1. (See Figure 7.) If the capacity constraint of the hydrocracking unit is lower than that of the distilling unit, total MC is vertical at Q_2.

Chapter 7—Appendix

1. a. Returns to scale refers to the relationship between output and proportional increases in all inputs. If $F(\lambda L,\lambda K) > \lambda F(L,K)$, there are increasing returns to scale; if $F(\lambda L,\lambda K) = \lambda F(L,K)$, there are constant returns to scale; if $F(\lambda L,\lambda K) < \lambda F(L,K)$, there are decreasing returns to scale. Applying this to $F(L,K) = K^2L$, $F(\lambda L,\lambda K) = (\lambda K)^2(\lambda L) = \lambda^3K^2L = \lambda^3F(L,K) > \lambda F(L,K)$. So, this production function exhibits increasing returns to scale.
   b. $F(\lambda L,\lambda K) = 10\lambda K + 5\lambda L = \lambda F(L,K)$. The production function exhibits constant returns to scale.
   c. $F(\lambda L,\lambda K) = (\lambda K\lambda L)^{0.5} = (\lambda^2)^{0.5} = (KL)^{0.5} = \lambda(KL)^{0.5} = \lambda F(L,K)$. The production function exhibits constant returns to scale.

2. The marginal product of labor is $100K$. The marginal product of capital is $100L$. The marginal rate of technical substitution is K/L. Set this equal to the ratio of the wage rate to the rental rate of capital: $K/L = 30/120$ or $L = 4K$. Then substitute for L in the production function and solve for a K that yields an output of 1000 units: $1000 = 100K \cdot 4K$. So, $K = 2.5^{0.5}$, $L = 4 \cdot 2.5^{0.5}$, and total cost is equal to \$379.20.

Chapter 8

1. The table below shows the firm's revenue and cost information when the price falls to \$35. At a price of \$35 the firm should produce 7 units to maximize profit.

3. a. Profit is maximized where marginal cost (MC) is equal to marginal revenue (MR). Here, MR is equal to \$60. Setting MC equal to 60 yields a profit-maximizing quantity of 30.
   b. Profit is equal to total revenue (PQ) minus total cost. So profit $= PQ - 100 - Q^2$. At $P = 60$ and $Q = 30$, profit $= 800$.
   c. The firm produces in the short run if its revenues are greater than its variable costs. The firm's short-run supply curve is its MC curve about minimum AVC. Here, AVC is equal to variable cost, Q^2, divided by quantity, Q. So, $AVC = Q$. Also, MC is equal to $2Q$. So, MC is greater than AVC for any quantity greater than 0. This means that the firm produces in the short run as long as price is positive.

5. a. With the imposition of a \$1 tax on a single firm, all its cost curves shift up by \$1.
   b. Because the firm is a price taker, the imposition of the tax on only one firm does not change the market price. Given that the firm's short-run supply curve is its marginal cost curve (above average variable cost), and that the marginal cost curve has shifted up (or inward), the firm supplies less to the market at every price.
   c. If the tax is placed on a single firm, that firm will go out of business unless it was earning a positive economic profit before the tax.

Chapter 9

1. a. In free-market equilibrium, $L^S = L^D$. Solving, $w = \$4$ and $L^S = L^D = 40$. If the minimum wage is \$5, then $L^S = 50$ and $L^D = 30$. The number of people employed will be given by the labor demand. So employers will hire 30 million workers.
   b. With the subsidy, only $w - 1$ is paid by the firm. The labor demand becomes $L^{D*} = 80 - 10(w - 1)$. So $w = \$4.50$ and $L = 45$.

4. a. Equating demand and supply, $28 - 2P = 4 + 4P$. $P^* = 4$ and $Q^* = 20$.
   b. The 25-percent reduction required by the new Payment-In-Kind Program would imply that farmers produce 15 billion bushels. To encourage farmers to withdraw their land from cultivation, the government must give them 5 billion bushels that they can sell on the market. Since the total supply to the market is still 20 billion bushels, the market price remains at \$4 per bushel. The farmers gain \$20 billion (\$4 × 5 billion bushels) from the PIK Program, while consumers are not affected.
   c. Taxpayers gain because the government does not have to pay to store the wheat for a year and then ship it to an underdeveloped country. The PIK Program can last only as long as wheat reserves last. But PIK assumes that the land removed from production can be restored to production at such times as the stockpiles are exhausted. If this cannot be done, consumers may eventually pay more for wheat-based products. Finally, farmers enjoy a windfall profit because they have no production costs.

9. The supply and demand curves for natural gas can be approximated as follows: $Q^S = 14 + 2P_G + 0.25P_O$, $Q^D = -5P_G + 3.75P_O$. With the price of oil at \$12 per barrel, these curves become $Q^S = 17 + 2P_G$ and $Q^D = 45 - 5P_G$. Setting $Q^D = Q^S$, $17 + 2P_G = 45 - 5P_G$, $P = \$4$. At this price, equilibrium quantity is 25 Tcf. If a ceiling of \$1 is imposed, producers would supply 19 Tcf and consumers would demand 40 Tcf. Consumers gain area A − B = 57 − 3.6 = \$53.4 billion in the figure. Producers lose area A − C = −57 − 9 = \$66.0 billion. Deadweight loss is equal to 53.4 − 66 = \$12.6 billion.

Output (units)	Price (\$/unit) P = 40	Revenue (\$) P = 40	Total Cost	Profit (\$) P = 40	Marginal Cost P = 40	Marginal Revenue P = 40	Revenue (\$/unit) P = 35	Marginal Revenue P = 35	Profit (\$) P = 35
0	40	0	50	−50	—	—	0	—	−50
1	40	40	100	−60	50	40	35	35	−65
2	40	80	128	−48	28	40	70	35	−58
3	40	120	148	−28	20	40	105	35	−43
4	40	160	162	−2	14	40	140	35	−22
5	40	200	180	20	18	40	175	35	−5
6	40	240	200	40	20	40	210	35	10
7	40	280	222	58	22	40	245	35	23
8	40	320	260	60	38	40	280	35	20
9	40	360	305	55	45	40	315	35	10
10	40	400	360	40	55	40	350	35	−10
11	40	440	425	15	65	40	385	35	−40

13. No, it would not. The clearest case is where labor markets are competitive. With either design of the tax, the wedge between supply and demand must total 12.4 percent of the wage paid. It does not matter whether the tax is imposed entirely on the workers (shifting the effective supply curve up by 12.4 percent) or entirely on the employers (shifting the effective demand curve down by 12.4 percent). The same applies to any combination of the two that sums to 12.4 percent.

Chapter 10

2. There are three important factors: (1) How similar are the products offered by Caterpillar's competitors? If they are close substitutes, a small increase in price could induce customers to switch to the competition. (2) What is the age of the existing stock of tractors? A 5-percent price increase induces a smaller drop in demand with an older population of tractors. (3) As a capital input in agricultural production, what is the expected profitability of the agricultural sector? If expected farm incomes are falling, an increase in tractor prices induces a greater decline in demand than one would estimate with information on past sales and prices.

4. a. Optimal production is found by setting marginal revenue equal to marginal cost. If the demand function is linear, $P = a - bQ$ (here, $a = 100$ and $b = 0.01$), so that $MR = a - 2bQ = 100 - 2(0.01)Q$. Total cost $= 30{,}000 + 50Q$, so $MC = 50$. Setting MR = MC implies $100 - 0.02Q = 50$, so $Q = 2500$. Substituting into the demand function, $P = 100 - 0.01 \cdot 2500 = 75$ cents. Total profit is $-30{,}000 + 50 \cdot 2500 - 0.01 \cdot (2500)^2 = -30{,}000 + 125{,}000 - 62{,}500$, or \$325 per week.
   b. Suppose initially that the consumers must pay the tax. Since the price (including the tax) that consumers would be willing to pay remains unchanged, the demand function can be written $P + t = 100 - 0.01Q - t$. Because the tax increases the price of each unit, total revenue for the monopolist decreases by tQ, and marginal revenue decreases by t: $MR = 100 - 0.02Q - t$, where $t = 10$ cents. To determine the profit-maximizing output with tax, equate marginal revenue and marginal cost: $100 - 0.02Q - 10 = 50$, or $Q = 2000$ units.

   From the demand function, average revenue $= 100 - 0.01 \cdot 2000 - 10 = 70$ cents. Total profit is $70 \cdot (2000) - (30{,}000 + 50(2000)) = 10{,}000$ cents or \$100 per week.

8. a. Pro: Although Alcoa controlled about 90 percent of primary aluminum production in the United States, secondary aluminum production by recyclers accounted for 30 percent of the total aluminum supply. It should be possible for a much larger proportion of aluminum supply to come from secondary sources. Therefore the price elasticity of demand for

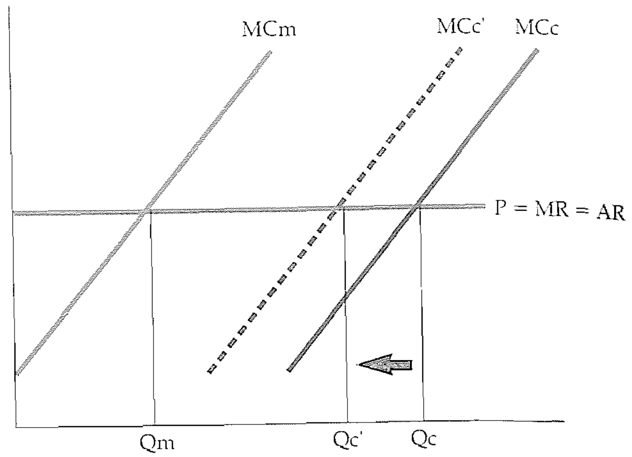

Figure 10

Alcoa's primary aluminum is much higher than we would expect. In many applications, other metals, such as copper and steel, are feasible substitutes for aluminum. Here, the demand elasticity Alcoa faces may be lower than we would otherwise expect.
   b. Con: The stock of potential supply is limited. Therefore, by keeping a stable high price, Alcoa could reap monopoly profits. Furthermore, since Alcoa had originally produced the metal reappearing as recycled scrap, it would have taken into account in its output decisions the effect of scrap reclamation on future prices. Hence, it exerted effective monopolistic control over the secondary metal supply.
   c. Alcoa was not ordered to sell any of its U.S. production facilities. Rather, (1) it was barred from bidding for two primary aluminum plants constructed by the government during World War II; and (2) it was ordered to divest itself of its Canadian subsidiary, which became Alcan.

11. No, you should not. In a competitive market, a firm views price as being horizontal and equal to average revenue, which is equal to marginal revenue. If Connecticut's marginal cost increases, price will still be equal to Massachusetts's marginal cost, total marginal cost, and marginal revenue. Only Connecticut's quantity is reduced (which, in turn, reduces overall quantity), as shown in Figure 10 above.

Chapter 11

1. a. The Saturday-night requirement separates business travelers, who prefer to return home for the weekend, from tourists, who travel on the weekend.
   b. By basing prices on the buyer's location, sorting is done by geography. Then prices can reflect transportation charges, which the customer pays for whether delivery is received at the buyer's location or at the cement plant.
   c. Rebate coupons with food processors separate consumers into two groups: (1) customers who are less price sensitive (those who have a lower elasticity of demand) do not request the rebate; and (2) customers who are more price sensitive (those who have a higher demand elasticity) request the rebate.
   d. A temporary price cut on bathroom tissue is a form of intertemporal price discrimination. Price-sensitive customers buy more tissue than they would otherwise during the price cut, while non-price-sensitive consumers buy the same amount.
   e. The plastic surgeon can distinguish a high-income patient from a low-income patient by negotiation. Arbitrage is no problem because plastic surgery cannot be transferred from low-income patients to high-income patients.

8. a. A monopolist with two markets should pick quantities in each market so that the marginal revenues in both markets are equal to one another and equal to marginal cost. Marginal cost is the slope of the total cost curve, 30. To determine marginal revenues in each market, we solve for price as a function of quantity. Then we substitute this expression for price into the equation for total revenue. $P_{NY} = 150 - 3Q_{NY}$, and $P_{LA} = 120 - (3/2)Q_{LA}$. Then total revenues are $TR_{NY} = Q_{NY}P_{NY} = Q_{NY}(150 - 3Q_{NY})$, and $TR_{LA} = Q_{LA}P_{LA} = Q_{LA}(120 - (3/2)Q_{LA})$. The marginal revenues are the slopes of the total revenue curves: $MR_{NY} = 150 - 6Q_{NY}$ and $MR_{LA} = 120 - 3Q_{LA}$. Next, we set each marginal revenue to marginal cost (= 30), implying $Q_{NY} = 20$ and $Q_{LA} = 30$. With these quantities, we solve for price in each market: $P_{NY} = 150 - 3 \cdot 20 = 90$ and $P_{LA} = 120 - (3/2) \cdot 30 = 75$.
   b. With the new satellite, Sal can no longer separate the two markets. The total demand function is the horizontal summation of the two markets. Above a price of 120, the total demand is just the New York demand function. Below a price of 120, we add the two demands: $Q_T = 50 - (1/3)P + 80 - (2/3)P = 130 - P$. Sal maximizes profit by choosing a quantity so that MR = MC. Total revenue is QP, where $P = 130 - Q$. $TR = Q(130 - Q)$, so marginal revenue is $130 - 2Q$. Setting this equal to marginal cost implies a profit-maximizing quantity of 50 with a price of 80. In the New York market, quantity is equal to $50 - (1/3)80 = 23\frac{1}{3}$, and in the Los Angeles market, quantity is equal to $80 - (2/3)80 = 26\frac{2}{3}$. Together, 50 units are purchased at a price of 80.
   c. Under the market conditions in (a) profit is equal to the sum of revenues from each market minus the cost of producing quantity for both markets: $Q_{NY}P_{NY} + Q_{LA}P_{LA} - 1000 - 30(Q_{NY} + Q_{LA}) = 20 \cdot 90 + 30 \cdot 75 - 1000 + 30(20 + 30) = 1550$. Under the market conditions in (b) profit is equal to the total revenue minus the cost of producing quantity for both markets: $QP - (1000 + 30Q) = 50 \cdot 80 - (1000 + 30 \cdot 50) = 1500$. So Sal makes more money when the two markets are separated.

   Under the market conditions in (a), in the New York market consumer surplus is $(150 - 90) \cdot 20(1/2) = 600$, and in the Los Angeles market consumer surplus is $(120 - 75) \cdot 30(1/2) = 675$. Under the conditions in (b), in the New York market consumer surplus is $(150 - 80) \cdot 23\frac{1}{3}(1/2) = 817$, and in the Los Angeles market consumer surplus is $(120 - 80) \cdot 26\frac{2}{3}(1/2) = 533$. The New Yorkers prefer (b) because the equilibrium price is 80 instead of 90, so their consumer surplus is higher. But the customers in Los Angeles prefer (a) because the equilibrium price is 75 instead of 80.

10. a. With individual demands of $Q_1 = 6 - P$, individual consumer surplus is equal to \$18 per week, or \$936 per year. An entry fee of \$936 captures all consumer surplus, even though no court fee would be charged, since marginal cost is equal to zero. Weekly profits would be equal to the number of serious players, 1000, times the weekly entry fee, \$18, minus \$5000, the fixed costs, or \$13,000 per week.
   b. When there are two classes of customers, the club owner maximizes profits by charging court fees above marginal cost and by setting the entry fee equal to the remaining consumer surplus of the consumer with the smaller demand—the occasional player. The entry fee, T, is equal to the consumer surplus remaining after the court fee is assessed: $T = (Q_2 - 0)(6 - P)(1/2)$, where $Q_2 = 3 - (1/2)P$, or $T = [3 - (1/2)P](6 - P)(1/2) = 9 - 3P + P^2/4$. Entry fees for all players would be $2000(9 - 3P + P^2/4)$. Revenues from court fees equals $P(Q_1 + Q_2) = P[1000(6 - P) + 1000(3 - P/2)] = 9000P - 1500P^2$. Then total revenue $= TR = 2000[(9 - 3P + P^2/4) + 9000P - 1500P^2 = 18{,}000 + 3000P - 1000P^2$. Marginal cost is zero, and marginal revenue is given by the slope of the total revenue curve: $\Delta TR/\Delta P = 3000 - 2000P$. Equating marginal revenue and marginal cost implies a price of \$1.50 per hour. Total revenue is equal to \$20,250. Total cost is equal to fixed costs of \$5000. So, profit is \$15,250 per week, which is greater than the \$13,000 per week when only professional players become members.
   c. An entry fee of \$18 per week would attract only serious players. With 3000 serious players, total revenues would be \$54,000, and profits would be \$49,000 per week. With both serious and occasional players, entry fees would be equal to 4000 times the consumer surplus of the occasional player: $T = 4000(9 - 3P + P^2/4)$. Court fees are $P[(6 - P)3000 +$

$(3 - P/2)1000] = (21P - 3.5P^2)1000$. Then TR = $[4(9 - 3P + P^2/4) + (21P - 3.5P^2)]1000 = (36 + 9P - 2.5P^2)1000$. Equating marginal revenue and marginal cost implies a price of (9/5), or \$1.80 per hour. Then total revenue is equal to \$44,100. Total cost is equal to fixed costs of \$5000. Profit with a two-part tariff is \$39,100 per week, which is less than the \$49,000 per week with only professional players. The club owner should set annual dues at \$936 and earn profits of \$2.548 million per year.

11. Mixed bundling is often the ideal strategy when demands are only somewhat negatively correlated and/or when marginal production costs are significant. The following tables present the reservation prices of the three consumers and the profits from the three strategies:

	Reservation Price		
	For 1	For 2	Total
Consumer A	\$ 3.25	\$ 6.00	\$ 9.25
Consumer B	8.25	3.25	11.50
Consumer C	10.00	10.00	20.00

	Price 1	Price 2	Bundled	Profit
Sell separately	\$ 8.25	\$6.00	—	\$28.50
Pure bundling	—	—	\$ 9.25	27.75
Mixed bundling	10.00	6.00	11.50	29.00

The profit-maximizing strategy is to use mixed bundling.

15. a. For each strategy, the optimal prices and profits are

	Price 1	Price 2	Bundled	Profit
Sell separately	\$40.00	\$40.00	—	\$240.00
Pure bundling	—	—	\$100.00	400.00
Mixed bundling	59.95	59.95	100.00	319.90

Pure bundling dominates mixed bundling because with marginal costs of zero, there is no reason to exclude purchases of both goods by all customers.

b. With marginal cost of \$35, the optimal prices and profits are

	Price 1	Price 2	Bundled	Profit
Sell separately	\$90.00	\$90.00	—	\$110.00
Pure bundling	—	—	\$100.00	120.00
Mixed bundling	59.95	59.95	100.00	110.00

Pure bundling still dominates all other strategies.

Chapter 11—Appendix

1. We examine each case, then compare profits.

a. Optimal quantities and prices with no external market for engines are $Q_E = Q_A = 2000$, $P_E = \$8000$, and $P_A = \$18{,}000$. For the engine-building division, TR = 2000 · \$8000 = \$16M, TC = $2(2000)^2$ = \$8M, and π = \$8M. For the automobile-assembly division, TR = 2000 · \$18,000 = \$36M, TC = \$8000 · 2000 + 16M = \$32M, and π = \$4M. Total profits are \$12M.

b. Optimal quantities and prices with an external market for engines are $Q_E = 1500$, $Q_A = 3000$, $P_E = \$6000$, and $P_A = \$17{,}000$. For the engine-building division, TR = 1500 · \$6000 = \$9M, TC = $2(1500)^2$ = \$4.5M, and π = \$4.5M. For the automobile-assembly division, TR = 3000 · \$17,000 = \$51M, TC = (8000 + 6000)3000 = \$42M, and π = \$9M. Total profits are \$13.5M.

c. Optimal quantities and prices with a monopoly market for engines are $Q_E = 2200$, $Q_A = 1600$, $P_E = \$8800$, and $P_A = \$18{,}400$, with 600 engines sold in the monopolized market for \$9400. For the engine-building division, TR = 1600 · \$8800 + 600 · 9400 = \$19.72M, TC = $2(2200)^2$ = \$9.68M, and π = \$10.04M. For the automobile-assembly division, TR = 1600 · \$18,400 = TR = 1600 · \$18,400 = \$29.44M, TC = (8000 + 8800)1600 = \$26.88M, and π = \$2.56M. Total profits are \$12.6M.

The upstream division, building engines, earns the most profit when it has a monopoly on engines. The downstream division, building automobiles, earns the most when there is a competitive market for engines. Given the high cost of engines, the firm does best when engines are produced at the lowest cost with an external, competitive market for engines.

Chapter 12

1. Each firm earns economic profit by distinguishing its brand from all other brands. If these competitors merge into a single firm, the resulting monopolist would not produce as many brands as would have been produced before the merger. But, producing several brands with different prices and characteristics is one method of splitting the market into sets of customers with different price elasticities.

3. a. To maximize profit $\pi = 53Q - Q^2 - 5Q$, we find $\Delta\pi/\Delta Q = -2Q + 48 = 0$. $Q = 24$, so $P = 29$. Profit is equal to 576.

b. $P = 53 - Q_1 - Q_2$, $\pi_1 = PQ_1 - C(Q_1) = 53Q_1 - Q_1^2 - Q_1Q_2 - 5Q_1$ and $\pi_2 = PQ_2 - C(Q_2) = 53Q_2 - Q_1Q_2 - Q_2^2 - 5Q_2$.

c. The problem facing Firm 1 is to maximize profit, given that the output of Firm 2 will not change in reaction to the output decision of Firm 1. Therefore, Firm 1 chooses Q_1 to maximize π_1, as above. The change in π_1 with respect to a change in Q_1 is $53 - 2Q_1 - Q_2 - 5 = 0$, implying $Q_1 = 24 - Q_2/2$. Since the problem is symmetric, the reaction function for Firm 2 is $Q_2 = 24 - Q_1/2$.

d. Solve for the values of Q_1 and Q_2 that satisfy both reaction functions: $Q_1 = 24 - (1/2)(24 - Q_1/2)$. So, $Q_1 = 16$ and $Q_2 = 16$. The price is $P = 53 - Q_1 - Q_2 = 21$. Profit is $\pi_1 = \pi_2 = P \cdot Q_i - C(Q_i) = 256$. Total profit in the industry is $\pi_1 + \pi_2 = 512$.

5. *True.* The reaction curve of Firm 2 will be $q_2 = 7.5 - 1/2\,q_1$ and the reaction curve of Firm 1 will be $q_1 = 15 - 1/2q_2$. Substituting yields $q_2 = 0$ and $q_1 = 15$. The price will be 15, which is the monopoly price.

11. a. To determine the Nash equilibrium we calculate the reaction function for each firm, then simultaneously solve for price. *Assuming marginal cost is zero,* profit for Firm 1 is $P_1Q_1 = P_1(20 - P_1 + P_2) = 20P_1 - P_1^2 + P_2P_1$. $MR_1 = 20 - 2P_1 + P_2$. At the profit-maximizing price, $MR_1 = 0$. So, $P_1 = (20 + P_2)/2$. Because Firm 2 is symmetric to Firm 1, its profit-maximizing price is $P_2 = (20 + P_1)/2$. We substitute Firm 2's reaction function into that of Firm 1: $P_1 = [20 + (20 + P_1)/2]/2 = 15 + P_1/4$. $P_1 = 20$. By symmetry $P_2 = 20$. Then $Q_1 = 20$, and by symmetry $Q_2 = 20$. Profit for Firm 1 is $P_1Q_1 = 400$, and profit for Firm 2 is also 400.

b. If Firm 1 sets its price first, it takes Firm 2's reaction function into account. Firm 1's profit is $\pi_1 = P_1[20 - P_1 + (20 + P_1)/2]$. Then, $d\pi_1/dP_1 = 20 - 2P_1 + 10 + P_1$. Setting this expression equal to zero, $P_1 = 30$. We substitute for P_1 in Firm 2's reaction function, $P_2 = 25$. At these prices, $Q_1 = 20 - 30 + 25 = 15$ and $Q_2 = 20 + 30 - 25 = 25$. Profit is $\pi_1 = 30 \cdot 15 = 450$ and $\pi_2 = 25 \cdot 25 = 625$.

c. Your first choice should be (iii), and your second choice should be (ii). Setting prices above the Cournot equilibrium values is optional for both firms when Stackelberg strategies are followed. From the reaction functions, we know that the price leader provokes a price increase in the follower. But the follower increases price less than the price leader, and hence undercuts the leader. Both firms enjoy increased profits, but the follower does best, and both do better than they would in the Cournot equilibrium.

Chapter 13

1. If games are repeated indefinitely and all players know all payoffs, rational behavior will lead to apparently collusive outcomes. But, sometimes the payoffs of other firms can only be known by engaging in extensive information exchanges.

Perhaps the greatest problem to maintaining a collusive outcome is exogenous changes in demand and in the prices of inputs. When new information is not available to all players simultaneously, a rational reaction by one firm could be interpreted as a threat by another firm.

2. Excess capacity can arise in industries with easy entry and differentiated products. Because downward-sloping demand curves for each firm lead to outputs with average cost above minimum average cost, increases in output result in decreases in average cost. The difference between the resulting output and the output at minimum long-run average cost is excess capacity, which can be used to deter new entry.

4. a. There are two Nash equilibria: (100,800) and (900,600).

b. Both managers will follow a high-end strategy, and the resulting equilibrium will be (50,50), yielding less profit to both parties.

c. The cooperative outcome (900,600) maximizes the joint profit of the two firms.

d. Firm 1 benefits the most from cooperation. Compared to the next best opportunity, Firm 1 benefits by 900 − 100 = 800, whereas Firm 2 loses 800 − 600 = 200 under cooperation. Therefore, Firm 1 would need to offer Firm 2 at least 200 to compensate for Firm 2's loss.

6. a. Yes, there are two: (1) Given Firm 2 chooses *A*, Firm 1 chooses *C*; given Firm 1 chooses *C*, Firm 2 chooses *A*. (2) Given Firm 2 chooses *C*, Firm 1 chooses *A*; given Firm 1 chooses *A*, Firm 2 chooses *C*.

b. If both firms choose according to maximin, Firm 1 will choose Product *A* and Firm 2 will choose Product *A*, resulting in −10 payoff for both.

c. Firm 2 will choose Product *C* in order to maximize payoffs at 10, 20.

12. Although antique auctions often have private-value elements, they are primarily common value because dealers are involved. Our antique dealer is disappointed in the nearby town's public auction because estimates of the value of the antiques vary widely and she has suffered from the winner's curse. At home, where there are fewer well-informed bidders, the winner's curse has not been a problem.

Chapter 14

1. The budget line for workers under this program is a straight line at \$5000. There is no incentive to work under the new program. Only wages yielding incomes greater than \$10,000 will yield a positive labor supply.

5. The demand for labor is given by the marginal revenue product of labor; $MRP_L = MR \cdot MP_L$. In a competitive market, price is equal to marginal revenue, so MR = 10. The marginal product of labor is equal to the slope of the production function $Q = 12L - L^2$. This slope is equal to $12 - 2L$. The firm's profit-maximizing quantity of labor occurs where $MRP_L = w$, the wage rate. If $w = 30$, solving for L yields 4.5 hours per day. Similarly, if $w = 60$, solving for L yields 3 hours per day.

8. Economic rent is the difference between total wages for all employed workers minus the amount that would have induced these workers to work. Total wages are equal to $wL_D = 1200w - 10w^2$. The total income that workers would have accepted is the area under the labor supply curve up to the labor demanded at w. From the supply function, we know that $L = 20w$, or $w = (1200 - 10w)/20$. Then this area is a triangle equal

to $L_D \cdot w_s \cdot 1/2 = (1200 - 10w)[(1200 - 10w)/20] \cdot 1/2 = 36{,}000 - 600w + 2.5w^2$. If the union's goal is to maximize rent, then it would choose a w to maximize $(1200w - 10w^2) - (36{,}000 - 600w + 2.5w^2) = -36{,}000 = 1800w - 12.5w^2$. The slope with respect to w is $1800 - 25w$. The maximum occurs where this slope is equal to zero, or $w = 72$. At a wage rate of \$72, 480 union members are employed. They would have been willing to work for a total income of \$5760 $(0.5 \cdot 480 \cdot 480/20)$. They receive \$34,560 and enjoy economic rents of \$28,800 (34,560 − 5760).

Chapter 15

1. The present discounted value of the first \$80 payment one year from now is PDV $= 80/(1 + 0.10)^1 = \$72.73$. The value of all these coupon payments can be found the same way: PDV $= 80[1/(1.10)^1 + 1/(1.10)^2 + 1/(1.10)^3 + 1/(1.10)^4 + 1/(1.10)^5] = \$303.26$. The present value of the final payment of \$1000 in the sixth year is $1000/1.1^6 = \$564.47$. So the present value of this bond is \$303.26 + \$564.47 = \$867.73. With an interest rate of 15 percent, PDV = \$700.49.

3. Redefining terms, the net present value equation becomes $\text{NPV} = -5 - 5(1 + R)^{-1} - 1(1 + R)^{-2} - 0.5(1 + R)^{-3} + 0.96[(1 + R)^{-4} + (1 + R)^{-5} + (1 + R)^{-6} + (1 + R)^{-7}] + 0.96[(1 + R^{-8}) + (1 + R)^{-9} + (1 + R)^{-10} + (1 + R)^{-11}] + 0.96[(1 + R)^{-12} + (1 + R)^{-13} + (1 + R)^{-14} + (1 + R)^{-15}] + 0.96[(1 + R)^{-16} + (1 + R)^{-17} + (1 + R)^{-18} + (1 + R)^{-19}] + 0.96(1 + R)^{-20} + 1(1 + R)^{-20}$. With an interest of 4 percent, the NPV becomes $-5 - 4.8075 - 0.9246 - 0.4445 + 3.0978 + 2.6482 + 2.2637 + 1.9349 + 0.4381 + 0.4564 = -\$320{,}000$. The investment is not worthwhile.

7. a. If we buy a bottle and sell it after t years, we pay \$100 now and receive $100t^{0.5}$ when it is sold. The NPV of this investment is $\text{NPV} = -100 + e^{-rt}100t^{0.5} = -100 + e^{-0.1t}100t^{0.5}$.

   If we do buy a bottle, we will choose t to maximize the NPV. The necessary condition is $d\text{NPV}/dt = e^{-0.1t}(50 - t^{-0.5}) - 0.1e^{-0.1t} \cdot 100t^{0.5} = 0$. Solving, $t = 5$. If we hold the bottle 5 years, the NPV is $-100 + e^{-0.1 \cdot 5}100 \cdot 5^{0.5} = 35.62$. Since each bottle is a good investment, we should buy all 100 bottles.

   b. You get \$130 now, but lose the $\$100 \cdot 5^{0.5}$ you would get for selling in five years. The NPV of the offer is $\text{NPV} = 130 - (e^{(0.1)(5)})(100)(5^{(0.5)}) = -239 < 0$. Therefore, you should not sell.

   c. If the interest rate changes from 10 percent to 5 percent the NPV calculation changes to $\text{NPV} = -100 + e^{-0.05t} \cdot 100t^{0.5}$. If we hold the bottle 10 years, the maximum NPV is $-100 + e^{-0.05 \cdot 10} \cdot 100 \cdot 10^{0.5} = \$91.80$.

9. a. Compare buying the car to leasing the car, with $r = 0.04$. The present value net cost of buying is $-15{,}000 + 6000/(1 + 0.04)^3 = -9666.02$. The present value cost of leasing the car is $-3600/(1 + 0.04) - 3600/(1 + 0.04)^2 - 3600/(1 + 0.04)^3 = -9990.33$. You are better off buying the car if $r = 4$ percent.

   b. Again, compare buying to leasing: $-15{,}000 + 6000/(1 + 0.12)^3 = -10{,}729.32$ with buying, versus $-3600/(1 + 0.12) - 3600/(1 + 0.12)^2 - 3600/(1 + 0.12)^3 = -8{,}646.6$ with leasing. You are better off leasing the car if $r = 12$ percent.

   c. Consumers will be indifferent when the present value cost of buying and later selling the car equals the present value cost of leasing: $-15{,}000 + 6000/(1 + r)^3 = -3600/(1 + r) - 3600/(1 + r)^2 - 3600/(1 + r)^3$. This is true when $r = 4.96$ percent. You can solve this equation using a graphing calculator or computer spreadsheet, or by trial and error.

Chapter 16

1. Even with identical preferences, the contract curve may or may not be a straight line. This can easily be shown graphically. For example, when both individuals have utility functions $U = x^2y$, the marginal rate of substitution is given by $2y/x$. It is not difficult to show that the MRS's of both individuals are equal for all points on the contract curve $y = (Y/X)/x$, where X and Y are the total quantities of both goods. One example in which the contract curve is not a straight line is when the two individuals have different incomes and one good is inferior.

2. The marginal rate of transformation is equal to the ratio of the marginal costs of producing the two goods. Most production possibilities frontiers are "bowed outward." However, if the two goods are produced with identical production functions, the production possibilities frontier is a straight line.

6. A change from a constant-returns-to-scale production process to a sharply-increasing-returns-to-scale process does not imply a change in the shape of the isoquants. One can simply redefine the quantities associated with each isoquant such that proportional increases in inputs yield greater than proportional increases in outputs. Under this assumption, the marginal rate of technical substitution would not change, and there would be no change in the production contract curve.

Chapter 17

5. a. In the recent past, American automobiles appeared to customers to be of low quality. To reverse this trend, American companies invested in quality control, improving the potential repair records of their products. They signaled the improved quality of their products through improved warranties.

   b. Moral hazard occurs when the party to be insured (the owner of an American automobile with an extensive warranty) can influence the probability or the magnitude of the event that triggers payment (the repair of the automobile). Covering all parts and labor associated with mechanical problems reduces the incentive to maintain the automobile. Hence, a moral hazard problem is created with extensive warranties.

Chapter 18

4. One needs to know the value to homeowners of swimming in the river, and the marginal cost of abatement. The choice of a policy tool will depend on the marginal benefits and costs of abatement. If firms are charged an equal rate effluent fee, the firms will reduce effluent to the point where the marginal cost of abatement is equal to the fee. If this reduction is not high enough to permit swimming, the fee could be increased.

   The setting of a standard will be efficient only if the policy maker has complete information regarding the marginal costs and benefits of abatement. Further, the standard will not encourage firms to reduce effluent further if new filtering technologies become available. A transferable effluent permit system still requires the policymaker to determine the efficient effluent standard. Once the permits are distributed, a market will develop and firms with a higher cost of abatement will purchase permits from firms with lower abatement costs. However, unless permits are sold initially, no revenue will be generated.

6. a. Profit is maximized when marginal revenue is equal to marginal cost. With a constant marginal revenue of \$20 and a marginal cost of $10 + 2Q$, $Q = 5$.

   b. If bees are not forthcoming, the farmer must pay \$10 per acre for artificial pollination. Since the farmer would be willing to pay up to \$10 to the beekeeper to maintain each additional hive, the marginal social benefit of each is \$30, which is greater than the marginal private benefit of \$20. Equating the marginal social benefit to the marginal cost, $Q = 10$.

   c. The most radical change that would lead to more efficient operations would be the merger of the farmer's business with the beekeeper's business. This merger would internalize the positive externality of bee pollination. Short of a merger, the farmer and beekeeper should enter into a contract for pollination services.

INDEX

C

F

G

H

I

J

K

L

M

N

O

P