CHAPTER 6

Production

In the last three chapters, we focused on the demand side of the market-the preferences and behavior of consumers. Now we turn to the supply side and examine the behavior of producers. We will see how firms can organize their production efficiently and how their costs of production change as input prices and the level of output change. We will also see that there are strong similarities between the optimizing decisions of firms and those of consumers-understanding consumer behavior will help us understand producer behavior.

The theory of production and cost is central to the economic management of the firm. Just consider some of the problems that a company like General Motors faces regularly. How much assembly-line machinery and how much labor should it use in its new automobile plants? If it wants to increase production, should it hire more workers, or should it also construct new plants? Does it make more sense for one automobile plant to produce different models, or should each model be manufactured in a separate plant? What should GM expect its costs to be during the coming year, and how are these costs likely to change over time and be affected by the level of production? These questions apply not only to business firms, but also to other producers of goods and services, such as governments and nonprofit agencies.

In this chapter we study the firm's production technology-the physical relationship that describes how inputs (such as labor and capital) are transformed into outputs (such as cars and televisions). We do this in several steps. First, we show how the production technology can be represented in the form of a production function-a compact description that facilitates the analysis. Then, we use the production function to show how the firm's output changes when first one and then all the inputs are varied. We will be particularly concerned with the scale of the firm's operation. Are there technological advantages that make the firm more productive as its scale increases?

6.1 *The Technology of Production*

In the production process, firms turn *inputs,* which are also called *factors of production,* into *outputs* (or products). For example, a bakery uses inputs that include the labor of its workers; raw materials, such as flour and sugar; and the capital invested in its ovens, mixers, and other equipment to produce such outputs as bread, cakes, and pastries.

We can divide inputs into the broad categories of labor, materials, and capital, each of which might include more narrow subdivisions. Labor inputs include skilled workers (carpenters, engineers) and unskilled workers (agricultural workers), as well as the entrepreneurial efforts of the firm's managers. Materials include steel, plastics, electricity, water, and any other goods that the firm buys and transforms into a final product. Capital includes buildings, equipment, and inventories.

The relationship between the inputs to the production process and the resulting output is described by a production function. A *production function* indicates the output Q that a firm produces for every specified combination of inputs. For simplicity, we will assume that there are two inputs, labor L and capital K. We can then write the production function as

$$Q=F(K,L) \tag{6.1}$$

This equation relates the quantity of output to the quantities of the two inputs, capital and labor. For example, the production function might describe the number of personal computers that can be produced each year with a 10,000-square-foot plant and a specific amount of assembly-line labor employed during the year. Or it might describe the crop that a farmer can obtain with a specific amount of machinery and workers.

The production function allows for inputs to be combined in varying proportions to produce an output in many ways. For example, wine can be produced in a labor-intensive way by people stomping the grapes, or in a capital-intensive way by machines squashing the grapes. Note that equation (6.1) applies to a *given technology* (i.e., a given state of knowledge about the various methods that might be used to transform inputs into outputs). As the technology becomes more advanced and the production function changes, a firm can obtain more output for a given set of inputs. For example, a new, faster computer chip may allow a hardware manufacturer to produce more high-speed computers in a given period of time.

Production functions describe what is *technically feasible* when the firm operates *efficiently;* that is, when the firm uses each combination of inputs as effectively as possible. Because production functions describe the maximum output feasible for a given set of inputs in a *technically efficient* manner, it follows that inputs will not be used if they decrease output. The presumption that production is always technically efficient need not always hold, but it is reasonable to expect that profit-seeking firms will not waste resources.

6.2 Isoquants

Let's begin by examining the firm's production technology when it uses two inputs and can vary both of them. Suppose, for example, that the inputs are labor and capital, and that they are used to produce food. Table 6.1 tabulates the output achievable for various combinations of inputs.

Labor inputs are listed across the top row, capital inputs down the column on the left. Each entry in the table is the maximum (technically efficient) output that can be produced per time period (say, a year) with each combination of labor and capital used over that time period. (For example, 4 units of labor per year and 2 units of capital per year yield 85 units of food per year.) Reading along each row, we see that output increases as labor inputs are increased, with capital inputs fixed. Reading down each column, we see that output also increases as capital inputs are increased, with labor inputs fixed.

The information contained in Table 6.1 can also be represented graphically using isoquants. *An isoquant is a curve that shows all the possible combinations of inputs that yield the same output.* Figure 6.1 shows three isoquants. (Each axis in the figure measures the quantity of inputs.) These isoquants are based on the data in Table 6.1, but have been drawn as smooth curves to allow for the use of fractional amounts of inputs.

For example, isoquant Q_1 shows all combinations of labor per year and capital per year that together yield 55 units of output per year. Two of these points, *A* and *D*, correspond to Table 6.1. At *A*, 1 unit of labor and 3 units of capital yield 55 units of output; whereas at *D*, the same output is produced from 3 units of labor and 1 unit of capital. Isoquant Q_2 shows all combinations of inputs that yield 75 units of output and corresponds to the four combinations of

TABLE 6.1 Production with Two Variable Inputs

	Labor Input				
Capital Input	1	2	3	4	5
1	20	40	55	65	75
2	40	60	75	85	90
3	55	75	90	100	105
4	65	85	100	110	115
5	75	90	105	115	120

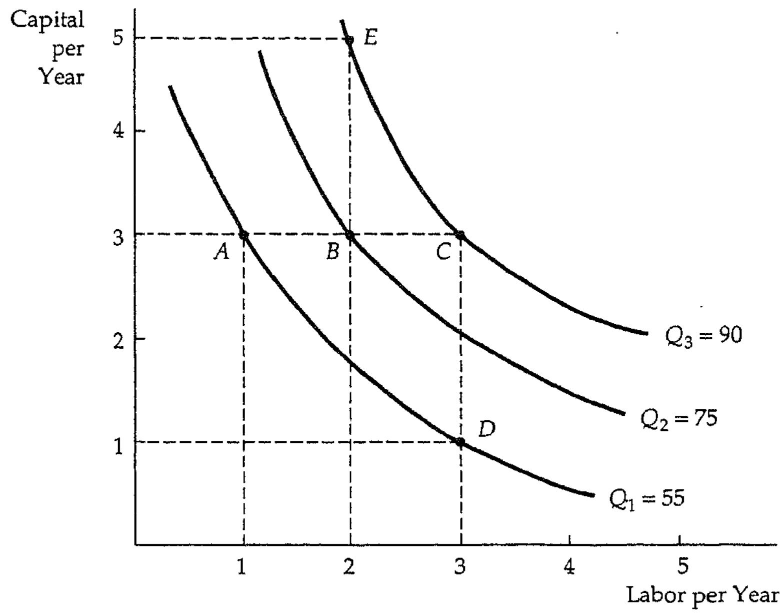

FIGURE 6.1 Production with Two Variable Inputs. Production isoquants show the various combinations of inputs necessary for the firm to produce a given output. A set of isoquants, or isoquant map, describes the firm's production function. Output increases as one moves from isoquant Q_1 (55 units per year) to isoquant Q_2 (75 units per year) and to isoquant Q_3 (90 units per year).

labor and capital italicized in the table (e.g., at *B*, where 2 units of capital and 3 units of labor are combined). Isoquant Q_2 lies above and to the right of Q_1 because it takes more labor and/or capital to obtain a higher level of output. Finally, isoquant (33 shows labor-capital combinations that yield 90 units of output. Point *C* involves 3 units of labor and 3 units of capital, while Point *E* involves only 2 units of labor and 5 units of capital. Note that inputs and output are *flows.* The firm uses certain amounts of labor and capital *each year* to produce an amount of output over that year. To simplify, we will frequently ignore the reference to time, and just refer to amounts of labor, capital, and output.

Isoquants are similar to the indifference curves that we used to study consumer theory. Where indifference curves order levels of satisfaction from low to high, isoquants order levels of output. However, unlike indifference curves, each isoquant is associated with *a specific level of output. By* contrast, the numerical labels attached to indifference curves are meaningful only in an ordinal way-higher levels of utility are associated with higher indifference curves, but we cannot measure a specific level of utility the way we can measure a specific level of output with an isoquant

An *isoquant map* is a set of isoquants, each of which shows the maximum output that can be achieved for any set of inputs. An isoquant map is another way of describing a production function, just as an indifference map is a way of describing a utility function. Each isoquant corresponds to a different level of output, and the level of output increases as you move up and to the right in the figure.

Isoquants show the flexibility that firms have when making production decisions-firms can usually obtain a particular output using various combinations of inputs. It is important for the managers of a firm to understand the nature of this flexibility. For example, fast-food restaurants have recently faced shortages of young, low-wage employees. The companies have responded by automating, for example by adding salad bars or by introducing more sophisticated cooking equipment. They have also recruited older people to fill these positions. As we discuss in Chapters 7 and 8, by taking this flexibility in the production process into account, managers can choose input combinations that minimize cost and maximize profit.

The Short Run Versus the Long Run

It is important to distinguish between the short and long run when analyzing production. The *short run* refers to a period of time in which one or more factors of production cannot be changed. Factors that cannot be varied over this period are called *fixed inputs*. A firm's capital, for example, usually requires time to change-a new factory must be planned and built, machinery and other equipment must be ordered and delivered, all of which can take a year or more. The *long run* is the amount of time needed to make all inputs variable. In the short run, firms vary the intensity with which they utilize a given plant and machinery; in the long run, they vary the size of the plant. All fixed inputs in the short run represent the outcomes of previous long-run decisions based on firms' estimates of what they could profitably produce and sell.

There is no specific time period, such as one year, that separates the short run from the long run. Rather, one must distinguish them on a case-by-case basis. For example, the long run can be as brief as a day or two for a child's lemonade stand, or as long as five or ten years for a petrochemical producer or an automobile manufacturer.

6.3 *Production with One Variable Input (Labor)*

Let's consider the case in which capital is fixed, but labor is variable, so that the firm can produce more output by increasing its labor input. Imagine, for example, that you are managing a clothing plant. You have a fixed amount of equipment, but you can hire more labor or less to sew and to run the machines. You have to decide how much labor to hire and how much clothing to produce. To make the decision, you will need to know how the amount of output Q increases (if at all), as the input of labor L increases.

Table 6.2 gives this information. The first three columns show the amount of output that can be produced in one month with different amounts of labor, and with capital fixed at ten units. (The first column shows the amount of labor, the second the fixed amount of capital, and the third output.) When labor input is zero, output is also zero. Then output increases as labor is increased up to an input of eight units. Beyond that point, total output declines: While initially each unit of labor can take greater and greater advantage of the existing machinery and plant, after a certain point, additional labor is no longer useful and indeed can be counterproductive. (Five people can run an assembly line better than two, but ten people may get in each other's way.)

Average and Marginal Products

The contribution that labor makes to the production process can be described in terms of the average and marginal products of labor. The fourth column in Table 6.2 shows the *average product of labor* AP_L, which is the output per unit of labor input. The average product is calculated by dividing the total output Q by the total input of labor, L. In our example the average product increases initially but falls when the labor input becomes greater than 4. The fifth column shows the *marginal product of labor* MP_L. This is the *additional* output produced as the labor input is increased by one unit. For example, with capital fixed at 10 units, when the labor input increases from 2 to 3, total output increases from 30 to 60, creating an additional output of 3Q (60 - 30) units. The marginal product of labor can be written as $\Delta Q/\Delta L$ (i.e., the change in output ΔQ resulting from a one-unit increase in labor input ΔL).

TABLE 6.2 Production with One Variable Input

Amount of Labor (L)	Amount of Capital (K)	Total Output (Q)	Average Product (Q/L)	Marginal Product ($\Delta Q/\Delta L$)
0	10	0		—
1	10	10	10	10
2	10	30	15	20
3	10	60	20	30
4	10	80	20	20
5	10	95	19	15
6	10	108	18	13
7	10	112	16	4
8	10	112	14	0
9	10	108	12	-4
10	10	100	10	-8

Remember that the marginal product of labor depends on the amount of capital used. If the capital input increased from 10 to 20, for example, the marginal product of labor would most likely increase. The reason is that additional workers are likely to be more productive if they have more capital to use. Like the average product, the marginal product first increases then falls, in this case after the third unit of labor.

To summarize:

Average Product of Labor	= Output/Labor Input = Q/L
Marginal Product of Labor	= Change in Output/Change in Labor Input = $\Delta Q/\Delta L$

Figure 6.2 plots the information contained in Table 6.2. (We have connected all the points in the figure with solid lines.) Figure 6.2a shows that output increases until it reaches the maximum output of 112; thereafter it diminishes. That portion of the total output is dashed to denote that production past an output of 8 is not technically efficient and therefore is not part of the production function; technical efficiency rules out negative marginal products. Figure 6.2b shows the average and marginal product curves. (The units of the vertical axis have changed from output to output per unit of labor.) Note that the marginal product is always positive when output is increasing, and it is negative when output is decreasing.

It is no coincidence that the marginal product curve crosses the horizontal axis of the graph at the point of maximum total product. This happens because adding a worker to a production line in a manner that slows up the line and decreases total output implies a negative marginal product for that worker.

The average product and marginal product curves are closely related. *When the marginal product is greater than the average product, the average product is increasing,* as shown between outputs 1 and 4 in Figure 6.2b. For example, suppose that the only employee of an advertising firm can write 10 advertisements per day, so that initially 10 is the average product of labor. Now, a more productive employee is hired who can produce 20 ads per day. The marginal product of labor, 20 ads, is greater than the average, 10. And because both workers combine to produce 30 ads in two days of labor, the new average product has increased to 15 ads.

Similarly, *when the marginal product is less than the average product, the average product is decreasing,* as shown between outputs 4 and 10 in Figure 6.2b. In our previous example, had the first worker been the more productive of the two, the marginal product of the first worker would have been 20 ads and the marginal product of the second worker 10 ads. Since the marginal product (10 ads) would then be less than the average product (20 ads), the new average product would have fallen to 15 ads.

Because the marginal product is above the average product when the average product is increasing, and below the average product when the average prod-

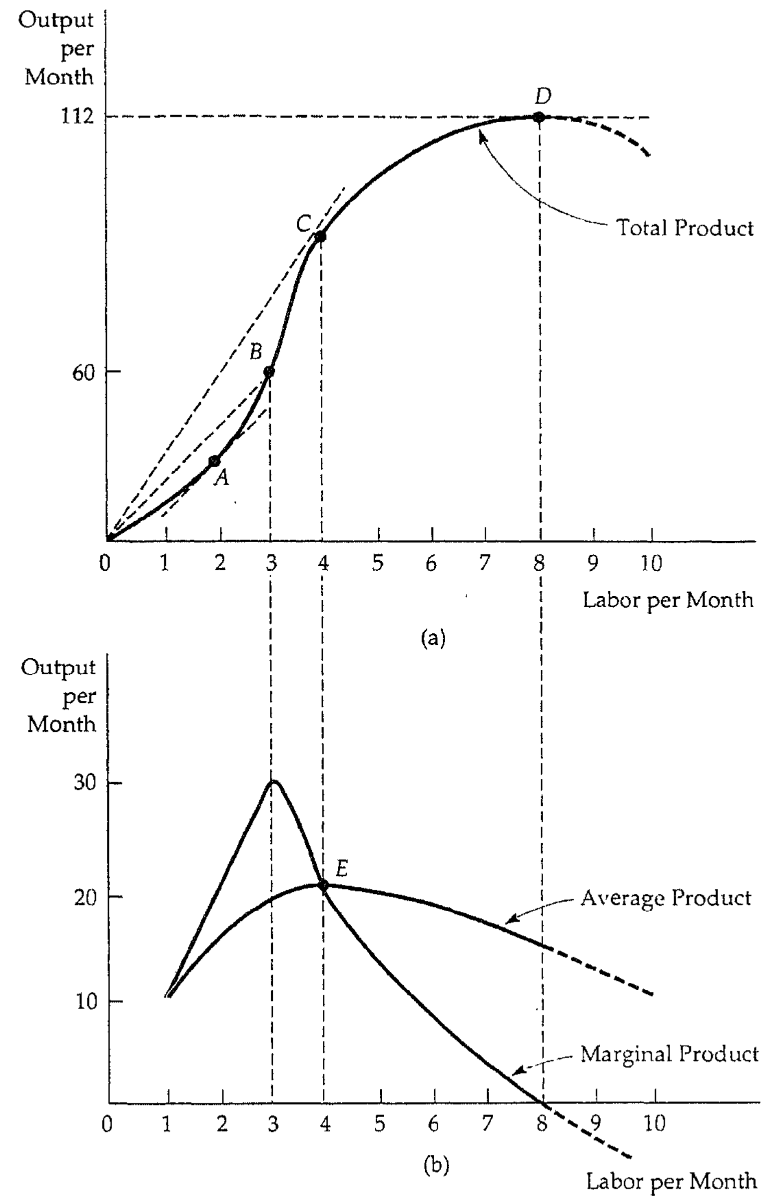

Figure 6.2 Production with One Variable Input. When all inputs other than labor are fixed, the total product curve in (a) shows the output produced for different amounts of labor input. The average and marginal products in (b) are obtained directly from the total product curve. At point B in (a) the average product of labor is given by the slope of the line from the origin to B.

uct is decreasing, it follows that the marginal product must equal the average product when the average product reaches its maximum. This happens at point *E* in Figure 6.2b.

The geometric relationship between the total product and the average and marginal product curves is shown in Figure 6.2a. The average product of labor is the total product divided by the quantity of labor input. For example, at *B* the average product is equal to the output of 60 divided by the input of 3, or 20 units of output per unit of labor input. But this is just the slope of the line running from the origin to B in Figure 6.2a. In general, *the average product of labor is given by the slope of the line drawn from the origin to the corresponding point on the total product curve.*

The marginal product of labor is the change in the total product resulting from an increase of one unit of labor. For example, at *A* the marginal product is 20 because the tangent to the total product curve has a slope of 20. In general, *the marginal product of labor at a point is given by the slope of the total product at that point.* We can see in Figure 6.2a that the marginal product of labor increases initially, peaks at an input of 3, and then declines as we move up the total product curve to *C* and *D*. At *D*, when total output is maximized, the slope of the tangent to the total product curve is 0, as is the marginal product. Beyond that point, the marginal product becomes negative.

Note the graphical relationship between average and marginal products. At *B*, the marginal product of labor (the slope of the tangent to the total product curve at *B*-not shown explicitly) is greater than the average product (dashed line *OB*). As a result, the average product of labor increases as we move from *B* to *C*. At *C*, the average and marginal products of labor are equal-the average product is the slope of the line from the origin *OC*, while the marginal product is the tangent to the total product curve at C (note the equality of the average and marginal products at point E in Figure 6.2b). Finally, as we move beyond *C* toward *D*, the average marginal product falls below the average product; you can check that the slope of the tangent to the total product curve at any point between *C* and *D* is lower than the slope of the line from the origin.

The Law of Diminishing Returns

A diminishing marginal product of labor (and a diminishing marginal product of other inputs) holds for most production processes; the phrase "the law of diminishing returns" is often used to describe this phenomenon. The *law of diminishing returns* states that as the use of an input increases (with other inputs fixed), a point will eventually be reached at which the resulting additions to output decrease. When the labor input is small (and capital is fixed), small increments in labor input add substantially to output as workers are allowed to develop specialized tasks. Eventually, however, the law of diminishing returns applies. When there are too many workers, some workers become ineffective, and the marginal product of labor falls.

The law of diminishing returns usually applies to the short run where at least one input is fixed. However, it can also apply to the long run. Even though all inputs are variable in the long run, a manager may still want to analyze production choices for which one or more inputs are unchanged. Suppose, for example, that only two plant sizes are feasible, and a manager must decide which to build. Then, the manager would want to know when diminishing returns will set in for each of the two options.

Do not confuse the law of diminishing returns with possible changes in the quality of labor as labor inputs are increased (as, for example, if the most highly qualified laborers are hired first, and the least qualified last). In our analysis of production, we have assumed that all labor inputs are of equal quality; diminishing returns result from limitations on the use of other fixed inputs (e.g., machinery), not from declines in worker quality. Also, do not confuse diminishing returns with negative returns. The law of diminishing returns describes a *declining* marginal product, but not necessarily a negative one.

The law of diminishing returns applies to a given production technology. Over time, however, inventions and other improvements in technology may allow the entire total product curve in Figure 6.2a to shift upward, so that more output can be produced with the same inputs. Figure 6.3 illustrates this.

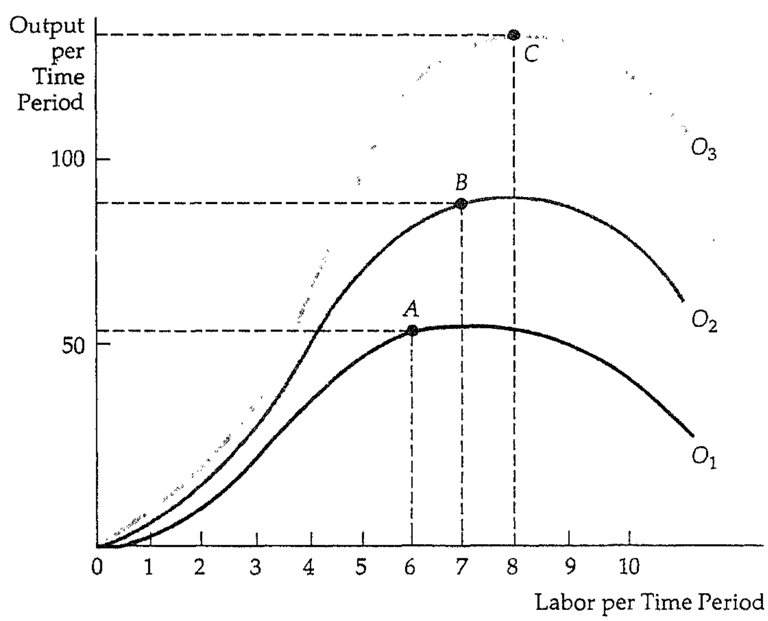

FIGURE 6.3 The Effect of Technological Improvement. Labor productivity (output per unit of labor) can increase if there are improvements in the technology, even though any given production process exhibits diminishing returns to labor. As we move from point *A* on curve O_1 to *B* on curve O_2 to *C* on curve O_3 over time,labor productivity increases.

Initially the output curve is given by O_1, but improvements in technology may allow the curve to shift upward, first to O_2 and later to O_3.

Suppose that over time as labor is increased in production, technological improvements are also being made. Then, output changes from *A* (with an input of 6 on curve O_1) to B (with an input of 7 on curve O_2) to *C* (with an input of 8 on curve O_3). The move from A to *B* to *C* relates an increase in labor input to an increase in output and makes it appear that there are no diminishing returns when there are. For inputs greater than 6, each of the individual product curves exhibits diminishing returns to labor.

The shifting of the total product curve hides the presence of diminishing returns, and suggests that they need not have any negative long-run implications for economic growth. In fact, as we discuss in Example 6.1, the failure to account for improvements in technology in the long run led British economist Thomas Malthus wrongly to predict dire consequences from continued population growth.

EXAMPLE 6.1 MALTHUS AND THE FOOD CRISIS

The law of diminishing returns was central to the thinking of economist Thomas Malthus (1766-1834).[1] Malthus believed that the limited amount of land on our globe would not be able to supply enough food as population grew and more laborers began to farm the land. Eventually as both the marginal and average productivity of labor fell and there were more mouths to feed, mass hunger and starvation would result. Fortunately, Malthus was wrong (although he was right about the diminishing returns to labor).

TABLE 6.3 Index of World Food Consumption Per Capita[2]

Year	Index
1948-1952	100
1955	109
1960	115
1965	116
1970	123
1978	128
1987	133
1991	142

[1] Thomas Malthus, *Essay on the Principle of Population,* 1798.

[2] All but the data for 1987 and 1991 appear as Table 4-1 in Julian Simon, *The Ultimate Resource* (Princeton: Princeton University Press, 1981). The original source for all the data is the UN Food and Agriculture Organization, *Production Yearbook,* and *World Agricultural Situation.*

Over the past century, technological improvements have dramatically altered the production of food in most countries (including developing countries, such as India), so that the average product of labor has increased. These improvements include new high-yielding and disease-resistant strains of seeds, better fertilizers, and better harvesting equipment. As Table 6.3 shows, overall food production throughout the world has outpaced population growth more or less continually since the end of World War II.

Some of the increase in food production has been due to small increases in the amount of land devoted to farming. For example, from 1961 to 1975, the percentage of land devoted to agriculture increased from 32.9 percent to 33.3 percent in Africa, from 19.6 percent to 22.4 percent in Latin America, and from 21.9 percent to 22.6 percent in the Far East.[3] However, during the same perioa. the percentage of land devoted to agriculture fell from 26.1 percent to 25.5 percent in North America, and from 46.3 percent to 43.7 percent in Western Europe. Clearly most of the improvement in food output is due to improved technology and not to increases in land used for agriculture.

Hunger remains a severe problem in some areas, such as the Sahel region of Africa, in part because of the low productivity of labor there. Although other countries produce an agricultural surplus, mass hunger still occurs because of the difficulty of redistributing foods from more to less productive regions of the world, and because of the low incomes of those less productive regions.

Labor Productivity

We sometimes measure the average product of labor for an industry or for the economy as a whole; then we refer to the results as *labor productivity.* Because the average product measures output per unit of labor input, it is relatively easy to measure (because total labor input and total output are the only pieces of information you need), and can provide useful comparisons across industries and for one industry over a long period. But productivity is especially important because it determines the real standard of living that a country can achieve for its citizens.

There is a simple link betweenproductivity and the standard of living. In any particular year, the aggregate value of goods and services produced by an economy is equal to the payments made to all factors of production, including wages, rental payments to capital, and profit to firms. But consumers ultimately receive these factor payments, whatever their form, As a result, consumers ill the aggregate can increase their rate of consumption in the long run only by increasing the total amount they produce.

[3] See Julian Simon, *The Ultimate Resource,* 83.

TABLE 6.4 Labor Productivity in Developed Countries[4]

	France	West Germany	Japan	United Kingdom	United States
			Output per Person (1990)		
	$17,431	$18,291	$17,634	$15,720	$21,449
Years		Annual Rate of Growth of Labor Productivity (%)			
1960-1973	5.4	4.5	8.6	3.6	2.2
1973-1991	2.5	2.1	2.8	1.8	0.4

As Table 6.4 shows, the level of output per person in the United States in 1990 was substantially higher than in other leading developed nations. But two patterns over the post-World War II period have been disturbing for Americans. First, productivity growth in the United States has been less rapid than productivity growth in most other developed nations. Second, productivity growth in the past two decades has been substantially lower in all developed countries than it has been in the past. Both these patterns can be seen clearly in the table.

Throughout the period 1960 to 1991, the rate of productivity growth in Japan has been the highest, followed by West Germany and France. United States productivity growth has been the lowest, even lower than that of the United Kingdom. How can this slowdown in growth be explained? And why has productivity growth in the United States been lower than elsewhere? The most important source of growth in labor productivity is the growth in the *stock of capital.* An increase in capital means more and better machinery, so that each worker can produce more output for each hour worked. Differences in the rate of growth of capital help to explain much of the data in Table 6.4. The greatest capital growth during the postwar period was in Japan and France, which were rebuilt substantially after World War II. To some extent, therefore, the lower rate of growth of productivity in the United States as compared with Japan, France, and West Germany is the result of these countries catching up after the war.

Productivity growth is also tied to the natural resource sector of the economy. As oil and other resources began to be depleted, output per worker fell somewhat. Environmental regulations (e.g., the need to restore land to its original condition after strip mining for coal) magnified this effect as the public became more concerned with the importance of cleaner air and water.

[4] See Angus Maddison, "Growth and Slowdown in Advanced Capitalist Countries," *Journal of Economic Literature* 25 (1987): 649-698. The more recent growth numbers are adjusted based on data from *Industrial Policy in OECD Countries, Annual Review,* 1992, and The Supplement to the OECD *Observer,* 1992, No. 176 (June/July 1992).

These factors explain part, but not all, of productivity growth over time and. across different countries. A full understanding of these differences remains an important research problem in economics.

EXAMPLE 6.2 WILL THE STANDARD OF LIVING IN THE UNITED STATES IMPROVE?

Will the standard of living in the United States continue to improve, or will the economy barely keep future generations from being worse off than we are today? The answer depends on the labor productivity of U.S. workers, because the real incomes of U.S. consumers increase only as fast as productivity does.

From 1979 to 1990, productivity growth in the United States was 0.4 percent, the lowest of all major developed countries.[5] What does this mean for the average U.S. worker? In a competitive international economy, this low growth will eventually lead to lower increases in workers' wages; otherwise, higher wages would have to be matched by higher prices. But these higher prices would not be competitive in today's world economy. The result is that workers will have to absorb most of the impact of low productivity growth.

We have seen how slow growth in capital investment leads to low productivity growth. But the decline in productivity growth in the United States has other causes particular to this country. This can best be understood if we look at three major production sectors of the economy. First, during 1945-1965, many workers left farms and entered manufacturing. Agriculture has lower productivity than manufacturing, so the shift created productivity growth. (The ratio of agricultural to industrial productivity was about 0.40 in 1948, and has not changed much since.) By 1965 few people were left on the farms who could move to manufacturing, so this source of growth was exhausted.

Second, the productivity of the U.S. construction sector has declined substantially. There is no consensus about the source of this decline-it may be due in part to problems with nuclear reactor construction, and in part to problems with the interstate highway system. Whatever the cause, construction productivity has fallen and does not seem likely to increase.

Third, the movement of workers into the service sector of the economy has also dampened productivity growth in the United States, since productivity in the service sector is approximately 60 percent of the national average. By 1990, for example, over 35 percent of all hours of work were spent on service industry jobs, with a substantial portion of these hours devoted to nursing and health care and to lawyers and accountants.

Overall, this suggests that much of the slowdown in productivity growth was inevitable, and that not all of it was bad. Nursing care may be a low-productivity industry, but it is one that our society considers important. Other sources of low productivity growth include a relatively inexperienced labor

[5] This discussion is based in part on Lester Thurow, 'The Productivity Problem,"*Technology Reviczu* (1980): 40-51; and Martin N. Baily, "What Has Happened to Productivity Growth," *Science* (Oct. 1986): 443-451.

force due to the postwar baby boom, and the inhibiting effects of government regulations involving health, safety, and the environment. Because the sources of low productivity growth are varied and complex, the standard of living cannot be increased simply by reversing what has happened in the past. But the future need not be bleak- Capital can be increased by tax policies that stimulate investment, and greater and more creative efforts can be made to encourage productivity-enhancing research and development.

6.4 *Production with Two Variable Inputs*

Now that we have seen the relationship between production and productivity, let's consider the firm's production technology in the long run, where both capital and labor inputs (instead of just labor) are variable. We can examine alternative ways of producing by looking at the shape of a series of isoquants.

Recall that an isoquant describes all combinations of inputs that yield the same level of output. The isoquants shown in Figure 6.4 are reproduced from Figure 6.1; they all slope downward because both labor and capital have positive marginal products. More of either input increases output; so if output is to be kept constant as more of one input is used, less of the other input must be used.

Diminishing .Returns

There are diminishing returns to both labor and capital in this example. To see why there are diminishing returns to labor, draw a horizontal line at a particular level of capital, say 3. Reading the levels of output from each isoquant as labor is increased, we note that each additional unit of labor generates less and less additional output. For example, when labor is increased from 1 unit to 2 (from *A* to *B*), output increases by 20 (from 55 to 75). However, when labor is increased by an additional unit (from *B* to *C*), output increases by only 15 (from 75 to 90). Thus,there are diminishing returns to labor both in the long and short run. Because adding one factor while holding the other factor constant eventually leads to lower and lower increments to output, the isoquant must become steeper, as more capital is added in place of labor, and flatter when labor is added in place of capital.

There are also diminishing returns to capital. With labor fixed, the marginal product of capital decreases as capital is increased. For example, when capital is increased from 1 to 2 and labor is held constant at 3, the marginal product of capital is initially 20 (75 - 55), but the marginal product falls to 15 (90 - 75) when capital is increased from 2 to 3.

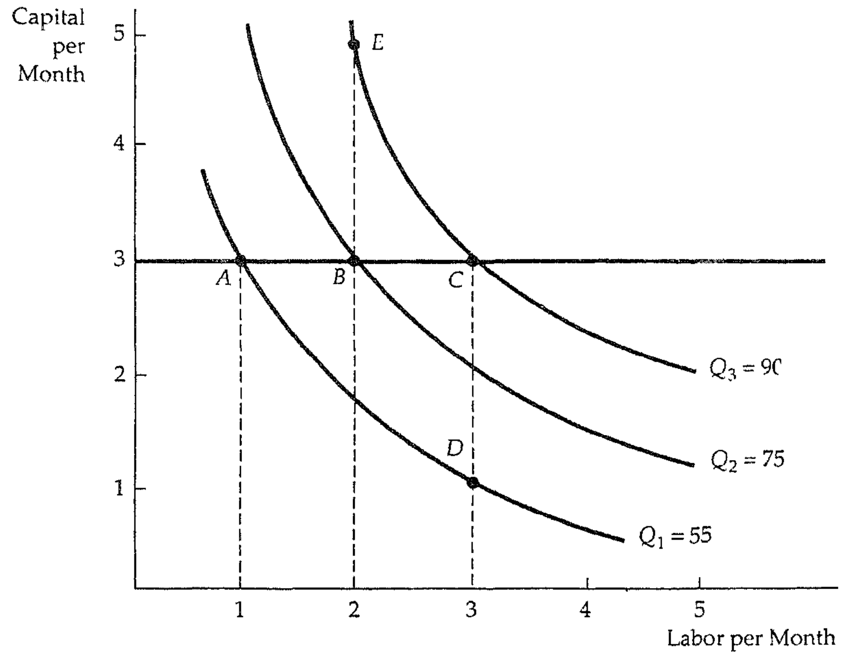

FIGURE 6.4 The Shape of Isoquante. In the long run when both labor and capital are variable, both factors of production can exhibit diminishing returns. As we move from *A* to *C*, there are diminishing returns to labor, and as we move from *D* to *C*, there are diminishing returns to capital.

Substitution Among Inputs

With two inputs that can be varied, a manager will want to consider substituting one input for another. The slope of each isoquant indicates how the quantity of one input can be traded off against the quantity of the other, while keeping output constant. When the negative sign is removed, we call the slope the marginal rate of technical substitution (MRTS). 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. This is analogous to the marginal rate of substitution (MRS) in consumer theory. Like the MRS, the MRTS is always measured as a positive quantity. In formal terms,

$$\begin{aligned} \text{MRTS} &= -\text{Change in Capital Input/Change in Labor Input} \\ &= -\Delta K/\Delta L \text{ (for a fixed level of } Q) \end{aligned}$$

where ΔK and ΔL are small changes in capital and labor along an isoquant.

Note that in Figure 6.5 the MRTS is equal to 2 when labor increases from 1 unit to 2, and output is fixed at 75. However, the MRTS falls to 1 when labor

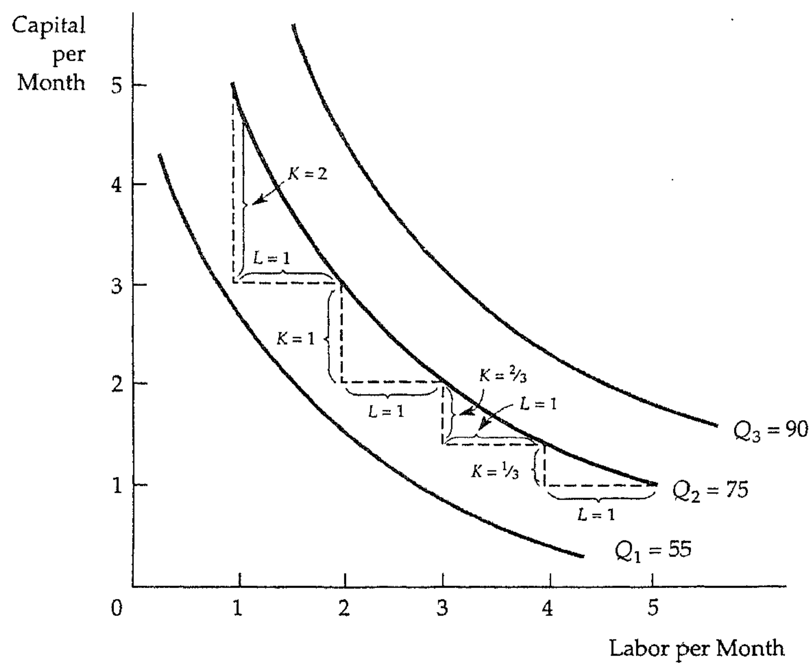

FIGURE 6.5 Marginal Rate of Technical Substitution. Isoquants are downward sloping and convex like indifference curves. The slope of the isoquant at any point measures the marginal rate of technical substitution, the ability of the firm to replace capital with labor while maintaining the same level of output. On isoquant Q_2, the marginal rate of technical substitution falls from 2 to 1 to 2/3 to 1/3.

is increased from 2 units to 3, and then declines to ⅔ and to ⅓. Clearly, as more and more labor replaces capital, labor becomes less productive and capital becomes relatively more productive. So less capital needs to be given up to keep constant the output from production, and the isoquant becomes flatter.

Isoquants are convex-the MRTS diminishes as. we move down along an isoquant. The diminishing MRTS tells us that the productivity that any one input can have is limited. As a lot of labor is added to the production process in place of capital, the productivity of labor falls. Similarly, when a lot of capital is added in place of labor, the productivity of capital falls. Production needs a balanced mix of both inputs.

As our discussion has just suggested, the MRTS is closely related to the marginal products of labor MP_L and capital MP_K. To see how, imagine adding some labor and reducing the amount of capital to keep output constant. The addition to output resulting from the increased labor input is equal to the additional output per unit of additional labor (the marginal product of labor) times the number of units of additional labor:

$$\text{Additional Output from Increased Use of Labor} = (\text{MP}_L)(\Delta L)$$

Similarly, the decrease in output resulting from the reduction in capital is the loss of output per unit reduction in capital (the marginal product of capital) times the number of units of capital reduction:

$$\text{Reduction in Output from Decreased Use of Capital} = (\text{MP}_K)(\Delta K)$$

Because we are keeping output constant by moving along an isoquant, the total change in output must be zero. Thus,

$$(\text{MP}_L)(\Delta L) + (\text{MP}_K)(\Delta K) = 0$$

Now, by rearranging terms we see that

$$(\text{MP}_L)/(\text{MP}_K) = -(\Delta K/\Delta L) = \text{MRTS} \tag{6.2}$$

Equation (6.2) tells us that as we move along an isoquant, continually replacing capital with labor in the production process, the marginal product of capital increases and the marginal product of labor decreases. The combined effect of both these changes is for the marginal rate of technical substitution to decrease as the isoquant becomes flatter.

Production Functions-Two Special Cases

Two extreme cases of production functions show the possible range of input substitution in the production process. In the first case, shown in Figure 6.6, inputs to production are *perfectly substitutable* for one another. Here the MRTS is constant at all points on an isoquant. As a result the same output (say Q_3) can be produced with mostly capital (at *A*), mostly labor (at *C*), or a balanced combination of both (at *B*). For example, a toll booth on a road or bridge might be run automatically or manned by a toll collector. Another example is musical instruments, which can be manufactured almost entirely with machine tools or with very few tools and highly skilled labor.

Figure 6.7 illustrates the opposite extreme, the *fixed-proportions production function*. In this case it is impossible to make any substitution among inputs. Each level of output requires a specific combination of labor and capital. Additional output cannot be obtained unless more capital and labor are added in specific proportions. As a result, the isoquants are L-shaped. An example is the reconstruction of concrete sidewalks using jackhammers. It takes one person to use a jackhammer-neither two people and one jackhammer nor one person and two jackhammers is likely to increase production.

In Figure 6.7 points *A*, *B*, and *C* represent technically efficient combinations of inputs. For example, to produce output Q_1, a quantity of labor L_1 and capital K_1 can be used, as at *A*. If capital stays fixed at K_1, adding more labor does not change output. Nor does adding capital with labor fixed at L_1. Thus, on the vertical and the horizontal segments of the L-shaped isoquants, either the

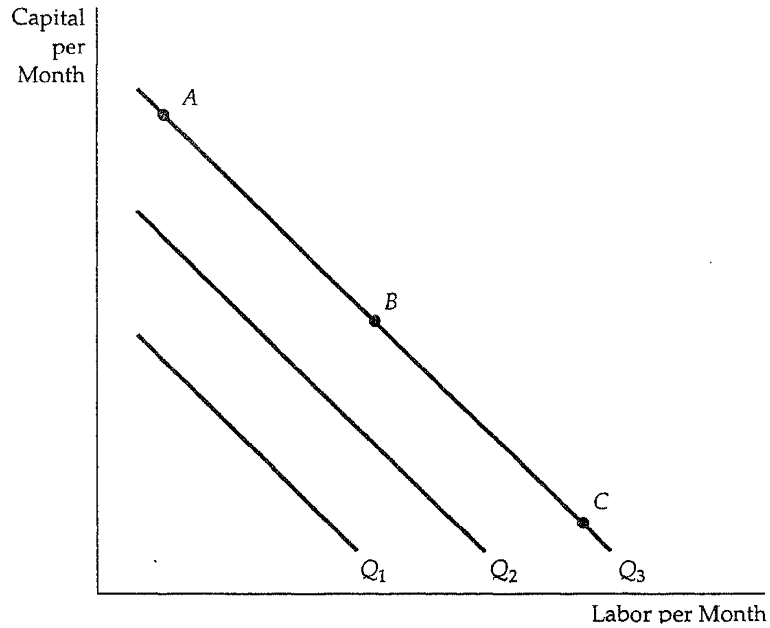

FIGURE 6.6 Isoquants When Inputs Are Perfectly Substitutable. When the isoquants are straight lines, the MRTS is constant. Hence the rate at which capital and labor can be substituted for each other is the same whatever level of inputs is being used.

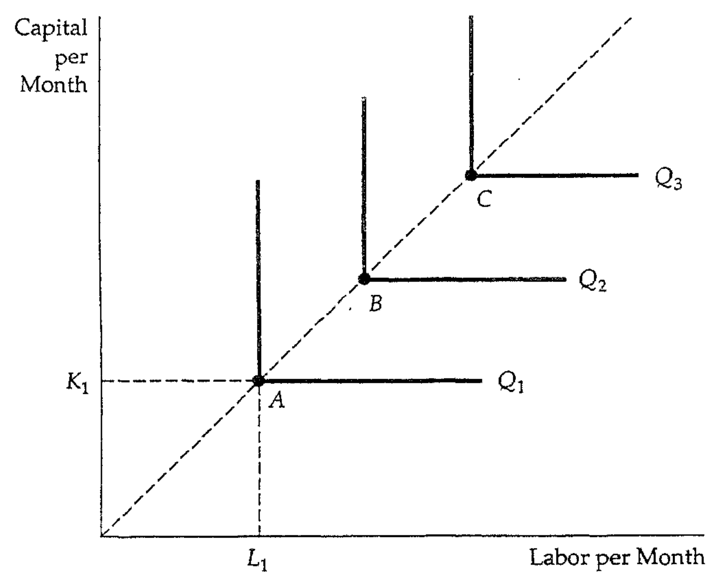

FIGURE 6.7 Fixed-Proportions Production Function. When the isoquants are L-shaped, only one combination of labor and capital can be used to produce a given output. Adding more labor does not increase output, nor does adding more capital alone.

marginal product of capital or the marginal product of labor is zero. Higher output results only when both labor and capital are added, as in the move from input combination A to input combination B.

The fixed-proportions production function describes situations in which the methods of production available to firms are limited. For example, the production of a television show might involve a certain mix of capital (camera and sound equipment, etc.) and labor (producer, director, actors, etc.). To make more television shows, all inputs to production must be increased proportionally. In particular, it would be difficult to increase capital inputs at the expense of labor, since actors are necessary inputs to production (except perhaps for animated films). Likewise, it would be difficult to substitute labor, for capital, since filmmaking today requires sophisticated film equipment.

EXAMPLE 6.3 A PRODUCTION FUNCTION FOR WHEAT

Crops can be produced using different methods. Food grown on large farms in the United States is usually produced with a *capital-intensive technology,* which involves substantial investments in capital, such as buildings and equipment, and relatively little input of labor. However, food can also be produced using very little capital (a hoe) and a lot of labor (several people with the patience and stamina to work the soil). One way to describe the agricultural production process is to show one isoquant (or more) that describes the combination of inputs that generates a given level of output (or several output levels). The description that follows comes from a production function for wheat that was estimated statistically.[6]

Figure 6.8 shows one isoquant, associated with the production function, corresponding to an output of 13,800 bushels of wheat per year. The manager of the farm can use this isoquant to decide whether it is profitable to hire more labor or use more machinery. Assume the farm is currently operating at A, with a labor input L of 500 hours and a capital input K of 100 machine-hours. The manager decides to experiment by using fewer hours of machine time. To produce the same crop per year, he finds that he needs to replace this machine time by adding 260 hours of labor.

The results of this experiment tell the manager about the shape of the wheat production isoquant. When comparing points A (where $L = 500$ and $K = 100$) and B (where $L = 760$ and $K = 90$) in Figure 6.8, both of which are on the same isoquant, the manager finds that the marginal rate of technical substitution is equal to 0.04 ($-\Delta K/\Delta L = -(-10)/260 = .04$).

The MRTS tells the manager the nature of the trade-off between adding labor and reducing the use of farm machinery. Because the MRTS is substantially less than 1 in value, the manager knows that when the wage of a laborer is equal to the cost of running a machine, he ought to use more capital. (At

[6] The food production function on which this example is based is given by the equation $Q = 100(K_8L_2)$, where Q is the rate of output in bushels of food per year, K is the quantity of machines in use per year, and L is the number of hours of labor per year.

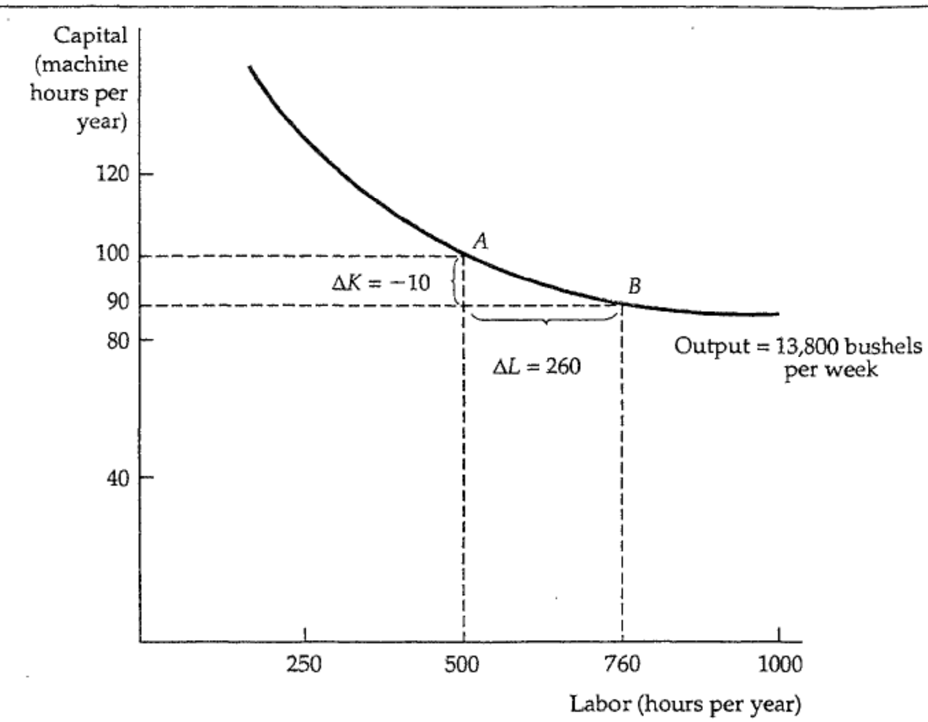

FIGURE 6.8 Isoquant Describing the Production of Wheat. A wheat output of 13/800 bushels per week can be produced with different combinations of labor and capital. The more capital-intensive production process is shown as point *A*, and the more labor-intensive process as *B*. The marginal rate of technical substitution between *A* and *B* is 10/260=0.04.

his current level of production, he needs 260 units of labor to substitute for 10 units of capital.) In fact, he knows that unless labor is substantially less expensive than the use of a machine, his production process ought to become more capital-intensive.

The decision about how many laborers to hire and machines to use cannot be fully resolved until we discuss the costs of production in the next chapter. However, this example illustrates how knowledge about production isoquants and the marginal rate of technical substitution can help a manager. It also suggests why most farms in the United States and Canada, where labor is relatively expensive, operate in the range of production in which the MRTS is relatively high (with a high capital-to-labor ratio), while farms in developing countries in which labor is cheap operate with a lower MRTS (and a lower capital-to-labor ratio).[7] The exact labor/capital combination to use depends on the input prices, a subject we discuss in Chapter 7.

6.5 Returns to Scale

The measure of increased output associated with increases in *all* inputs is fundamental to the long-run nature of the firm's production process. How does the output of the firm change as its inputs are proportionately increased? If output more than doubles when inputs are doubled, there are *increasing returns to scale.* This might arise because the larger scale of operation allows managers and workers to specialize in their tasks and make use of more sophisticated, large-scale factories and equipment. The automobile assembly line is a famous example of increasing returns.

The presence of increasing returns to scale is an important issue from a public policy perspective. If there are increasing returns, then it is economically advantageous to have one large firm producing (at relatively low cost) than to have many small firms (at relatively high cost). Because this large firm can control the price that it sets, it may need to be regulated. For example, increasing returns in the provision of electricity is one reason why we have large, regulated power companies.

A second possibility with respect to the scale of production is that output may double when inputs are doubled. In this case, we say there are *constant returns to scale.* With constant returns to scale, the size of the firm's operation does not affect the productivity of its factors. The average and marginal productivity of the firm's inputs remains constant whether the plant is small or large. With constant returns to scale, one plant using a particular production process can easily be replicated, so that two plants produce twice as much output. For example, a large travel agency might provide the same service per client and use the same ratio of capital (office space) and labor (travel agents) as a small travel agency that services fewer clients.

Finally, output may less than double when all inputs double. This case of *decreasing returns to scale is* likely to apply to any firm with large-scale operations. Eventually, difficulties of management associated with the complexities of organizing and running a large-scale operation may lead to decreased productivity of both labor and capital. Communication between workers and managers can become difficult to monitor and the workplace more impersonal. Thus, the decreasing-returns case is likely to be associated with the problems of coordinating tasks and maintaining a useful line of communication between management and workers. Or it may result because individuals cannot exhibit their entrepreneurial abilities in a large-scale operation.

[7] With the production function given in footnote 6, it is not difficult (using calculus) to show that the marginal rate of technical substitution is given by MRTS = (MP_L/MP_K) = ^(¼)(*K/L*). Thus, the MRTS decreases as the capital-to-labor ratio falls. For an interesting study of agricultural production in Israel, see Richard E. Just, David Zilberman, and Eithan Hochman, "Estimation of Multicrop Production Functions," *American Journal of Agricultural Economics* 65 (1983): 770-780.

The presence or absence of returns to scale is seen graphically in Figure 6.9. The production process is one in which labor and capital are used as inputs in the ratio of 5 hours of labor to 1 hour of machine time. The ray OB from the origin describes the various combinations of labor and capital that can be used to produce output when the input proportions are kept constant.

At relatively low output levels, the firm's production function exhibits increasing returns to scale, as shown in the range from 0 to *A*. When the input combination is 5 hours of labor and 1 hour of machine time, 10 units of output are produced (as shown in the lowest isoquant in the figure). When both inputs double, output triples from 10 to 30 units. Then when inputs increase by one-half again (from 10 to 15 hours of labor and 2 to 3 hours of machine time), output doubles from 30 to 60 units.

At higher output levels, the production function exhibits decreasing returns to scale, as shown in the range from *A* to *B*. When the input combination increases by one-third, from 15 to 20 hours of labor and from 3 to 4 machine hours, output increases only by one-sixth, from 60 to 70 units. And when inputs increase by one-half, from 20 to 30 hours of labor and from 4 to 6 machine hours, output increases by only one-seventh, from 70 to 80 units.

Figure 6.9 shows that with increasing returns to scale, isoquants become closer and closer to one another as inputs increase proportionally. However,

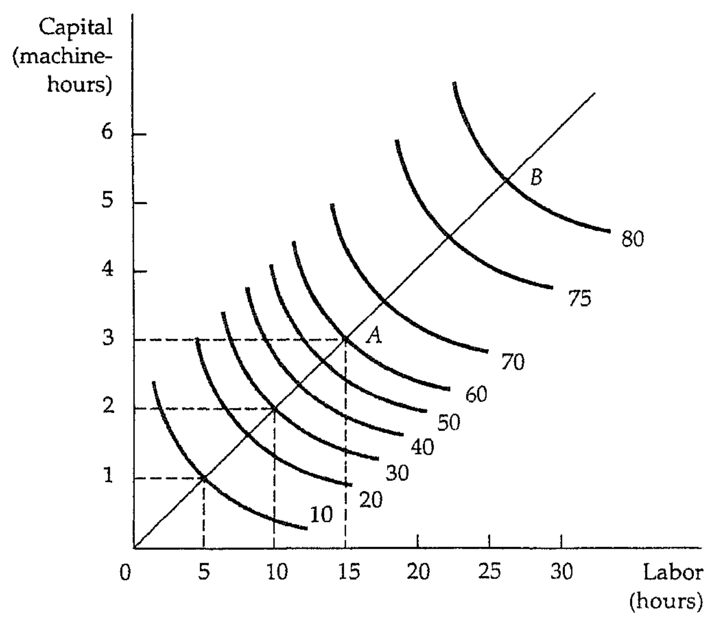

FIGURE 6.9 Returns to Scale. When a firn's production process exhibits increasing returns to scale as shown by a movement from 0 to *A* along ray *OB*, the isoquants get closer and closer to one another. However, when there are decreasing returns to scale as shown by a move from *A* to *B*, the isoquants get farther apart.

with decreasing returns to scale, isoquants become farther and farther from one another, because more and more inputs are needed. When there are constant returns to scale (not shown in Figure 6.9), isoquants are equally spaced.

Returns to scale vary considerably across firms and industries. Other things equal, the greater the returns to scale, the larger firms in an industry are likely to be. Manufacturing industries are more likely to have increasing returns to scale than service-oriented industries because manufacturing involves large investments in capital equipment. Services are more labor-intensive and can usually be provided as efficiently in small quantities as they can on a large scale.

EXAMPLE 6.4 RETURNS TO SCALE IN THE RAIL INDUSTRY

During most of this century, railroads have grown larger and larger, yet their financial problems have continued to mount.[8] Does this increase in size make good economic sense? If so, why do railroads continue to have difficulty competing with other forms of transportation? We can get some insight into these questions by looking at the economics of rail freight transportation.

To see whether there are increasing returns to scale, we will measure input as *freight density,* the number of tons of railroad freight that are run per unit of time along a particular route. Output is given by the amount of a particular commodity shipped along this route within the specified time. Then we can ask whether the amount that can be shipped increases more than proportionately as we add to freight tonnage. We might expect increasing returns initially because as more freight is shipped, the railroad management can use its planning and organization to design the appropriate scheduling of the freight system efficiently. However, decreasing returns will arise at some point when there are so many freight shipments that scheduling gets difficult and rail speeds are reduced.

Most studies of the railroad industry indicate increasing returns to scale at low and moderate freight densities, but decreasing returns to scale begin to set in after a certain point (called the *efficient density).* Only when the density gets quite large is this phenomenon important, however. One study, for example, indicated increasing returns to scale up to the range of 8 to 10 million tons (per year) per route-mile, a very large freight density.

To see the practical importance of these numbers, we have tabulated the 1980 freight densities of major U.S. railroads in Table 6.5. Some railroads such as Colorado & Southern and Union Pacific have reached or surpassed the point of minimum efficient size (the point at which increasing returns to scale disappear). But many railroads operate at freight densities below this.

[8]This example draws on the analysis of railroad freight regulation by Theodore Keeler, *Railroads Freight, and Public Policy (Washington D.C.:* The Brookings Institution, 1983), Chapter 3.

TABLE 6.5 Freight Densities of Major Railroads (Million Tons Per Route-Mile)

Railroad	Density
Atchison, Topeka & Santa Fe	6.03
Baltimore & Ohio	4.46
Burlington Northern	6.11
Chicago and Northwestern	3.10
Colorado & Southern	10.66
Fort Worth & Denver	6.55
Kansas City Southern	5.96
Missouri Pacific	5.01
Southern Pacific	5.35
Union Pacific	7.87
Western Pacific	3.20

Since most rail companies have not surpassed their optimum size, it appears that growth has been economically advantageous. The financial problems of the railroad industry relate more to competition from other forms of transportation than to the nature of the production process itself.

Summary

1. A *production function* describes the maximum output a firm can produce for each specified combination of inputs.
2. An *isocquant* is a curve that shows all combinations of inputs that yield a given level of output. A firm's production function can be represented by a series of isoquants associated with different levels of output.
3. In the short run, one or more inputs to the production process are fixed, whereas in the long run all inputs are potentially variable.
4. Production with one variable input, labo, can be usefully described in terms of the *average product of labor* (which measures the productivity of the average worker), and the *marginal product of labor* (which measures the productivity of the last worker added to the production process).
5. According to the "law of diminishing returns," when one or more inputs are fixed, a variable input (usually labor) is likely to have a marginal product that eventually diminishes as the level of input increases.

6. Isoquants always slope downward because the marginal product of all inputs is positive. The shape of each isoquant can be described by the marginal rate of technical substitution at each point on the isoquant. The *marginal rate of technical substitution of labor for capital* (MRTS) 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.
7. The standard of living that a country can attain for its citizens is closely related to its level of labor productivity. Recent decreases in the rate of productivity growth in developed countries are due in part to the lack of growth of capital investment.
8. The possibilities for substitution among inputs in the production process range from a production function in which inputs are perfectly substitutable to one in which the proportions of inputs to be used are fixed (a fixed-proportions production function).
9. In the long-run analysis, we tend to focus on the firm's choice of its scale or size of operation. Constant returns to scale means that doubling all inputs leads to doubling output. Increasing returns to scale occurs when output more than doubles when inputs are doubled, whereas decreasing returns to scale applies when output less than doubles.

Questions for Review

1. What is a production function? How does a long-run production function differ from a short-run production function?

2. Why is the marginal product of labor likely to increase and then decline in the short run?

3. Diminishing returns to a single factor of production and constant returns to scale are not inconsistent. Discuss.

4. You are an employer seeking to fill a vacant position on an assembly line. Are you more concerned with the average product of labor or the marginal product of labor for the last person hired? If you observe that your average product is just beginning to decline, should you hire any more workers? What does this situation imply about the marginal product of your last worker hired?

5. Faced with constantly changing conditions, why would a firm ever keep *any* factors fixed? What determines whether a factor is fixed or variable?

6. How does the curvature of an isoquant relate to the marginal rate of technical substitution?

7. Can a firm have a production function that exhibits increasing returns to scale, constant returns to scale, and decreasing returns to scale as output increases? Discuss.

8. Give an example of a production process in which the short run involves a day or a week, and the long run any period longer than a week.

Exercises

1. Suppose a chair manufacturer is producing in the short run when equipment is fixed. The manufacturer knows that as the number of laborers used in the production process increases from 1 to 7, the number of chairs produced changes as follows: 10, 17, 22, 25, 26, 25, 23.

a. Calculate the marginal and average product of labor for this production function.
b. Does this production function exhibit diminishing returns to labor? Explain.
c. Explain intuitively what might cause the marginal product of labor to become negative.

2. Fill in the gaps in the table below.

Quantity of Variable Input	Total Output	Marginal Product of Variable Input	Average Product of Variable Input
0	0	—	—
1	150		
2			200
3		200	
4	760		
5		150	
6 .			150

3. A political campaign manager has to decide whether to emphasize television advertisements or letters to potential voters in a reelection campaign. Describe the production function for campaign votes. How might information about this function (such as the shape of the isoquants) help the campaign manager to plan strategy?

4. A firm has a production process in which the inputs to production are perfectly substitutable in the long run. Can you tell whether the marginal rate of technical substitution is high or low, or is further information necessary? Discuss.

5. The marginal product of labor is known to be greater than the average product of labor at a given level of employment. Is the average product increasing **or** decreasing? Explain.

6. The marginal product of labor in the production of computer chips is 50 chips per hour. The marginal rate of technical substitution of hours of labor for hours of machine-capital is ¼. What is the marginal product of capital?

7. Do each of the following production functions exhibit decreasing, constant, or increasing returns to scale.
a. $Q = .5KL$
b. $Q = 2K + 3L$

8. The production function for the personal computers of DISK, Inc., is given by $Q = 10K_5L_5$, where Q is the number of computers produced per day, K is hours of machine time, and L is hours of labor input. DISK'S competitor, FLOPPY, Inc., is using the production function $Q = 10K_6L_4$.
a. If both companies use the same amounts of capital and labor, which will generate more output?
b. Assume that capital is limited to 9 machine hours/but labor is unlimited in supply. In which company is the marginal product of labor greater? Explain.

9. In Example 6.3, wheat is produced according to the production function $Q = 100(K_8L_2)$.
a. Beginning with a capital input of 4 and a labor input of 49, show that the marginal product of labor and the marginal product of capital are both decreasing.
b. Does this production function exhibit increasing, decreasing, or constant returns to scale?