CHAPTER 6

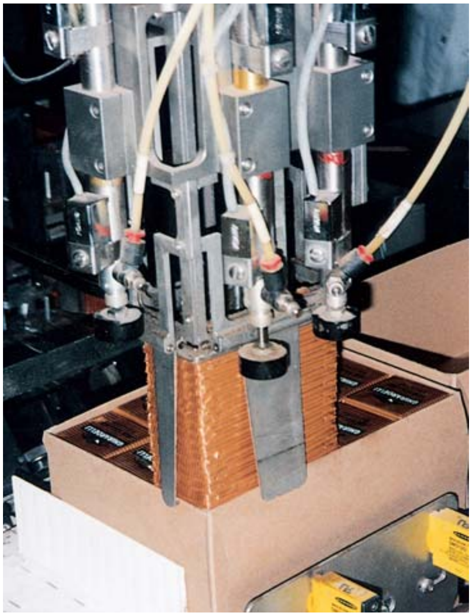
Firms and Production

Hard work never killed anybody, but why take a chance?

—Charlie McCarthy

The Ghirardelli Chocolate Company converts chocolate and other inputs into an output of 144,000 wrapped chocolate bars and 340,000 wrapped chocolate squares a day. The *material inputs* include chocolate, other food products, and various paper goods for wrapping and boxing the candy. The *labor inputs* include chefs, assembly-line workers, and various mechanics and other technicians. The *capital inputs* are the manufacturing plant, the land on which the plant is located, conveyor belts, molds, wrapping machines, and various other types of equipment.

Over time, Ghirardelli has changed how it produces its finished product, increasing the ratio of machines to workers. Several years ago, to minimize employees' risk of repetitive motion injuries, the company spent $300,000 on robots, which pack the wrapped chocolate and put it on pallets. The use of robotic arms resulted in greatly reduced downtime, increased production, and improved working conditions.



By using robotic equipment to pack finished, wrapped chocolate, the Ghirardelli Chocolate Company benefits from reduced downtime and increased production.

This chapter looks at the types of decisions that the owners of firms have to make. First, a decision must be made as to how a firm is owned and managed. Ghirardelli, for example, is a corporation—it is not owned by an individual or partners—and is run by professional managers. Second, the firm must decide how to produce. Ghirardelli now uses relatively more machines and robots and fewer workers than in the past. Third, if a firm wants to expand output, it must decide how to do that in both the short run and the long run. In the short run, Ghirardelli can expand output by extending the workweek to six or seven days and using extra materials. To expand output more, Ghirardelli would have to install more equipment (such as extra robotic arms), hire more workers, and eventually build a new plant, all of which take time. Fourth, given its ability to change its output level, a firm must determine how large to grow. Ghirardelli determines its current investments on the basis of its beliefs about demand and costs in the future.

In this chapter, we examine how a firm chooses its inputs so as to produce efficiently. In Chapter 7, we examine how the firm chooses the least costly among all possible efficient production processes. In Chapter 8, we combine this information about costs with information about revenues to determine how a firm picks the output level that maximizes profit.

The main lesson of this chapter and the next is that firms are not black boxes that mysteriously transform inputs (such as labor, capital, and material) into outputs. Economic theory explains how firms make decisions about production processes, types of inputs to use, and the volume of output to produce.

In this chapter, we examine six main topics

1. **The ownership and management of firms:** Decisions must be made as to how a firm is owned and run.
2. **Production:** A firm converts inputs into outputs using one of possibly many available technologies.
3. **Short-run production: one variable and one fixed input:** In the short run, only some inputs can be varied, so the firm changes its output by adjusting its variable inputs.
4. **Long-run production: two variable inputs:** The firm has more flexibility in how it produces and how it changes its output level in the long run when all factors can be varied.
5. **Returns to scale:** How the ratio of output to input varies with the size of the firm is an important factor in determining the size of a firm.
6. **Productivity and technical change:** The amount of output that can be produced with a given amount of inputs varies across firms and over time.

6.1 THE OWNERSHIP AND MANAGEMENT OF FIRMS

A **firm** is an organization that converts *inputs* such as labor, materials, and capital into *outputs*, the goods and services that it sells. U.S. Steel combines iron ore, machinery, and labor to create steel. A local restaurant buys raw food, cooks it, and serves it. A landscape designer hires gardeners and machines, buys trees and shrubs, transports them to a customer's home, and supervises the work.

Most goods and services produced in Western countries are produced by firms. In the United States, firms produce 82% of national production (U.S. gross domestic product); the government, 12%; and nonprofit institutions (such as some universities and hospitals) and households, 6% (*Survey of Current Business*, 2002). In developing countries, the government's share of total national production can be much higher, reaching 37% in Ghana, 38% in Zambia, 40% in Sudan, and 90% in Algeria, though it is as low as 3% in Bangladesh, Paraguay, and Nepal (United Nations, *Industry and Development: Global Report 1992/93*). In this book, we focus on production by for-profit firms rather than by nonprofit organizations and governments.

The Ownership of Firms

In most countries, for-profit firms have one of three legal forms: sole proprietorships, partnerships, and corporations.

Sole proprietorships are firms owned and run by a single individual.

Partnerships are businesses jointly owned and controlled by two or more people. The owners operate under a partnership agreement. If any partner leaves, the partnership agreement ends. For the firm to continue to operate, a new partnership agreement must be written.

Corporations are owned by *shareholders* in proportion to the numbers of shares of stock they hold. The shareholders elect a board of directors who run the firm. In turn, the board of directors usually hires managers who make short-term decisions and long-term plans.

Corporations differ from the other two forms of ownership in terms of personal liability for the debts of the firm. Sole proprietors and partners are personally responsible for the debts of their firms. All of an owner's personal wealth—not just that invested in the firm—is at risk if the business becomes bankrupt and is unable to pay its bills. Even the assets of partners who are not responsible for the failure can be taken to cover the firm's debts.

Corporations have **limited liability**: The personal assets of the corporate owners cannot be taken to pay a corporation's debts if it goes into bankruptcy. Because of the limited liability of corporations, the most that shareholders can lose if the firm goes bankrupt is the amount they paid for their stock, which becomes worthless if the corporation fails. Sole proprietors have unlimited liability—that is, even their personal assets can be taken to pay the firm's debts. Partners share liability: Even the assets of partners who are not responsible for the failure can be taken to cover the firm's debts. General partners can manage the firm but have unlimited liability. Limited partners are prohibited from managing but are liable only to the extent of their investment in the business.[1]

In the United States, 87% of business sales are made by corporations, even though fewer than 20% of all firms are corporations. Nearly 72% of all firms are sole proprietorships. Sole proprietorships tend to be small, however, so they are responsible for only 5% of all sales. Partnerships account for 8% of all firms and make 8% of sales (*Statistical Abstract of the United States*, 2001).

The Management of Firms

In a small firm, the owner usually manages the firm's operations. In larger firms, typically corporations and larger partnerships, a manager or team of managers usually runs the company. In such firms, owners, managers, and lower-level supervisors are all decision makers.

A recent revelations about Enron and WorldCom illustrate, the various decision makers may have conflicting objectives. What is in the best interest of the owners may not be the same as what is in the best interest of managers or other employees.

---

[1]Due to changes in corporate and tax laws over the last decade, *limited liability companies* (LLCs) have become common in the United States. Owners are liable only to the extent of their investment (as in a corporation) and can play an active role in management (as in a partnership or sole proprietorship). When an owner leaves, the LLC does not have to dissolve as with a partnership.

For example, a manager may want a fancy office, a company car, a company jet, and other perks, but the owner would likely oppose these drains on profit.

The owner replaces the manager if the manager pursues personal objectives rather than the firm's objectives. In a corporation, the board of directors is supposed to ensure that managers do not stray. If the manager and the board of directors run the firm badly, the shareholders can fire both or directly change some policies through votes at the corporation's annual meeting for shareholders. Until Chapter 20, we'll ignore the potential conflict between managers and owners and assume that the owner *is* the manager of the firm and makes all the decisions.

What Owners Want

Economists usually assume that a firm's owners try to maximize profit. Presumably, most people invest in a firm to make money—lots of money, they hope. They want the firm to earn a positive profit rather than make a loss (a negative profit). A firm's **profit**, π, is the difference between its revenue, R, which is what it earns from selling the good, and its cost, C, which is what it pays for labor, materials, and other inputs:

$$\pi = R - C.$$

Typically, revenue is p, the price, times q, the firm's quantity: $R = pq$.

In reality, some owners have other objectives, such as having as big a firm as possible or a fancy office or keeping risks low. In Chapter 8, however, we show that a competitive firm is likely to be driven out of business if it doesn't maximize profits.

To maximize profits, a firm must produce as efficiently as possible. A firm engages in **efficient production** (achieves **technological efficiency**) if it cannot produce its current level of output with fewer inputs, given existing knowledge about technology and the organization of production. Equivalently, the firm produces efficiently if, given the quantity of inputs used, no more output could be produced using existing knowledge.

If the firm does not produce efficiently, it cannot be profit maximizing—so efficient production is a *necessary condition* for profit maximization. Even if a firm produces a given level of output efficiently, it is not maximizing profit if that output level is too high or too low or if it is using excessively expensive inputs. Thus efficient production alone is not a *sufficient condition* to ensure that a firm's profit is maximized.

A firm may use engineers and other experts to determine the most efficient ways to produce with a known method or technology. However, this knowledge does not indicate which of the many technologies, each of which uses different combinations of inputs, allows for production at the lowest cost or with the highest possible profit. How to produce at the lowest cost is an economic decision, typically made by the firm's manager.

6.2 PRODUCTION

A firm uses a *technology* or *production process* to transform *inputs* or *factors of production* into *outputs*. Firms use many types of inputs. Most of these inputs can be grouped into three broad categories:

- **Capital (K):** Long-lived inputs such as land, buildings (factories, stores), and equipment (machines, trucks),
- **Labor (L):** Human services such as those provided by managers, skilled workers (architects, economists, engineers, plumbers), and less-skilled workers (custodians, construction laborers, assembly-line workers),
- **Materials (M):** Raw goods (oil, water, wheat) and processed products (aluminum, plastic, paper, steel).

The output can be a *service*, such as an automobile tune-up by a mechanic, or a *physical product*, such as a computer chip or a potato chip.

Production Functions

Firms can transform inputs into outputs in many different ways. Candy-manufacturing companies differ in the skills of their workforce and the amount of equipment they use. While all employ a chef, a manager, and relatively unskilled workers, some candy firms also use skilled technicians and modern equipment. In small candy companies, the relatively unskilled workers shape the candy, decorate it, package it, and box it by hand. In slightly larger firms, the relatively unskilled workers use conveyor belts and other equipment that was invented decades ago. In modern, large-scale plants, the relatively unskilled laborers work with robots and other state-of-the art machines, which are maintained by skilled technicians. Before deciding which production process to use, a firm needs to consider its various options.

The various ways inputs can be transformed into output are summarized in the **production function**: the relationship between the quantities of inputs used and the *maximum* quantity of output that can be produced, given current knowledge about technology and organization. The production function for a firm that uses only labor and capital is

$$q = f(L, K), \tag{6.1}$$

where q units of output (wrapped candy bars) are produced using L units of labor services (days of work by relatively unskilled assembly-line workers) and K units of capital (the number of conveyor belts).

The production function shows only the *maximum* amount of output that can be produced from given levels of labor and capital, because the production function includes only efficient production processes. A profit-maximizing firm is not interested in production processes that are inefficient and waste inputs: Firms do not want to use two workers to do a job that can be done as efficiently by one worker.

Time and the Variability of Inputs

A firm can more easily adjust its inputs in the long run than in the short run. Typically, a firm can vary the amount of materials and of relatively unskilled labor it uses comparatively quickly. However, it needs more time to find and hire skilled workers, order new equipment, or build a new manufacturing plant.

The more time a firm has to adjust its inputs, the more factors of production it can alter. The **short run** is a period of time so brief that at least one factor of production cannot be varied practically. A factor that cannot be varied practically in the short run is called a **fixed input**. In contrast, a **variable input** is a factor of production whose quantity can be changed readily by the firm during the relevant time period. The **long run** is a lengthy enough period of time that all inputs can be varied. There are no fixed inputs in the long run—all factors of production are variable inputs.

Suppose that a painting company gets more work than usual one day. Even if it wanted to do so, the firm does not have time to buy or rent an extra truck and buy another compressor to run a power sprayer; these inputs are fixed in the short run. To get the work done that afternoon, the firm uses the company's one truck to drop off a temporary worker, equipped with only a brush and a can of paint, at the last job. In the long run, however, the firm can adjust all its inputs. If the firm wants to paint more houses every day, it hires more full-time workers, gets a second truck, purchases more compressors to run the power sprayers, and buys a computer to keep track of all its projects.

How long it takes for all inputs to be variable depends on the factors a firm uses. For a janitorial service whose only major input is workers, the long run is a very brief period of time. In contrast, an automobile manufacturer may need many years to build a new manufacturing plant or to design and construct a new type of machine. A pistachio farmer needs the better part of a decade before newly planted trees yield a substantial crop of nuts.

For many firms over, say, a month, materials and often labor are variable inputs. However, labor is not always a variable input. Finding additional highly skilled workers may take substantial time. Similarly, capital may be a variable or fixed input. A firm can rent small capital assets (trucks and personal computers) quickly, but it may take the firm years to obtain larger capital assets (buildings and large, specialized pieces of equipment).

To illustrate the greater flexibility that a firm has in the long run than in the short run, we examine the production function in Equation 6.1, in which output is a function of only labor and capital. We look at first the short-run and then the long-run production process.

6.3 SHORT-RUN PRODUCTION: ONE VARIABLE AND ONE FIXED INPUT

In the short run, we assume that capital is a fixed input and labor is a variable input, so the firm can increase output only by increasing the amount of labor it uses. In the short run, the firm's production function is

$$q = f(L, \bar{K}), \tag{6.2}$$

where q is output, L is workers, and $\bar{K}$ is the fixed number of units of capital.

To illustrate the short-run production process, we consider a firm that assembles computers for a manufacturing firm that supplies it with the necessary parts, such as computer chips and disk drives. The assembly firm cannot increase its capital—eight workbenches fully equipped with tools, electronic probes, and other equipment for testing computers—in the short run, but it can hire extra workers or pay current workers extra to work overtime so as to increase production.

Total Product

The exact relationship between *output* or *total product* and *labor* can be illustrated by using a particular function, Equation 6.2, a table, or a figure. Table 6.1 shows the relationship between output and labor when capital is fixed for a firm. The first column lists the fixed amount of capital: eight fully equipped workbenches. As the number of workers, the amount of labor (second column), increases, total output, the number of computers assembled in a day (third column), first increases and then decreases.

With zero workers, no computers are assembled. One worker with access to the firm's equipment assembles five computers in a day. As the number of workers increases, so does output: 1 worker assembles 5 computers in a day, 2 workers assemble 18, 3 workers assemble 36, and so forth. The maximum number of computers that can be assembled with the capital on hand, however, is limited to 110 per day. That maximum can be produced with 10 or 11 workers. Adding extra workers beyond 11 lowers production as workers get in each other's way. The dashed line in the table indicates that a firm would not use more than 11 workers, as to do so would be inefficient. We can show how extra workers affect the total product by using two additional concepts: the marginal product of labor and the average product of labor.

Marginal Product of Labor

Before deciding whether to hire one more worker, a manager wants to determine how much this extra worker, $\Delta L = 1$, will increase output, Δq. That is, the manager wants to know the **marginal product of labor** (MP_L): the change in total output

Table 6.1 Total Product, Marginal Product, and Average Product of Labor with Fixed Capital

Capital, $\bar{K}$	Labor, L	Output, Total Product of Labor, Q	Marginal Product of Labor, $MP_L = \Delta Q/\Delta L$	Average Product of Labor, $AP_L = Q/L$
8	0	0		
8	1	5	5	5
8	2	18	13	9
8	3	36	18	12
8	4	56	20	14
8	5	75	19	15
8	6	90	15	15
8	7	98	8	14
8	8	104	6	13
8	9	108	4	12
8	10	110	2	11
8	11	110	0	10
8	12	108	–2	9
8	13	104	–4	8

resulting from using an extra unit of labor, holding other factors (capital) constant. If output changes by Δq when the number of workers increases by ΔL, the change in output per worker is[2]

$$MP_L = \frac{\Delta q}{\Delta L}.$$

As Table 6.1 shows, if the number of workers increases from 1 to 2, $\Delta L = 1$, output rises by $\Delta q = 13 = 18 - 5$, so the marginal product of labor is 13.

Average Product of Labor

Before hiring extra workers, a manager may also want to know whether output will rise in proportion to this extra labor. To answer this question, the firm determines how extra workers affect the **average product of labor** (AP_L): the ratio of output to the number of workers used to produce that output,

$$AP_L = \frac{q}{L}.$$

Table 6.1 shows that 10 workers can assemble 110 computers in a day, so the average product of labor for 10 workers is 11 computers. The average product of labor for 9 workers is 12 computers per day; thus increasing from 9 to 10 workers lowers the average product per worker.

Graphing the Product Curves

Figure 6.1 and Table 6.1 show how output, the average product of labor, and the marginal product of labor vary with the number of workers. (The figures are smooth curves because the firm can hire a "fraction of a worker" by employing a worker for a fraction of a day.) The curve in panel a of Figure 6.1 shows how a change in labor affects the **total product of labor**—the amount of output (or *total product*) that can be produced by a given amount of labor. Output rises with labor until it reaches its maximum of 110 computers at 11 workers, point *C*; with extra workers, the number of computers assembled falls.

Panel b of the figure shows how the average product of labor and marginal product of labor vary with the number of workers. We can line up the figures in panels a and b vertically because the units along the horizontal axes of both figures, the number of workers per day, are the same. The vertical axes differ, however. The vertical axis is total product in panel a and the average or marginal product of labor—a measure of output per unit of labor—in panel b.

Effect of Extra Labor. In most production processes, the average product of labor first rises and then falls as labor increases. One reason the AP_L curve initially rises in Figure 6.1 is that it helps to have more than two hands when assembling a computer. One worker holds a part in place while another one bolts it down. As a

[2]The calculus definition of the marginal product of labor is $MP_L = \partial q/\partial L = \partial f(L, \bar{K})/\partial L$, where capital is fixed at $\bar{K}$.

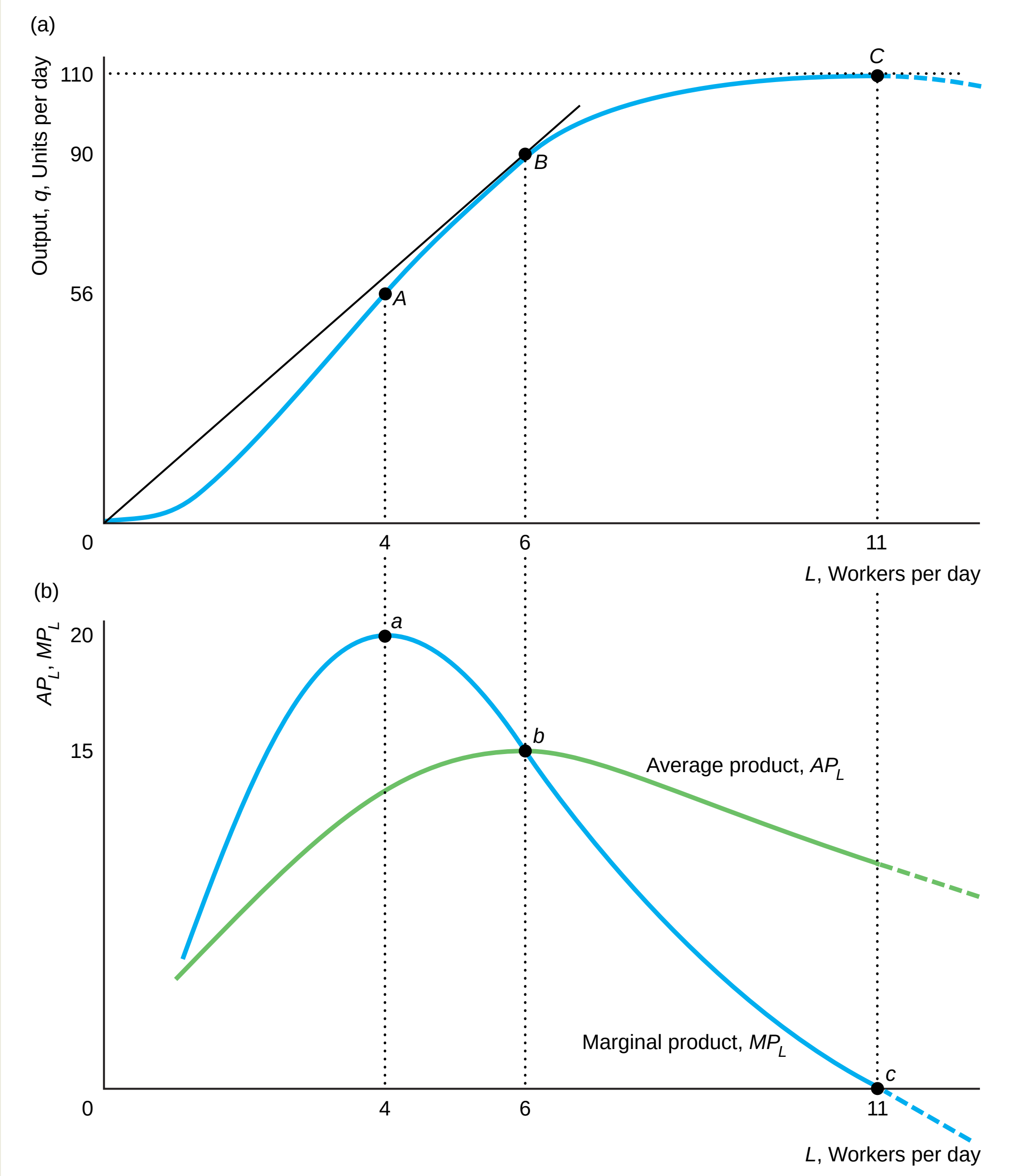


Figure 6.1 Production Relationships with Variable Labor. (a) The total product of labor curve shows how many computers, q, can be assembled with eight fully equipped workbenches and a varying number of workers, L, who work an eight-hour day (see columns 2 and 3 in Table 6.1). Where extra workers reduce the number of computers assembled, the total product curve is a dashed line, which indicates that such production is inefficient production and not part of the production function. The slope of the line from the origin to point B is the average product of labor for six workers. (b) The marginal product of labor ($MP_L = \Delta q/\Delta L$, column 4 of Table 6.1) equals the average product of labor ($AP_L = q/L$, column 5 of Table 6.1) at the peak of the average product curve.

result, output increases more than in proportion to labor, so the average product of labor rises. Doubling the number of workers from one to two more than doubles the output from 5 to 18 and causes the average product of labor to rise from 5 to 9, as Table 6.1 shows.

Similarly, output may initially rise more than in proportion to labor because of greater specialization of activities. With greater specialization, workers are assigned to tasks at which they are particularly adept, and time is saved by not having workers move from task to task.

As the number of workers rises further, however, output may not increase by as much per worker as they have to wait to use a particular piece of equipment or get in each other's way. In Figure 6.1, as the number of workers exceeds 6, total output increases less than in proportion to labor, so the average product falls.

If more than 11 workers are used, the total product curve falls with each extra worker as the crowding of workers gets worse. Because that much labor is not efficient, that section of the curve is drawn with a dashed line to indicate that it is not part of the production function. Similarly, the dashed portions of the average and marginal product curves are irrelevant because no firm would produce with that many workers.

Relationship of the Product Curves. The three curves are geometrically related. First we use panel b to illustrate the relationship between the average and marginal product of labor curves. Then we use panels a and b to show the relationship between the total product of labor curve and the other two curves.

The average product of labor curve slopes upward where the marginal product of labor curve is above it and slopes downward where the marginal product curve is below it. If an extra worker adds more output—that worker's marginal product—than the average product of the initial workers, the extra worker raises the average product. As Table 6.1 shows, the average product of 2 workers is 9. The marginal product for a third worker is 18—which is above the average product for two workers—so the average product rises from 9 to 12. As panel b shows, when there are fewer than 6 workers, the marginal product curve is above the average product curve, so the average product curve is upward sloping.

Similarly, if the marginal product of labor for a new worker is less than the former average product of labor, the average product of labor falls. In the figure, the average product of labor falls beyond 6 workers. Because the average product of labor curve rises when the marginal product of labor curve is above it and the average product of labor falls when the marginal product of labor is below it, the average product of labor curve reaches a peak, point *b* in panel b, where the marginal product of labor curve crosses it. (See Appendix 6A for a mathematical proof.)

The geometric relationship between the total product curve and the average and marginal product curves is illustrated in panels a and b of Figure 6.1. We can determine the average product of labor using the total product of labor curve. The average product of labor for L workers equals the slope of a straight line from the origin to a point on the total product of labor curve for L workers in panel a. The slope of this line equals output divided by the number of workers, which is the definition of the average product of labor. For example, the slope of the straight line drawn from the origin to point B ($L = 6$, $q = 90$) is 15, which equals the "rise" of $q = 90$ divided by the "run" of $L = 6$. As panel b shows, the average product of labor for 6 workers at point b is 15.

The marginal product of labor also has a geometric interpretation in terms of the total product curve. The slope of the total product curve at a given point, $\Delta q/\Delta L$, equals the marginal product of labor. That is, the marginal product of labor equals the slope of a straight line that is tangent to the total output curve at a given point. For example, at point C in panel a where there are 11 workers, the line tangent to the total product curve is flat, so the marginal product of labor is zero: A little extra labor has no effect on output. The total product curve is upward sloping when there are fewer than 11 workers, so the marginal product of labor is positive. If the firm is foolish enough to hire more than 11 workers, the total product curve slopes downward (dashed line), so the MP_L is negative: Extra workers lower output. Again, this portion of the MP_L curve is not part of the production function.

When there are 6 workers, the average product of labor equals the marginal product of labor. The reason is that the line from the origin to point B in panel a is tangent to the total product curve, so the slope of that line, 15, is the marginal product of labor and the average product of labor at point b in panel b.

Law of Diminishing Marginal Returns

Next to "supply equals demand," probably the most commonly used phrase of economic jargon is the "law of diminishing marginal returns." This law determines the shapes of the total product and marginal product of labor curves as the firm uses more and more labor.

The *law of diminishing marginal returns* (or *diminishing marginal product*) holds that, *if a firm keeps increasing an input, holding all other inputs and technology constant, the corresponding increases in output will become smaller eventually*. That is, if only one input is increased, *the marginal product of that input will diminish eventually*.

In Table 6.1, if the firm goes from 1 to 2 workers, the marginal product of labor is 13. If 1 or 2 more workers are used, the marginal product rises: The marginal product for 3 workers is 18, and the marginal product for 4 workers is 20. However, if the firm increases the number of workers beyond 4, the marginal product falls: The marginal product of 5 workers is 19, and that for 6 workers is 15. Beyond 4 workers, each extra worker adds less and less extra output, so the total product of labor curve rises by smaller increments. At 11 workers, the marginal product is zero. In short, the law of diminishing marginal returns says that if a firm keeps adding one more unit of an input, the extra output it gets grows smaller and smaller. This diminishing return to extra labor may be due to too many workers sharing too few machines or to crowding, as workers get in each other's way. Thus as the amount of labor used grows large enough, the marginal product curve approaches zero and the corresponding total product of labor curve becomes nearly flat.

Unfortunately, many people, when attempting to cite this empirical regularity, overstate it. Instead of talking about "diminishing *marginal* returns," they talk about "diminishing returns." The two phrases have different meanings. Where there are "diminishing marginal returns," the MP_L curve is falling—beyond 4 workers in panel b of Figure 6.1—but it may be positive, as the solid MP_L curve between 4 and 11 workers shows. With "diminishing returns," extra labor causes *output* to fall. There are diminishing (total) returns for more than 11 workers—a dashed MP_L line in panel b.

Thus saying that there are diminishing returns is much stronger than saying that there are diminishing marginal returns. We often observe firms producing where there are diminishing marginal returns to labor, but we rarely see firms operating where there are diminishing total returns. Only a firm that is willing to lose money would operate so inefficiently that it has diminishing returns. Such a firm could produce more output by using fewer inputs.

A second common misinterpretation of this law is to claim that marginal products must fall as we increase an input without requiring that technology and other inputs stay constant. If we increase labor while simultaneously increasing other factors or adopting superior technologies, the marginal product of labor may rise indefinitely. Thomas Malthus provided the most famous example of this fallacy.

Application

MALTHUS AND MASS STARVATION

In 1798, Thomas Malthus—a clergyman and professor of modern history and political economy—predicted that (unchecked) population would grow more rapidly than food production because the quantity of land was fixed. The problem, he believed, was that the fixed amount of land would lead to diminishing marginal product of labor, so output would rise less than in proportion to the increase in farmworkers. Malthus grimly concluded that mass starvation would result. Brander and Taylor (1998) argue that such a disaster may have occurred on Easter Island around 500 years ago.

Since Malthus's day, world population has increased nearly 800%. Why haven't we starved to death? The simple explanation is that fewer workers using less land can produce much more food today than was possible when Malthus was alive. Two hundred years ago, most of the population had to work in agriculture to prevent starvation. Today, less than 2% of the U.S. population works in agriculture, and the share of land devoted to farming is constantly falling. Yet U.S. food production continues to grow faster than the U.S. population. Since World War II, world population doubled but food production tripled.

Two key factors (in addition to birth control) are responsible for the rapid increase in food production per capita in most countries. First, agricultural technology—such as disease-resistant seeds and better land management practices—has improved substantially, so more output can be produced with the same inputs. Second, although the amounts of land and labor used have remained constant or fallen in most countries in recent years, the use of other inputs such as fertilizer and tractors has increased significantly, so output per acre of land has risen.

In the last three decades of the twentieth century, U.S. farm productivity (measured in value added per hour worked) rose at an average of 4.5% a year, about triple the rate of improvement of nonfarm business productivity. For example, although the nation's dairy herd had shrunk to about three-fourths its size in the late 1960s, milk production increased by more than a third. By 2002, one in seven U.S. farms was experimenting with robotic milking systems.[3]

In 1850, it took more than 80 hours of labor to produce 100 bushels of corn. Introducing mechanical power cut the labor required in half. Labor needs were again cut in half by the introduction of hybrid seed and chemical fertilizers, and then by the advent of herbicides and pesticides. Biotechnology, with the recent introduction of herbicide-tolerant and insect-resistant crops in 1996, has reduced the labor requirement today to about two hours of labor.

As countries adopt new products and methods, they benefit from the rapid growth of food production per capita. For example, from 1991 to 2001, per capita food production increased by 20% in developing countries.

However, parts of Africa have severe problems. Per capita food production has fallen in Africa over the past two decades. Worse, in several recent years, mass starvation has plagued some African countries. Although droughts have contributed, these tragedies appear to be primarily due to political problems such as wars and a breakdown of economic production and distribution systems. If these political problems cannot be solved, Malthus may prove to be right for the wrong reason.

6.4 LONG-RUN PRODUCTION: TWO VARIABLE INPUTS

Eternity is a terrible thought. I mean, where's it going to end?

—Tom Stoppard

We started our analysis of production functions by looking at a short-run production function in which one input, capital, was fixed, and the other, labor, was variable. In the long run, however, both of these inputs are variable. With both factors variable, a firm can usually produce a given level of output by using a great deal of labor and very little capital, a great deal of capital and very little labor, or moderate amounts of both. That is, the firm can substitute one input for another while continuing to produce the same level of output, in much the same way that a consumer can maintain a given level of utility by substituting one good for another.

Typically, a firm can produce in a number of different ways, some of which require more labor than others. For example, a lumberyard can produce 200 planks an hour with 10 workers using hand saws, with 4 workers using handheld power saws, or with 2 workers using bench power saws.

---

[3] See **www.aw.com/perloff**, Chapter 6, "Does that Compute Down on the Farm?" on the technical progress of farms due to their increased use of computers.

Table 6.2 Output Produced with Two Variable Inputs

	Labor, L					
Capital, K	1	2	3	4	5	6
1	10	14	17	20	22	**24**
2	14	20	**24**	28	32	35
3	17	**24**	30	35	39	42
4	20	28	35	40	45	49
5	22	32	39	45	50	55
6	**24**	35	42	49	55	60

We illustrate a firm's ability to substitute between inputs in Table 6.2, which shows the amount of output per day the firm produces with various combinations of labor per day and capital per day. The labor inputs are along the top of the table, and the capital inputs are in the first column. The table shows four combinations of labor and capital that the firm may use to produce 24 units of output. The firm may employ (a) 1 worker and 6 units of capital, (b) 2 workers and 3 units of capital, (c) 3 workers and 2 units of capital, or (d) 6 workers and 1 unit of capital.

Isoquants

These four combinations of labor and capital are labeled *a*, *b*, *c*, and *d* on the "q = 24" curve in Figure 6.2. We call such a curve an **isoquant**, which is a curve that shows the efficient combinations of labor and capital that can produce a single (*iso*) level of output (*quant*ity). If the production function is $q = f(L, K)$, then the equation for an isoquant where output is held constant at $\bar{q}$ is

$$\bar{q} = f(L, K).$$

An isoquant shows the flexibility that a firm has in producing a given level of output. Figure 6.2 shows three isoquants corresponding to three levels of output. These isoquants are smooth curves because the firm can use fractional units of each input.

We can use these isoquants to illustrate what happens in the short run when capital is fixed and only labor varies. As Table 6.2 shows, if capital is constant at 2 units, 1 worker produces 14 units of output (point *e* in Figure 6.2), 3 workers produce 24 units (point *c*), and 6 workers produce 35 units (point *f*). Thus if the firm holds one factor constant and varies another factor, it moves from one isoquant to another. In contrast, if the firm increases one input while lowering the other appropriately, the firm stays on a single isoquant.

Properties of Isoquants. Isoquants have most of the same properties as indifference curves. The biggest difference between indifference curves and isoquants is that an isoquant holds quantity constant, whereas an indifference curve holds utility constant. We now discuss three major properties of isoquants. Most of these properties result from firms' producing efficiently.

First, *the farther an isoquant is from the origin, the greater the level of output.* That is, the more inputs a firm uses, the more output it gets if it produces efficiently.

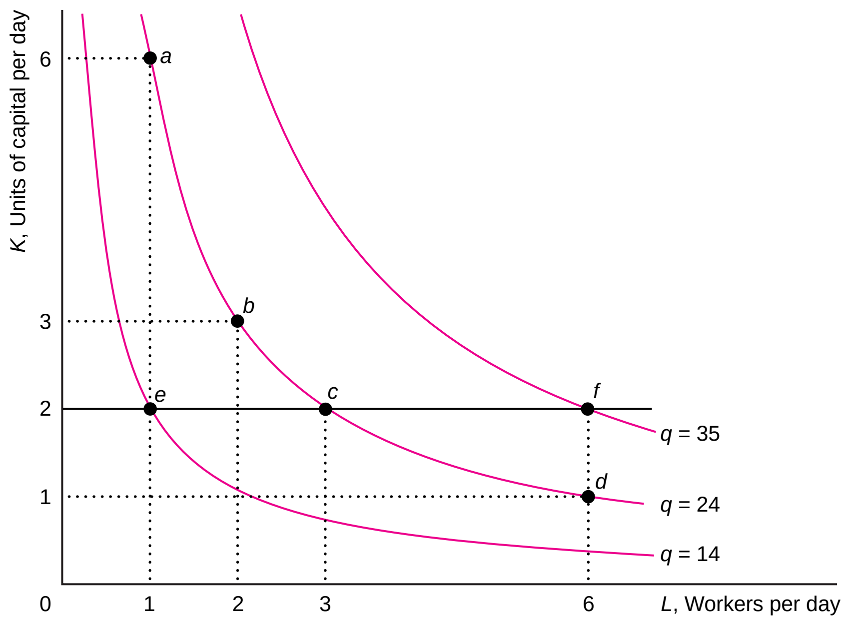


Figure 6.2 Family of Isoquants. These isoquants show the combinations of labor and capital that produce various levels of output. Isoquants farther from the origin correspond to higher levels of output. Points *a*, *b*, *c*, and *d* are various combinations of labor and capital the firm can use to produce $q = 24$ units of output. If the firm holds capital constant at 2 and increases labor from 1 (point *e*) to 3 (*c*) to 6 (*f*), it shifts from the $q = 14$ isoquant to the $q = 24$ isoquant and then to the $q = 35$ isoquant.

At point *e* in Figure 6.2, the firm is producing 14 units of output with 1 worker and 2 units of capital. If the firm holds capital constant and adds 2 more workers, it produces at point *c*. Point *c* must be on an isoquant with a higher level of output—here, 24 units—if the firm is producing efficiently and not wasting the extra labor.

Second, *isoquants do not cross*. Such intersections are inconsistent with the requirement that the firm always produces efficiently. For example, if the $q = 15$ and $q = 20$ isoquants crossed, the firm could produce at either output level with the same combination of labor and capital. The firm must be producing inefficiently if it produces $q = 15$ when it could produce $q = 20$. So that labor-capital combination should not lie on the $q = 15$ isoquant, which should include only efficient combinations of inputs. Thus efficiency requires that isoquants do not cross.

Third, *isoquants slope downward*. If an isoquant sloped upward, the firm could produce the same level of output with relatively few inputs or relatively many inputs. Producing with relatively many inputs would be inefficient. Consequently, because isoquants show only efficient production, an upward-sloping isoquant is impossible. Virtually the same argument can be used to show that isoquants must be thin, as you are asked to do in Question 4 at the end of this chapter.

Shape of Isoquants. The curvature of an isoquant shows how readily a firm can substitute one input for another. The two extreme cases are production processes in which inputs are perfect substitutes or in which they cannot be substituted for each other.

If the inputs are perfect substitutes, each isoquant is a straight line. Suppose either potatoes from Maine, x, or potatoes from Idaho, y, both of which are measured in pounds per day, can be used to produce potato salad, q, measured in pounds. The production function is

$$q = x + y.$$

One pound of potato salad can be produced by using 1 pound of Idaho potatoes and no Maine potatoes, 1 pound of Maine potatoes and no Idahoes, or 1/2 pound of each type of potato. Panel a of Figure 6.3 shows the q = 1, 2, and 3 isoquants.

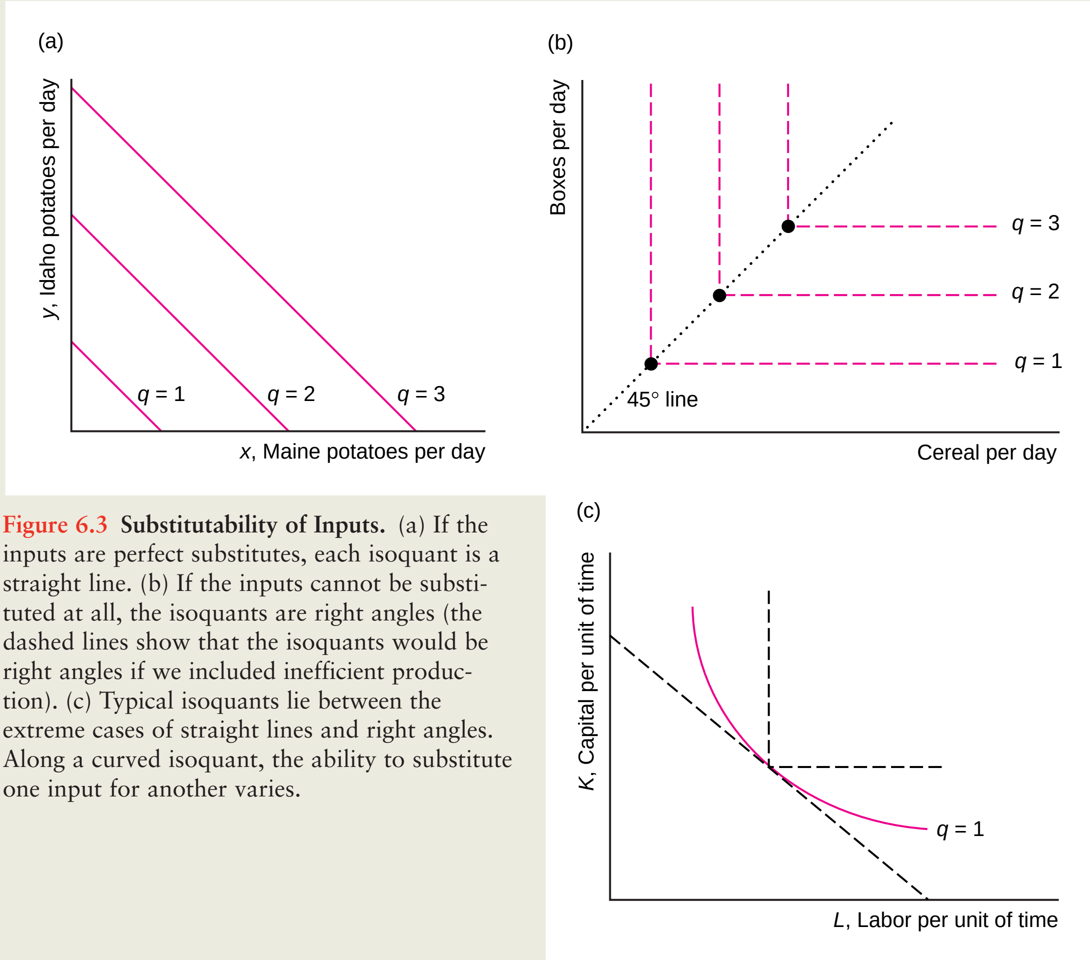


Figure 6.3 Substitutability of Inputs. (a) If the inputs are perfect substitutes, each isoquant is a straight line. (b) If the inputs cannot be substituted at all, the isoquants are right angles (the dashed lines show that the isoquants would be right angles if we included inefficient production). (c) Typical isoquants lie between the extreme cases of straight lines and right angles. Along a curved isoquant, the ability to substitute one input for another varies.

These isoquants are straight lines with a slope of –1 because we need to use an extra pound of Maine potatoes for every pound fewer of Idaho potatoes used.[4]

Sometimes it is impossible to substitute one input for the other: Inputs must be used in fixed proportions. For example, the inputs to produce a 12-ounce box of cereal, q, are cereal (in 12-ounce units per day) and cardboard boxes (boxes per day). If the firm has one unit of cereal and one box, it can produce one box of cereal. If it has one unit of cereal and two boxes, it can still make only one box of cereal. Thus in panel b, the only efficient points of production are the large dots along the 45° line.[5] Dashed lines show that the isoquants would be right angles if isoquants could include inefficient production processes.

Other production processes allow imperfect substitution between inputs. The isoquants are convex (so the middle of the isoquant is closer to the origin than it would be if the isoquant were a straight line). They do not have the same slope at every point, unlike the straight-line isoquants. Most isoquants are smooth, slope downward, curve away from the origin, and lie between the extreme cases of straight lines (perfect substitutes) and right angles (nonsubstitutes), as panel c illustrates.

A SEMICONDUCTOR INTEGRATED CIRCUIT ISOQUANT

We can show why isoquants curve away from the origin by deriving an isoquant for semiconductor integrated circuits (ICs, or "chips"). ICs—the "brains" of computers and other electronic devices—are made by building up layers of conductive and insulating materials on silicon wafers. Each wafer contains many ICs, which are subsequently cut into individual chips, called *dice*.

Semiconductor manufacturers ("fabs") buy the silicon wafers and then use labor and capital to produce the chips. A semiconductor IC's several layers of conductive and insulating materials are arranged in patterns that define the function of the chip.

During the manufacture of ICs, a track moves a wafer into a machine where it is spun and a light-sensitive liquid called photoresist is applied to its whole surface; the photoresist is then hardened. The wafer advances along the track to a point where photolithography is used to define patterns in the photoresist. In photolithography, light transfers a pattern from a template, called a photomask, to the photoresist, which is then "developed" like film, creating a pattern by removing the resist from certain areas. A subsequent process then can either add to or etch away those areas not protected by the resist.

In a repetition of this entire procedure, additional layers are created on the wafer. Because the conducting and insulating patterns in each layer interact with those in the previous layers, the patterns must line up correctly.

To align layers properly, firms use combinations of labor and equipment. In the least capital-intensive technology, employees use machines called *aligners*.

---

[4]The isoquant for $\bar{q} = 1$ pound of potato salad is $1 = x + y$, or $y = 1 - x$. This equation shows that the isoquant is a straight line with a slope of –1.

[5]This fixed-proportions production function is $q = \min(g, b)$, where g is the number of 12-ounce measures of cereal, b is the number of boxes used in a day, and the min function means "the minimum number of g or b." For example, if g is 4 and b is 3, q is 3.

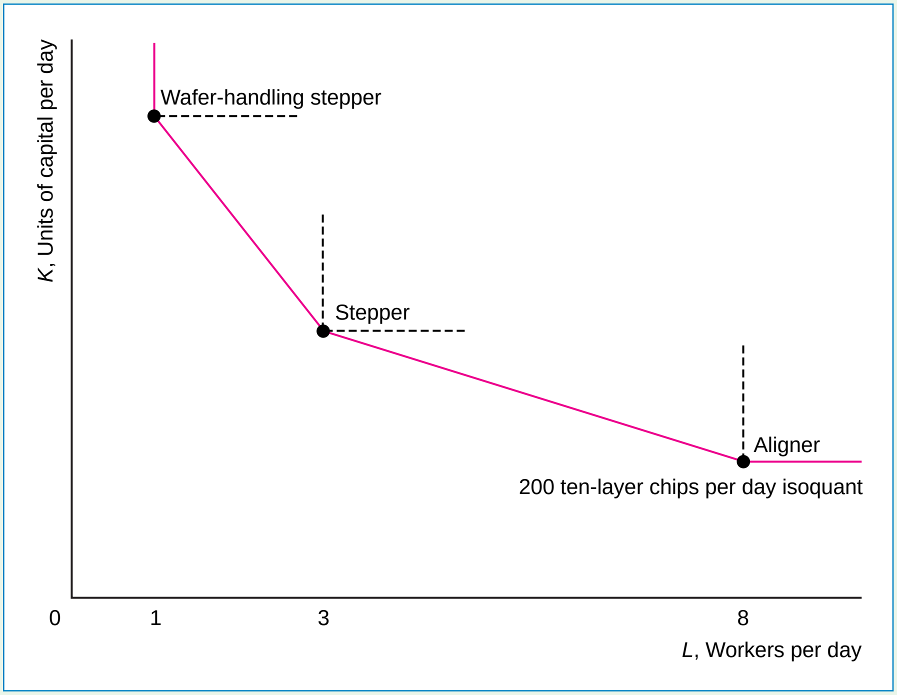
Operators look through microscopes and line up the layers by hand and then expose the entire surface. An operator running an aligner can produce 250 layers a day, or 25 ten-layer chips.

A second, more capital-intensive technology uses machines called *steppers*. The stepper picks a spot on the wafer, automatically aligns the layers, and then exposes that area to light. Then the machine moves—*steps* to other sections—lining up and exposing each area in turn until the entire surface has been aligned and exposed. This technology requires less labor: A single worker can run two steppers and produce 500 layers, or 50 ten-layer chips, per day.

A third, even more capital-intensive technology uses a stepper with wafer-handling equipment, which reduces the amount of labor even more. By linking the tracks directly to a stepper and automating the chip transfer process, human handling can be greatly reduced. A single worker can run 4 steppers with wafer-handling equipment and produce 1,000 layers, or 100 ten-layer chips, per day.

Only steppers can be used if the chip requires line widths of 1 micrometer or less. (Pathways on personal computer chips have narrowed from 1.5 micrometers on Intel Corporation's 386 chip, to 0.13 micrometers on its Pentium 4.) We show an isoquant for producing 200 ten-layer chips that have lines that are more than 1 micrometer wide, for which any of the three technologies can be used.

All three technologies use labor and capital in fixed proportions. To produce 200 chips takes 8 workers and 8 aligners, 3 workers and 6 steppers, or 1 worker and 4 steppers with wafer-handling capabilities. The accompanying graph shows the three right-angle isoquants corresponding to each of these three technologies.



Some fabs, however, employ a combination of these technologies; some workers use one type of machine while others use different types. By doing so, the fabs can produce using intermediate combinations of labor and capital, as the solid-line, kinked isoquant illustrates. The firm does *not* use a combination of the aligner and the wafer-handling stepper technologies because those combinations are less efficient than using the plain stepper (the line connecting the aligner and wafer-handling stepper technologies is farther from the origin than the lines between those technologies and the plain stepper technology).

New processes are constantly being invented. As they are introduced, the isoquant will have more and more kinks (one for each new process) and will begin to resemble the smooth, usual-shaped isoquants we've been drawing.

Substituting Inputs

The slope of an isoquant shows the ability of a firm to replace one input with another while holding output constant. Figure 6.4 illustrates this substitution using an isoquant for a Norwegian printing firm, which uses labor, L, and capital, K, to print its output, Q.[6] The isoquant shows various combinations of L and K that the firm can use to produce 10 units of output.

The firm can produce 10 units of output using the combination of inputs at a or b. At point a, the firm uses 2 workers and 39 units of capital. The firm could produce the same amount of output using $\Delta K = -18$ fewer units of capital if it used one more worker, $\Delta L = 1$, point b. If we drew a straight line from a to b, its slope would be $\Delta K/\Delta L = -18$. Thus this slope tells us how many fewer units of capital (18) the firm can use if it hires one more worker.[7]

The slope of an isoquant is called the *marginal rate of technical substitution* (*MRTS*):

$$MRTS = -\frac{\text{change in capital}}{\text{change in labor}} = \frac{\Delta K}{\Delta L}.$$

The **marginal rate of technical substitution** tells us how many units of capital the firm can replace with an extra unit of labor while holding output constant. Because isoquants slope downward, the *MRTS* is negative.

---

[6]This isoquant for $\bar{q} = 10$ is based on the estimated production function $q = 1.52L^{0.6}K^{0.4}$ (Griliches and Ringstad, 1971), where different units of measure are used. A unit of labor, L, is a worker-year (2,000 hours of work). Because capital, K, includes various types of machines, and output, q, reflects different types of printed matter, their units cannot be described by any common terms. This production function is an example of a Cobb-Douglas (Appendix 4A, Chapter 5) production function, whose properties are examined in Appendix 6B.

[7]The slope of the isoquant at a point equals the slope of a straight line that is tangent to the isoquant at that point. Thus the straight line between two nearby points on an isoquant has nearly the same slope as that of the isoquant.

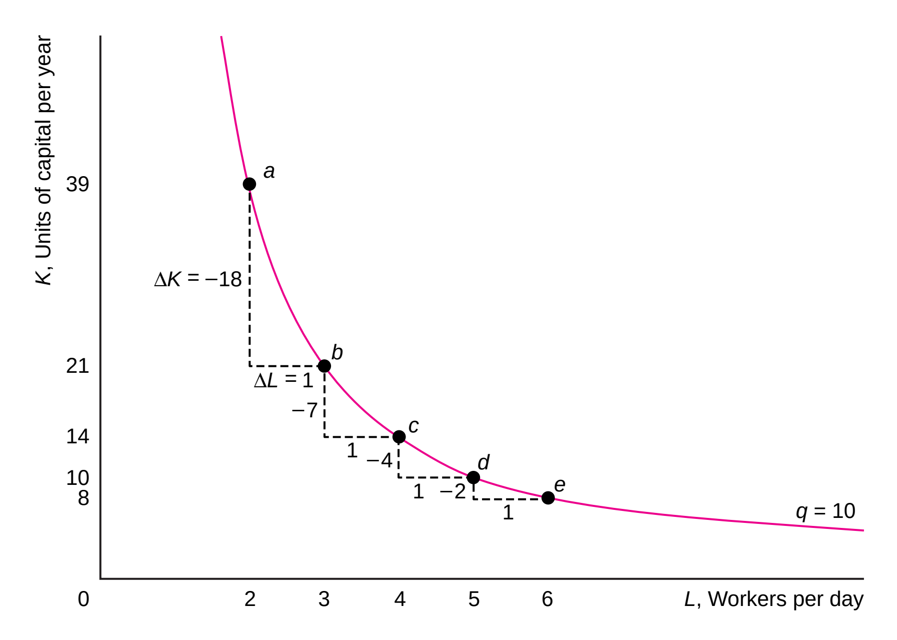


Figure 6.4 How the Marginal Rate of Technical Substitution Varies Along an Isoquant. Moving from point *a* to *b*, a Norwegian printing firm (Griliches and Ringstad, 1971) can produce the same amount of output, $q = 10$, using 18 fewer units of capital, $\Delta K = -18$, if it uses 1 more worker, $\Delta L = 1$. Thus its $MRTS = \Delta K/\Delta L = -18$. Moving from point *b* to *c*, its *MRTS* is −7. If it adds yet another worker, moving from point *c* to *d*, its *MRTS* is −4. If it adds one more worker, moving from *d* to *e*, its *MRTS* is −2. Thus because it curves away from the origin, this isoquant exhibits a diminishing marginal rate of technical substitution. That is, each extra worker allows the firm to reduce capital by a smaller amount as the ratio of capital to labor falls.

Substitutability of Inputs Varies Along an Isoquant. The marginal rate of technical substitution varies along a curved isoquant, as in Figure 6.4 for the printing firm. If the firm is initially at point *a* and it hires one more worker, the firm gives up 18 units of capital and yet remains on the same isoquant at point *b*, so the *MRTS* is −18. If the firm hires another worker, the firm can reduce its capital by 7 units and yet stay on the same isoquant, moving from point *b* to *c*, so the *MRTS* is −7. If the firm moves from point *c* to *d*, the *MRTS* is −4; and if it moves from point *d* to *e*, the *MRTS* is −2. This decline in the *MRTS* (in absolute value) along an isoquant as the firm increases labor illustrates *diminishing marginal rates of technical substitution*.

The curvature of the isoquant away from the origin reflects diminishing marginal rates of technical substitution. The more labor the firm has, the harder it is to replace the remaining capital with labor, so the *MRTS* falls as the isoquant becomes flatter.

In the special case in which isoquants are straight lines, isoquants do not exhibit diminishing marginal rates of technical substitution because neither input becomes more valuable in the production process: The inputs remain perfect substitutes. Solved Problem 6.1 illustrates this result.

Solved Problem 6.1

Does the marginal rate of technical substitution vary along the isoquant for the firm that produced potato salad using Idaho and Maine potatoes? What is the *MRTS* at each point along the isoquant?

Answer

1. *Determine the shape of the isoquant.* As panel a of Figure 6.3 illustrates, the potato salad isoquants are straight lines because the two types of potatoes are perfect substitutes.
2. *On the basis of the shape, conclude whether the MRTS is constant along the isoquant.* Because the isoquant is a straight line, the slope is the same at every point, so the *MRTS* is constant.
3. *Determine the MRTS at each point.* Earlier, we showed that the slope of this isoquant was −1, so the *MRTS* is −1 at each point along the isoquant. That is, because the two inputs are perfect substitutes, 1 pound of Idaho potatoes can be replaced by 1 pound of Maine potatoes.

Substitutability of Inputs and Marginal Products. The marginal rate of technical substitution—the degree to which inputs can be substituted for each other—equals the ratio of the marginal products of labor to the marginal product of capital, as we now show. The marginal rate of technical substitution tells us how much a firm can increase one input and lower the other while still staying on the same isoquant. Knowing the marginal products of labor and capital, we can determine how much one input must increase to offset a reduction in the other.

Because the marginal product of labor, $MP_L = \Delta q/\Delta L$, is the increase in output per extra unit of labor, if the firm hires ΔL more workers, its output increases by $MP_L \times \Delta L$. For example, if the MP_L is 2 and the firm hires one extra worker, its output rises by 2 units.

A decrease in capital alone causes output to fall by $MP_K \times \Delta K$, where $MP_K = \Delta q/\Delta K$ is the marginal product of capital—the output the firm loses from decreasing capital by one unit, holding all other factors fixed. To keep output constant, $\Delta q = 0$, this fall in output from reducing capital must exactly equal the increase in output from increasing labor:

$$(MP_L \times \Delta L) + (MP_K \times \Delta K) = 0.$$

Rearranging these terms, we find that[8]

$$-\frac{MP_L}{MP_K} = \frac{\Delta K}{\Delta L} = MRTS. \tag{6.3}$$

That is, the marginal rate of technical substitution, which is the absolute value of the change in capital relative to the change in labor, equals the ratio of the marginal products.

We can use Equation 6.3 to explain why marginal rates of technical substitution diminish as we move to the right along the isoquant in Figure 6.4. As we replace capital with labor (shift downward and to the right along the isoquant), the marginal product of capital increases—when there are few pieces of equipment per worker, each remaining piece is more useful—and the marginal product of labor falls, so the $MRTS = -MP_L/MP_K$ falls.[9]

6.5 RETURNS TO SCALE

So far, we have examined the effects of increasing one input while holding the other input constant (the shift from one isoquant to another) or decreasing the other input by an offsetting amount (the movement along an isoquant). We now turn to the question of *how much output changes if a firm increases all its inputs proportionately.* The answer helps a firm determine its *scale* or size in the long run.

In the long run, a firm can increase its output by building a second plant and staffing it with the same number of workers as in the first one. Whether the firm chooses to do so turns in part on whether its output increases less than in proportion, in proportion, or more than in proportion to its inputs.

Constant, Increasing, and Decreasing Returns to Scale

If, when all inputs are increased by a certain percentage, output increases by that same percentage, the production function is said to exhibit **constant returns to scale** (*CRS*). A firm's production process has constant returns to scale if, when the firm

---

[8]We can derive this result directly by totally differentiating an isoquant, $\bar{q} = f(L, K)$. As we change labor and capital, the output doesn't change, so

$$\mathrm{d}\bar{q} = 0 = \frac{\partial f}{\partial L}\mathrm{d}L + \frac{\partial f}{\partial K}\mathrm{d}K \equiv MP_L\,\mathrm{d}L + MP_K\,\mathrm{d}K.$$

Rearranging this expression, we find that $-MP_L/MP_K = \mathrm{d}K/\mathrm{d}L = MRTS$.

[9]Figure 6.4 shows the effects of fairly large changes in labor and capital along a printing firm's isoquant. Calculated exactly for small changes (see Appendix 6B), the printing firm's $MRTS = 1.5K/L$. As we move to the right along this isoquant, the amount of capital decreases and the amount of labor increases, so the capital-labor ratio falls, causing the $MRTS$ to fall.

doubles its inputs—builds an identical second plant and uses the same amount of labor and equipment as in the first plant—it doubles its output: $f(2L, 2K) = 2f(L, K)$.

We can check whether the potato salad production function has constant returns to scale. If a firm uses x_1 pounds of Idaho potatoes and y_1 pounds of Maine potatoes, it produces $q_1 = x_1 + y_1$ pounds of potato salad. If it doubles both inputs, using $x_2 = 2x_1$ Idaho and $y_2 = 2y_1$ Maine potatoes, it doubles its output:

$$q_2 = x_2 + y_2 = 2x_1 + 2y_1 = 2q_1.$$

Thus the potato salad production function exhibits constant returns to scale.

If output rises more than in proportion to an equal percentage increase in all inputs, the production function is said to exhibit **increasing returns to scale** (*IRS*). A technology exhibits increasing returns to scale if doubling inputs more than doubles the output: $f(2L, 2K) > 2f(L, K)$.

Why might a production function have increasing returns to scale? One reason is that, although it could duplicate a small factory and double its output, the firm might be able to more than double its output by building a single large plant, allowing for greater specialization of labor or capital. In the two smaller plants, workers have to perform many unrelated tasks such as operating, maintaining, and fixing the machines they use. In the large plant, some workers may specialize in maintaining and fixing machines, thereby increasing efficiency. Similarly, a firm may use specialized equipment in a large plant but not in a small one.

If output rises less than in proportion to an equal percentage increase in all inputs, the production function exhibits **decreasing returns to scale** (*DRS*). A technology exhibits decreasing returns to scale if doubling inputs causes output to rise less than in proportion: $f(2L, 2K) < 2f(L, K)$.

One reason for decreasing returns to scale is that the difficulty of organizing, coordinating, and integrating activities increases with firm size. An owner may be able to manage one plant well but may have trouble running two plants.[10] Another reason is that large teams of workers may not function as well as small teams, in which each individual takes greater personal responsibility.

One of the most widely estimated production functions is the Cobb-Douglas (Appendix 6B):

$$q = AL^{\alpha}K^{\beta}, \tag{6.4}$$

where A, α, and β are all positive constants. Solved Problem 6.2 shows that $\gamma = \alpha + \beta$ determines the returns to scale in a Cobb-Douglas production function.

[10]In some sense, the owner's difficulties in running a larger firm may reflect our failure to take into account some factor such as management in our production function. When the firm increases the various inputs, it does not increase the management input in proportion. If so, the "decreasing returns to scale" is really due to a fixed input.

Solved Problem 6.2

Under what conditions does a Cobb-Douglas production function, Equation 6.4, exhibit decreasing, constant, or increasing returns to scale?

Answer

1. *Show how output changes if both inputs are doubled*: If the firm initially uses L and K amounts of inputs, it produces

$$q_1 = AL^{\alpha}K^{\beta}.$$

When the firm doubles the amount of both labor and capital it uses, it produces

$$q_2 = A(2L)^{\alpha}(2K)^{\beta} = 2^{\alpha+\beta}AL^{\alpha}K^{\beta}.$$

Thus its output increases by

$$\frac{q_2}{q_1} = \frac{2^{\alpha+\beta}AL^{\alpha}K^{\beta}}{AL^{\alpha}K^{\beta}} = 2^{\alpha+\beta} \equiv 2^{\gamma}, \qquad (6.5)$$

where $\gamma \equiv \alpha + \beta$.

2. *Give a rule for determining the returns to scale*: The Cobb-Douglas production function has decreasing, constant, or increasing returns to scale as γ is less than, equal to, or greater than 1. For example, if $\gamma = 1$, doubling inputs doubles output, $q_2/q_1 = 2^{\gamma} = 2^1 = 2$, so the production function exhibits constant returns to scale.

RETURNS TO SCALE IN MANUFACTURING

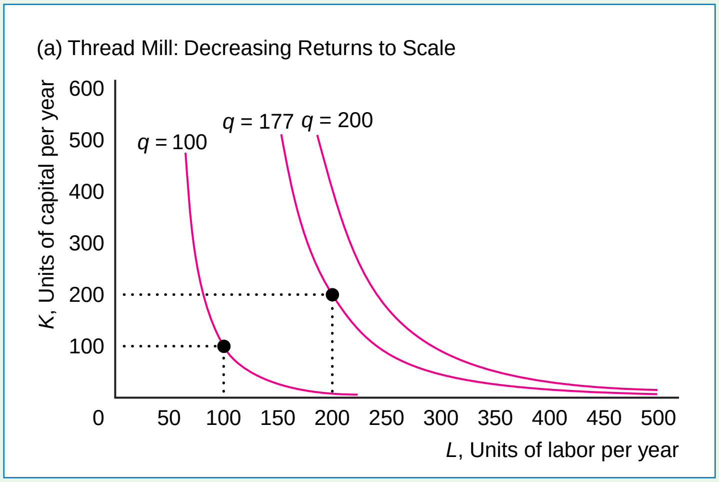
Increasing, constant, and decreasing returns to scale are commonly observed. The table shows estimates of Cobb-Douglas production functions and rates of returns in various Canadian manufacturing industries (Baldwin and Gorecki, 1986). The returns to scale measure in the table, γ, is an elasticity. It represents the percentage change in output for a 1% increase in all the inputs. Because the estimated returns to scale measures for a shoe firm is 1, a 1% increase in the inputs causes a 1% increase in output. Thus a shoe firm's production function exhibits constant returns to scale.

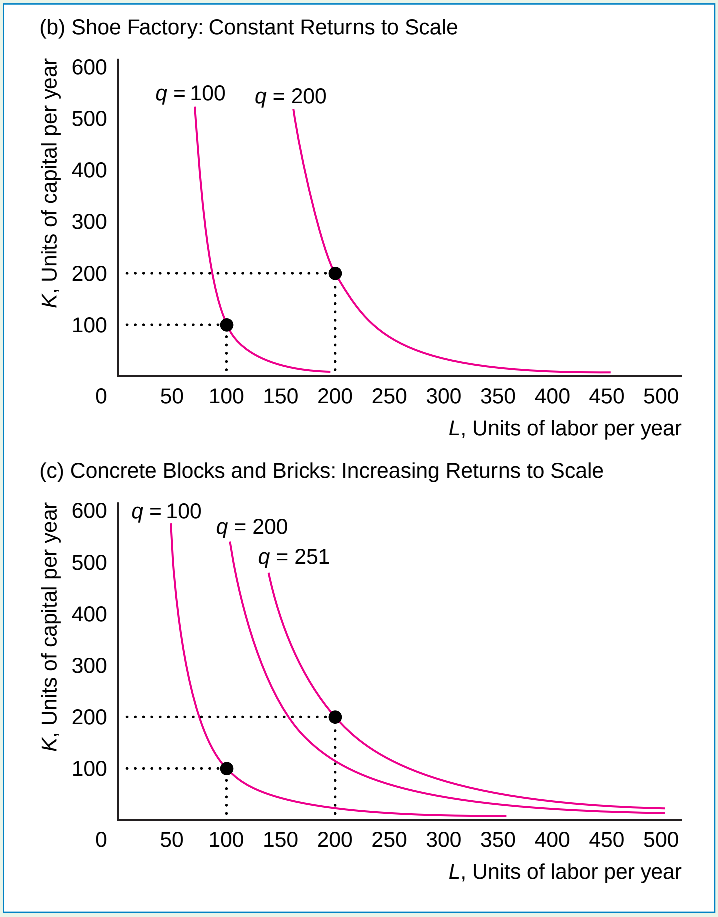
The estimated returns to scale measure for a thread mill is 0.82: A 1% increase in the inputs causes output to rise by 0.82%. Because output rises less than in proportion to the inputs, the thread mill production function exhibits decreasing returns to scale. In contrast, firms that make concrete blocks and

	Labor, α	Capital, β	Scale, $\gamma = \alpha + \beta$
Decreasing Returns to Scale			
Thread mill	0.64	0.18	0.82
Knitted fabrics	0.55	0.36	0.90
Lime manufacturers	0.60	0.25	0.84
Constant Returns to Scale			
Shoe factories	0.82	0.18	1.00
Hosiery mills	0.55	0.46	1.01
Jewelry and silverware	0.60	0.41	1.01
Increasing Returns to Scale			
Concrete blocks and bricks	0.93	0.40	1.33
Paint	0.71	0.61	1.32
Orthopedic and surgical appliances	0.30	0.99	1.30

bricks have increasing returns to scale production functions, in which a 1% increase in all inputs causes output to rise by 1.33%.

The accompanying graphs use isoquants to illustrate the returns to scale for the thread mill, shoe factory, and concrete block and bricks firm. We measure the units of labor, capital, and output so that, for all three firms, 100 units of labor and 100 units of capital produce 100 units of output on the $q = 100$ isoquant in the three panels. For the constant returns to scale shoe factory, panel b, if both labor and capital are doubled from 100 to 200 units, output doubles to 200 (= 100×2^1, multiplying the original output by the rate of increase using Equation 6.5).





That same doubling of inputs causes output to rise to only 177 ($\approx 100 \times 2^{0.82}$) for the thread mill, panel a. Because output rises less than in proportion to inputs, the production function has decreasing returns to scale. If the concrete block and brick firm doubles its inputs, panel c, its output more than doubles, to 251 ($\approx 100 \times 2^{1.33}$), so the production function has increasing returns to scale.

These graphs illustrate that the spacing of the isoquant determines the returns to scale. The closer together the q = 100 and q = 200 isoquants, the greater the returns to scale.

The returns to scale in these industries are estimated to be the same at all levels of output. A production function's returns to scale may vary, however, as the scale of the firm changes.

Varying Returns to Scale

Many production functions have increasing returns to scale for small amounts of output, constant returns for moderate amounts of output, and decreasing returns for large amounts of output. When a firm is small, increasing labor and capital allows for gains from cooperation between workers and greater specialization of workers and equipment—*returns to specialization*—so there are increasing returns to scale. As the firm grows, returns to scale are eventually exhausted. There are no more returns to specialization, so the production process has constant returns to scale. If the firm continues to grow, the owner starts having difficulty managing everyone, so the firm suffers from decreasing returns to scale.

We show such a pattern in Figure 6.5. Again, the spacing of the isoquants reflects the returns to scale. Initially, the firm has one worker and one piece of equipment, point a, and produces 1 unit of output on the $q = 1$ isoquant. If the firm doubles its inputs, it produces at b, where $L = 2$ and $K = 2$, which lies on the dashed line through the origin and point a. Output more than doubles to $q = 3$, so the

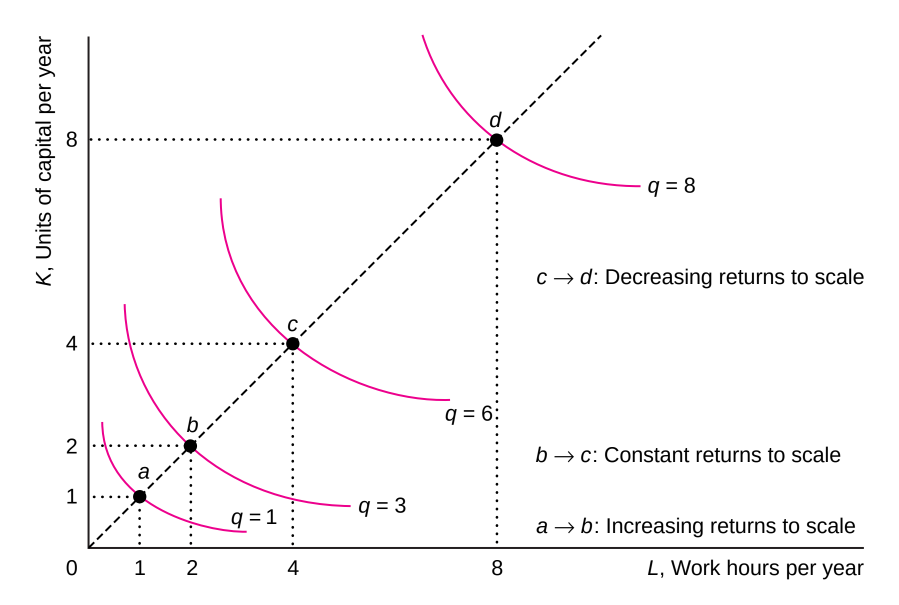


Figure 6.5 Varying Scale Economies. This production function exhibits varying returns to scale. Initially, the firm uses one worker and one unit of capital, point a. It repeatedly doubles these inputs to points b, c, and d, which lie along the dashed line. The first time the inputs are doubled, a to b, output more than doubles from $q = 1$ to $q = 3$, so the production function has increasing returns to scale. The next doubling, b to c, causes a proportionate increase in output, constant returns to scale. At the last doubling, from c to d, the production function exhibits decreasing returns to scale.

production function exhibits increasing returns to scale in this range. Another doubling of inputs to c causes output to double to 6 units, so the production function has constant returns to scale in this range. Another doubling of inputs to d causes output to increase by only a third, to $q = 8$, so the production function has decreasing returns to scale in this range.

6.6 PRODUCTIVITY AND TECHNICAL CHANGE

Because firms may use different technologies and different methods of organizing production, the amount of output that one firm produces from a given amount of inputs may differ from that produced by another firm. Moreover, after a technical or managerial innovation, a firm can produce more today from a given amount of inputs than it could in the past.

Relative Productivity

Throughout this chapter, we've assumed that firms produce efficiently. A firm must produce efficiently if it is to maximize its profit. Even if each firm in a market produces as efficiently as possible, however, firms may not be equally *productive*, in the sense that one firm can produce more than another from a given amount of inputs.

A firm may be more productive than others if its manager knows a better way to organize production or if it is the only firm with access to a new invention. Union-mandated work rules, government regulations, other institutional restrictions, or racial or gender discrimination that affect only some firms may lower the relative productivity of those firms.

We can measure the *relative productivity* of a firm by expressing the firm's actual output, q, as a percentage of the output that the most productive firm in the industry could have produced, q^*, from the same amount of inputs: $100q/q^*$. The most productive firm in an industry has a relative productivity measure of 100% (= $100q^*/q^*$ percent).

Caves and Barton (1990) report that the average productivity of firms across U.S. manufacturing industries ranges from 63% to 99%. That is, in the manufacturing industry with the most diverse firms, the average firm produces slightly less than two-thirds as much as the most productive firm, whereas in the manufacturing industry with the most homogeneous firms, all firms are nearly equally productive.

Differences in productivity across markets may be due to differences in the degree of competition. In competitive markets, in which many firms can enter and exit the market easily, less productive firms lose money and are driven out of business, so the firms that are actually producing are equally productive (as Chapter 8 shows). In a less competitive oligopoly market, with few firms and no possibility of entry by new firms, a less productive firm may be able to survive, so firms with varying levels of productivity are observed.

In communist and other government-managed economies, in which firms are not required to maximize profits, inefficient firms may survive. For example, a study of productivity in 48 medium-size, machine-building state enterprises in China (Kalirajan and Obwona, 1994) found that the productivity measure ranges from 21% to 100%, with an average of 55%.

Application

GERMAN VERSUS BRITISH PRODUCTIVITY

Even within a single company, a plant in one country may be more productive than a plant in another country operating under different institutional rules and with a different workforce. The Ford Motor Company has virtually identical plants in Saarlouis, Germany, and Halewood, England. Each plant manufactured Escorts on production lines equipped with robot welders and automated presses that punch out parts.

Although the plants appeared to be identical, the German plant was much more productive in 1981. Whereas the German plant produced 1,200 cars a day (more than the 1,015 that Ford planners predicted) using 7,762 workers, the British plant manufactured only 800 cars a day using 10,040 workers. Equivalently stated, it took only about 21 hours of labor ($L/q = 1/AP_L$) to produce an Escort in Germany compared to about 40 hours in Britain.

Moreover, the German plant turned out the highest-quality cars of any Ford plant. The German Escorts averaged only half as many demerits from quality inspectors as British- or American-made Escorts.

According to Ford officials, the difference was due to the attitudes of the workers. In Britain, under pressure from labor, management used two daily shifts so that no one worked on Friday nights; in Germany, there was no such pressure. Labor unrest was more common in Britain. There were 20 strikes throughout Britain in the first nine months of 1981—though that was considerably fewer than the 300 strikes in 1976—whereas no strikes occurred in Germany.

British labor leaders argued that the German plant was unsafe. The British union summoned a company doctor to rule, for example, that two workers were required to lift the hood onto a body—a job that a single worker performed at the German plant. Rejecting this claim, a reporter contended that one worker lifted and the other merely watched at the British plant.

Because the British plant was less productive than its German counterpart, we could conclude that the British plant was operating inefficiently. An alternative interpretation is that the British plant was operating as efficiently as possible given the institutional and union rules.

Innovations

In its production process, a firm tries to use the best available technological and managerial knowledge. An advance in knowledge that allows more output to be produced with the same level of inputs is called **technical progress**. The invention of new products is a form of technical innovation. The use of robotic arms increases the number of automobiles produced with a given amount of labor and raw materials. Better *management or organization of the production process* similarly allows the firm to produce more output from given levels of inputs.

Technical Progress. A technological innovation changes the production process. Last year a firm produced

$$q_1 = f(L, K)$$

units of output using L units of labor services and K units of capital service. Due to a new invention that the firm uses, this year's production function differs from last year's, so the firm produces 10% more output with the same inputs:

$$q_2 = 1.1f(L, K).$$

This firm has experienced *neutral technical change*, in which it can produce more output using the same ratio of inputs. For example, a technical innovation in the form of a new printing press allows more output to be produced using the same ratio of inputs as before: one worker to one printing press.

Nonneutral technical changes are innovations that alter the proportion in which inputs are used. If a printing press that required two people to operate is replaced by one that can be run by a single worker, the technical change is *labor-saving*. The ratio of labor to other inputs used to produce a given level of output falls after the innovation. Similarly, the ratio of output to labor, the average product of labor, rises.

In our neutral technical change example, the firm's rate of growth of output was $10\% = \Delta q/q_1 = [1.1f(L, K) - f(L, K)]/f(L, K)$ in one year due to the technical change. Table 6.3 shows estimates of the annual rate at which output grew for given levels of inputs in the United States and Poland. The rates of productivity growth differ across industries within a country and across countries for a particular industry. The faster growth in Poland is due to playing catch-up: Starting with antiquated technology and little capital, Poland invested heavily in modern technologies and capital during this period.

Table 6.3 Annual Rates of Productivity Growth

United States, 1949–1983[a]		Poland, 1962–1983[b]	
Industry	*Growth Rate, %*	*Industry*	*Growth Rate, %*
Food and kindred products	0.7	Food and tobacco	4.9
Tobacco manufactures	0.2		
Lumber and wood products	1.3	Wood and paper	2.5
Paper and allied products	0.9		
Chemicals and allied products	1.5	Chemicals	8.0
Total manufacturing	1.1	Light industry	1.3

[a]Gullickson and Harper (1987) calculated the annual percentage change in output for various U.S. manufacturing industries, holding labor, capital, energy, material, and business service inputs constant.
[b]Terrell (1993) estimated the annual percentage change in output for various Polish industries, holding labor, domestic capital, and Western capital constant.

Application

NONNEUTRAL TECHNICAL CHANGE IN PIN MANUFACTURING

The history of the humble pin illustrates how nonneutral technological change caused the average product of labor to rise over several centuries (Pratten, 1980). According to Adam Smith, a factory of 10 highly specialized workers using only handheld tools could produce 48,000 pins per day in 1776, so the average product of labor was 4,800 pins per day.

In 1824, Samuel W. Wright patented a solid-head pin machine in England. In 1830, this machine produced 45 pins a minute. Machines were substituted for workers. As a consequence of this mechanization, the average product of labor rose 70%.

Thanks to further technical advances, machines in 1900 produced 180 pins a minute, and current machines can turn out over 500 per minute. The number of machines a single worker can control has also increased. Currently, one worker can control as many as 24 machines. As a consequence, 50 people are now employed in Britain producing more pins than thousands of workers could manufacture in the early 1800s. By 1980, the average product of labor at one plant in Britain was 800,000 pins per day, an increase of 167 times over the average product in Adam Smith's day. Stated another way, the average product of labor has increased by nearly 2.7% each year for two centuries.

Organizational Change. Organizational changes may also alter the production function and increase the amount of output produced by a given amount of inputs. Organizational innovations have been very important in automobile manufacturing.

In the early 1900s, Henry Ford revolutionized mass production through two organizational innovations. First, he introduced interchangeable parts, which cut the time required to install parts because workers no longer had to file or machine individually made parts to get them to fit.

Second, Ford introduced a conveyor belt and an assembly line to his production process. Before Ford, workers walked around the car, and each worker performed many assembly activities. In Ford's plant, each worker specialized in a single activity such as attaching the right rear fender to the chassis. A conveyor belt moved the car at a constant speed from worker to worker along the assembly line. Because his workers gained proficiency from specializing in only a few activities and because the conveyor belts reduced the number of movements workers had to make, Ford could produce more automobiles with the same number of workers. By the early 1920s, Ford had cut the cost of a car by more than two-thirds and had increased production from fewer than a thousand cars per year to two million per year.

DELL COMPUTER'S ORGANIZATIONAL INNOVATIONS

Michael Dell, the president of Dell Computer, has become rich by innovating in organizational practices rather than by producing the most technologically advanced computers. Dell Computer is probably the world's most efficient

personal computer manufacturer due in large part to two organizational innovations: building to order and just-in-time delivery. Dell made its name by selling directly to customers and allowing them to specify the features they wanted on their personal computer. It has adopted and extended the use of just-in-time inventories, a practice developed by Toyota and other Japanese auto manufacturers. Upon getting an order, Dell uses the Internet to tell its suppliers which parts it needs, and receives delivery within an hour and a half. As Michael Dell writes, "Keep your friends close, and your suppliers closer." Its just-in-time strategy virtually eliminates the need for Dell to maintain any inventory of parts and finished products.

Consequently, Dell has eliminated warehouses in its factories, cutting the number of buildings it needs in each factory from two to one. If a particular component is not available, Dell executives instruct their salespeople to offer discounts on a computer with a better component or, in dire emergencies, a free upgrade. To further facilitate its manufacturing process, the company uses special hydraulic tools, conveyor belts, and tracks, cutting human intervention in half. Workers snap computer components into place, no longer having to use screwdrivers. They can assemble a computer in two to three minutes. From start to finish, Dell achieves a four-hour production cycle.

When it introduced its new built-to-order process in 2000, Dell expected to save $15 million in the first six months and $150 million within three years. Most of Dell's rivals *outsource* by having other firms manufacture computers for them. Reportedly, Dell's operating expenses as a fraction of its sales is only 9.9%, compared to Gateway's 27% and HP/Compaq's in the high teens.

Summary

1. **The ownership and management of firms:** Firms are either sole proprietorships, partnerships, or corporations. In smaller firms (particularly sole proprietorships and partnerships), the owners usually run the company. In large firms (such as most corporations), the owners hire managers to run the firms. Owners want to maximize profits. If managers have different objectives than owners, owners must keep a close watch over managers to ensure that profits are maximized.

2. **Production:** Inputs, or factors of production—labor, capital, and materials—are combined to produce output using the current state of knowledge about technology and management. To maximize profits, a firm must produce as efficiently as possible: It must get the maximum amount of output from the inputs it uses, given existing knowledge. A firm may have access to many efficient production processes that use different combinations of inputs to produce a given level of output. New technologies or new forms of organization can increase the amount of output that can be produced from a given combination of inputs. A production function shows how much output can be produced efficiently from various levels of

inputs. A firm can vary all its inputs in the long run but only some of them in the short run.

3. **Short-run production: one variable and one fixed input:** In the short run, a firm cannot adjust the quantity of some inputs, such as capital. The firm varies its output by adjusting its variable inputs, such as labor. If all factors are fixed except labor, and a firm that was using very little labor increases its use of labor, its output may rise more than in proportion to the increase in labor because of greater specialization of workers. Eventually, however, as more workers are hired, the workers get in each other's way or wait to share equipment, so output increases by smaller and smaller amounts. This latter phenomenon is described by the law of diminishing marginal returns: The marginal product of an input—the extra output from the last unit of input—eventually decreases as more of that input is used, holding other inputs fixed.
4. **Long-run production: two variable inputs:** In the long run, when all inputs are variable, firms can substitute between inputs. An isoquant shows the combinations of inputs that can produce a given level of output. The marginal rate of technical substitution is the absolute value of the slope of the isoquant. Usually, the more of one input the firm uses, the more difficult it is to substitute that input for another input. That is, there are diminishing marginal rates of technical substitution as the firm uses more of one input.
5. **Returns to scale:** If, when a firm increases all inputs in proportion, its output increases by the same proportion, the production process is said to exhibit constant returns to scale. If output increases less than in proportion to inputs, the production process has decreasing returns to scale; if it increases more than in proportion, it has increasing returns to scale. All three types of returns to scale are commonly seen in actual industries. Many production processes exhibit first increasing, then constant, and finally decreasing returns to scale as the size of the firm increases.
6. **Productivity and technical change:** Although all firms in an industry produce efficiently, given what they know and the institutional and other constraints they face, some firms may be more productive than others: They can produce more output from a given bundle of inputs. Due to innovations such as technical progress or new means of organizing production, a firm can produce more today than it could in the past from the same bundle of inputs. Such innovations change the production function.

Questions

QUIZ

1. If each extra worker produces an extra unit of output, how do the total product of labor, average product of labor, and marginal product of labor vary with labor?
2. Each extra worker produces an extra unit of output up to six workers. After six, no additional output is produced. Draw the total product of labor, average product of labor, and marginal product of labor curves.
3. What is the difference between an isoquant and an indifference curve?
4. Why must isoquants be thin? (*Hint*: See the explanation of why indifference curves must be thin in Chapter 4.)
5. Suppose that a firm has a fixed-proportions production function, in which one unit of output is produced using one worker and two units of capital. If the firm has an extra worker and no more capital, it still can produce only one unit of output. Similarly, one more unit of capital does the firm no good.
   a. Draw the isoquants for this production function.
   b. Draw the total product, average product, and marginal product of labor curves (you will probably want to use two diagrams) for this production function.
6. To produce a recorded CD, $Q = 1$, a firm uses one blank disk, $D = 1$, and the services of a recording machine, $M = 1$, for one hour. Draw an isoquant for this production process. Explain the reason for its shape.
7. Michelle's business produces ceramic cups using labor, clay, and a kiln. She can manufacture 25 cups a day with one worker and 35 with two

workers. Does her production process illustrate *diminishing returns to scale* or *diminishing marginal returns to scale*? What is the likely explanation for why output doesn't increase proportionately with the number of workers?

8. Draw a circle in a diagram with labor services on one axis and capital services on the other. This circle represents all the combinations of labor and capital that produce 100 units of output. Now draw the isoquant for 100 units of output. (*Hint*: Remember that the isoquant includes only the efficient combinations of labor and capital.)
9. In a manufacturing plant, workers use a specialized machine to produce belts. A new machine is invented that is laborsaving. With the new machine, the firm can use fewer workers and still produce the same number of belts as it did using the old machine. In the long run, both labor and capital (the machine) are variable. From what you know, what is the effect of this invention on the AP_L, MP_L, and returns to scale? If you require more information to answer this question, specify what you need to know.
10. Show in a diagram that a production function can have diminishing marginal returns to a factor and constant returns to scale.
11. If a firm lays off workers during a recession, how will the firm's marginal product of labor change?
12. During recessions, American firms lay off a larger proportion of their workers than Japanese firms do. (It has been claimed that Japanese firms continue to produce at high levels and store the output or sell it at relatively low prices during the recession.) Assuming that the production function remains unchanged over a period that is long enough to include many recessions and expansions, would you expect the average product of labor to be higher in Japan or the United States? Why?
13. Does it follow that, because we observe that the average product of labor is higher for Firm 1 than for Firm 2, Firm 1 is more productive in the sense that it can produce more output from a given amount of inputs? Why?
14. *Review* (Chapter 5): Melissa eats eggs and toast for breakfast and insists on having three pieces of toast for every two eggs she eats. If the price of eggs increases but we compensate Melissa to make her just as "happy" as she was before the price change, what happens to her consumption of eggs? Draw a graph and explain your diagram. Does the change in her consumption reflect a substitution or an income effect?

★15. *Review* (Chapters 4, 5, and 6): If we plot the profit of a firm against the number of vacation days the owner takes, we find that profit first rises with vacation days (a few days of vacation improve the owner's effectiveness as a manager the rest of the year) but eventually falls as the owner takes more vacation days. If the owner has usual-shaped indifference curves between profit and vacation days, will the manager take the number of vacation days that maximizes profit? If so, why? If not, what will the owner do and why?

Problems

16. Suppose that the production function is $q = L^{3/4}K^{1/4}$.
    a. What is the average product of labor, holding capital fixed at $\bar{K}$?
    b. What is the marginal product of labor? (*Hint*: Calculate how much q changes as L increases by 1 unit, or use calculus.)
    c. Does this production function have increasing, constant, or decreasing returns to scale?
17. What is the production function if L and K are perfect substitutes and each unit of q requires 1 unit of L or 1 unit of K (or a combination of these inputs that adds to 1)?
18. At $L = 4$, $K = 4$, the marginal product of labor is 2 and the marginal product of capital is 3. What is the marginal rate of technical substitution?
19. In the short run, a firm cannot vary its capital, $K = 2$, but can vary its labor, L. It produces output q. Explain why the firm will or will not experience diminishing marginal returns to labor in the short run if its production function is

a. $q = 10L + K$
b. $q = L^{\frac{1}{2}}K^{\frac{1}{2}}$

20. Under what conditions do the following production functions exhibit decreasing, constant, or increasing returns to scale?
a. $q = L + K$
b. $q = L^{\alpha}K^{\beta}$
c. $q = L + L^{\alpha}K^{\beta} + K$

21. Firm 1 and Firm 2 use the same type of production function, but Firm 1 is only 90% as productive as Firm 2. That is, the production function of Firm 2 is $q_2 = f(L, K)$, and the production function of Firm 1 is $q_1 = 0.9f(L, K)$. At a particular level of inputs, how does the marginal product of labor differ between the firms?