4

Technology and Cost

Anyone who has observed the extensive growth of electronic commerce in recent years cannot fail to have noticed the low prices charged by many Internet firms. Egreetings Network, Inc., an Internet firm selling e-mail greeting cards, offers a case in point. In just one year, 1998, the firm lowered its price per card from $2.50 to just $0.50. A year later the company reduced its fees still further. Currently the firm charges less than 10 cents for some cards and gives others away for free. Such minimal pricing strategies are not uncommon in digital commerce. Many e-sellers permit customers to download their products either freely or for a very modest charge. The question that naturally arises is how such behavior can be profitable? Surely, these firms incur costs in producing their goods and services. How can they cover such costs while selling at such low prices?

Production costs are an important factor explaining firm behavior, as well as an important determinant of the industry's structure. The four firms—*General Mills*, *Kelloggs*, *General Foods* (*Post*), and *Quaker Oats* currently account for about 80 percent of sales in the U.S. ready-to-eat breakfast cereal industry. By contrast, the largest four manufacturers of games and toys account for 35 percent to 45 percent of these products—less if video games are included. In this chapter we introduce key cost concepts that are relevant to understanding industry structure.[1]

4.1 PRODUCTION TECHNOLOGY AND COST FUNCTIONS FOR SINGLE PRODUCT FIRMS

What is a firm's technology? For our purposes, the firm's technology is a production relationship that describes how a given quantity of inputs is transformed into the firm's output. In this sense we adopt the traditional neoclassical approach to the firm in which a firm is solely envisioned as a production unit. The goal of this production unit is profit maximization, which, in turn, implies minimizing the cost of making any given level of output.

The neoclassical approach is not without its weaknesses. While it does indicate how the firm's production plans changes in response to changes in input and output prices, it says little about how that plan is actually implemented or managed. In other words, it says little about what happens inside the firm and, more specifically, about how the various competing

[1] Panzar (1989) presents a more extended review of this topic.

interests of management, workers, and shareholders are reconciled in the design and implementation of a production plan.[2]

Moreover, whatever happens within a firm it is clear that these internal relationships are different from the external ones that exist between the firm and those outside the firm such as customers and suppliers. A market typically mediates these external relationships. Customers and suppliers buy from and sell to the firm at market prices. Inside the firm, however, relationships are organized by non-market methods, such as hierarchical control. Thus, as eloquently argued by Nobel laureate Ronald Coase (1937), the boundary of the firm is really the boundary between the use of non-market business transactions and market ones. The question Coase then raised is what determines this boundary. Why is it that production of a good is distributed across many different firms instead of a few large ones? Indeed, what limits are there to having all production organized by one or a few giant, multidivisional and multiplant firms?

These are questions that the neoclassical view of the firm cannot fully answer because its focus on production costs narrowly defined leads it to ignore another important cost—the cost of transacting business. Coase (1937) was the first to raise the issue in his classic paper, "The Nature of the Firm," Hart and Moore (1990), Williamson (1975), and Hart (1990) are subsequent important contributions to these issues, as is Bolton and Scharfstein (1998). Yet while the neoclassical approach to firm size and market structure is not without its limitations, the approach does remain insightful. For our purposes, it is useful to be aware of the issues raised by the agency and transactions cost literature but to explore those concerns at all satisfactorily would take us beyond the boundary of this book! As long as its limitations are recognized, the neoclassical view of the firm will permit us to accomplish many of our objectives. So keep in mind throughout the following discussion that a firm is interpreted as simply a profit-maximizing production unit and not a complex organization.

4.1.1 Key Cost Concepts

Standard microeconomic theory describes a firm in terms of its production technology. A firm producing the quantity q of a single product is characterized by its production function $q = f(x_1, x_2, \ldots, x_k)$. This function specifies the quantity q that the firm produces from using k different inputs at levels x_1 for the first input, x_2 for the second input, and so on through the kth input of which x_k is used. The technology is reflected in the precise form of the function, $f(\ )$. In turn, the nature of this technology will be a central determinant of the firm's costs.

The firm is treated as a single decision-making unit that chooses output q and the associated inputs $x_1, x_2, \ldots, x_k$ to maximize profits. It is convenient to approach this choice by first identifying the relationship between a firm's output and its resulting production costs—which is simply the firm's cost function. That is, for any specific output $\bar{q}$ and given the prices $w_1, w_2, \ldots, w_k$ of the k inputs, there is a unique way to choose the level of each input $x_1, x_2, \ldots, x_k$ so as to minimize the total cost of producing $\bar{q}$. The firm obtains this solution by choosing that input combination that solves the problem:

$$\underset{x_i}{Minimize} \sum_{i=1}^{k} w_i x_i \tag{4.1}$$

subject to the constraint $f(x_1, x_2, \ldots, x_k) = \bar{q}$.

[2] See Milgrom and Roberts (1992) for a classic discussion of these issues.

If we solve this problem for different levels of output $\bar{q}$, we will obtain the minimum cost of each possible production level per unit of time. This relationship between costs and output is what is described by the cost function for the firm. We typically describe the firm's cost function by the expression $C(q) + F$, from which we can then derive three key cost concepts: fixed cost; average or unit cost; and marginal cost.

1. *Fixed cost*: The fixed cost concept is reflected in the term F. This term describes a given amount of expenditure that the firm must incur each period and that is unrelated to how much output the firm produces. That is, the firm must incur F whether it produces 0 or a 1,000 units, hence the term, fixed. This is distinct from the variable cost portion described by $C(q)$ that does vary as output changes. Costs that may be fixed include interest costs associated with financing a particular size of plant and advertising costs. Note, however, that often these costs may be fixed only in the short run. Over a longer period of time, the firm can adjust what plant size it wants to operate and its promotional efforts. If this is true, then these costs are not fixed over a longer period of time.
2. *Average cost*: The firm's average cost is simply a measure of the expenditure per unit of production and is given by total cost divided by total output. This cost measure does depend on output; hence its algebraic representation is $AC(q)$. Formally, $AC(q) = [C(q) + F]/q$. We may also decompose average cost into its fixed and variable components. Average fixed cost is simply total fixed cost per unit of output or F/q. Average variable cost $AVC(q)$ is similarly just the total variable cost per unit of output, $C(q)/q$. Alternatively, average variable cost is just average cost less average fixed cost, $AVC(q) = AC(q) - F/q$.
3. *Marginal cost*: The firm's marginal cost $MC(q)$ is calculated as the addition to total cost that is incurred in increasing output by one unit. Alternatively, marginal cost can be defined as the savings in total cost that is realized as the firm decreases output by one unit. More precisely, marginal cost is the slope of the total cost function and so is defined by the derivative term, $MC(q) = dC(q)/dq$.

We now add a fourth key cost concept—*sunk cost*. Like fixed cost, sunk cost is a cost that is unrelated to output. However, unlike fixed costs, which are incurred every period, sunk cost is a cost that is incurred just once—typically as a prerequisite for entry. For example, a doctor will need to acquire a license to operate. Similarly, a firm may need to do market and product research or install highly specialized equipment before it enters a market. The cost of the license, the research expenditures and the expenditures on specialized assets are likely to be unrelated to subsequent output, so in this sense they are fixed. More importantly, should the doctor or firm subsequently decide to close down, only part of these specialized expenditures will be recoverable? It might be possible to sell the license to another doctor but probably not at the price that the first doctor paid. Similarly, the research expenditures are unrecoverable on exit and it will not be possible to sell the specialized assets for anything like their initial acquisition costs. For example, the kilns that are needed to manufacture cement have almost no alternative use other than as scrap metal. Much of the capital cost that Toyota incurred in building its U.S. car manufacturing plants—production lines, robots, and other highly specialized machinery—have no other uses. By contrast, the airplanes used by JetBlue to open up a new route, say between Boston and Miami, can be redeployed if passenger traffic on that route turns out to be insufficient to continue its operation. Sunk costs, in other words, are initial entry costs that are unrecoverable if the doctor or firm chooses to exit the market.

4.1.2 Cost Variables and Output Decisions

Figure 4.1 depicts a standard textbook average cost function, $AC(q)$, and its corresponding marginal cost function, $MC(q)$. As discussed in Chapter 2, profit maximization over any period of time requires that the firm produce where marginal revenue is equal to marginal cost. Thus, with one important caveat, marginal cost is the relevant cost concept to determine how much the firm should produce. That caveat is that marginal cost is important for determining how much to produce *given* that the firm is going to produce any output at all.

Suppose for example that demand is very weak. In such a case, equating marginal cost to marginal revenue may result in price falling below average cost. If price is below average cost, the firm loses money on every unit that it sells. It cannot continue to do this in the long run. Hence, the firm will eventually shut down if price stays below average cost. Whether this shutdown happens sooner or later will depend on the relation between price and average variable cost, $AVC(q)$. If price exceeds average variable cost, the firm will continue to operate in the short run. If price is above average variable cost, the firm can make some operating profit on each unit that it sells and this provides funds to cover at least some of its fixed cost. However, if price is below average variable cost, then the firm will simply shut down immediately.

Consideration of price and average cost also allows us to identify the role played by sunk cost in the firm's decision-making. Again, profit per unit in any period is simply price less average cost, $P - AC(q)$. Total profit in any period is just the profit per unit times the number of units, $[P - AC(q)]q$. Before entering an industry, a firm must expect at least to break even. If entry incurs a sunk cost such as a licensing fee or research expense, then the firm will have to believe that it will earn enough profit in subsequent periods to cover that initial sunk cost. Otherwise, it will not enter the market. Formally, the discounted present value of the expected future profits must be at least as great as the sunk cost of entry. Note though, that once it has entered, the sunk cost is no longer relevant. Once the entry decision has been made and the sunk cost incurred, the best that the firm can do is to follow the prescription above: produce where marginal revenue equals marginal cost so long as in the short run price is greater than average variable cost, otherwise shut down. In the long run: produce where marginal revenue equals marginal cost so long as price is greater than average cost, otherwise exit. Sunk cost affects the entry decision—not the decision on how much to produce after entry has occurred nor the decision to exit.

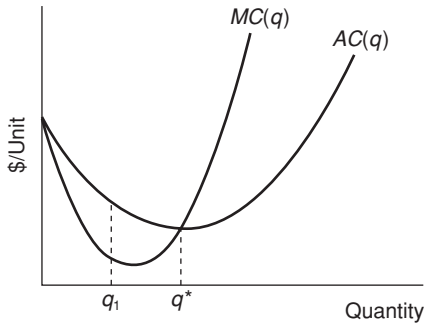

Figure 4.1 Typical average and marginal cost curves

In sum, the concept of average cost is relevant to whether the firm will produce positive output in the long run, and the concept of average variable cost is relevant to whether the firm will produce positive output in the short run. The concept of marginal cost is relevant to how much output the firm will produce given that it chooses to produce a positive amount. Sunk cost is relevant to the decision to enter the market in the first place.

4.1.3 Costs and Market Structure

Let's take a second, closer look at Figure 4.1. This figure illustrates an important relationship between average and marginal costs. Note that when marginal cost is less than average cost, as at output q_1, an expansion of output will lead to a reduction in average cost. Conversely, when marginal cost is greater than average cost, an expansion of output will lead to an increase in average cost. In the figure, marginal cost is less than average cost for all outputs less than q^*, and average cost falls throughout this range of output. Marginal cost is greater than average cost for outputs greater than q^* and average cost rises over this range of output. This feature is true for all cost functions. Average cost falls whenever marginal cost is less than average cost and rises whenever marginal cost exceeds average cost. A corollary of the just-described relationship between marginal cost and average cost is that the two are equal at the minimum point on the average cost function.

The basic cost relationships are illustrated with a hypothetical example in Table 4.1 (the parameter S in this table is explained below). This table provides measures of total, average and marginal cost data for an imaginary firm.[3] As that table documents, average cost falls when it lies above marginal cost; rises when it is below marginal cost and (because the numbers are an approximation) is essentially equal to marginal cost at the minimum average cost value. Intuitively, if marginal cost is below average cost when average cost is falling but crosses above average cost when average cost is rising, then the crossing point at which the two are equal must be at the minimum average cost.

As noted above, firms have to expect to break even in order for production to be profitable. This means that both average cost and sunk cost play a role in determining market structure. We consider average cost first.

Table 4.1 Average and marginal cost

Output	*Total cost ($)*	*Average cost ($/output)*	*Marginal cost (Δ$/Δ output)*	*Scale economy index (S)*
5	725	145	—	—
6	816	136	96	1.42
7	917	131	104	1.26
8	1,024	128	113	1.13
9	1,143	127	123	1.03
10	1,270	127	132	0.96
11	1,408	128	151	0.85
12	1,572	131	—	—

[3] Note: in Table 4.1, marginal cost is calculated as the average of the increase in cost associated with producing 1 unit more and the decrease in cost associated with producing 1 unit less.

Derivation Checkpoint

Average Cost, Marginal Cost, and Cost Minimization

Average cost is defined to be $AC(q) = C(q)/q$. Differentiate this with respect to output to yield:

$$\frac{dAC(q)}{dq} = \frac{q\dfrac{dC(q)}{dq} - C(q)}{q^2}$$

This can be simplified to:

$$\frac{dAC(q)}{dq} = \frac{q\left(MC(q) - \dfrac{C(q)}{q}\right)}{q^2} = \frac{[MC(q) = AC(q)]}{q}$$

The denominator of this term is positive. So, the slope of the average cost curve depends on the relation between marginal cost and average cost. If marginal cost exceeds average cost, the slope is positive. Raising output raises average cost. If average cost exceeds marginal cost, the slope is negative. Raising output lowers average cost. Minimum average cost is found where the slope of the average cost curve is zero. It is easy to see from the equation above that this occurs when average cost and marginal cost are equal.

Derivation of total and average cost functions assumes that firms produce each output level at minimum cost. A necessary condition for such minimization is that the following equation be satisfied for any pair of inputs i and j:

$$\frac{MP_i}{MP_j} = \frac{w_i}{w_j}; \text{ which is equivalent to } \frac{MP_i}{w_i} = \frac{MP_j}{w_j}$$

In other words, inputs should be used up to the point where the marginal product of the last dollar spent on input i equals the marginal product of the last dollar spent on input j.

The fact that average cost falls as output increases amounts to saying that the cost per unit of output declines as the scale of operations rises. It is natural to describe this state of affairs as one in which there are economies of scale. If, however, unit costs rise as production increases we say that there are diseconomies of scale. Fundamentally, the presence of scale economies or scale diseconomies reflects the underlying technology. Some factors of production simply cannot be scaled down to small levels of production. For example, provision of passenger rail service between Omaha and Lincoln, Nebraska, will require approximately 60 miles of track whether the number of trains per day is 1 or 20. As a result, a passenger train firm renting the track from the freight company that currently owns it, will have to pay the same rent whether it has many passengers or just a few.

Yet it is not just the presence of large fixed costs that give rise to scale economies. For many productive processes, there are efficiencies that come about just as a result of being larger. To begin with, size permits a greater division of labor, as Adam Smith noted over

two hundred years ago[4]. This in turn permits specialization and more efficient production. Sometimes, the simple mathematics of the activity gives rise to important scale effects. It is well known, for example, that the cost of a container will rise roughly in proportion to its surface area (essentially, the radius squared), whereas its capacity rises roughly in proportion to its volume (essentially, the radius cubed). Thus, while a $10 \times 10 \times 10$ cube will hold 1,000 cubic feet, a $20 \times 20 \times 20$ cube holds 8,000 cubic feet. Since the cost in terms of materials and labor depends on surface area but output depends on volume, it follows that as container size increases there is a less-than-proportional rise in the cost. In turn, this implies that unit cost declines as output increases. Specifically, unit cost will fall by about 3 percent for every 10 percent increase in output.[5] For a variety of processes, such as distributing natural gas via a pipeline or manufacturing glass products in which molten glass is kept in large ovens, this relationship suggests that it will be less expensive per unit to operate at a large volume.[6]

Whatever the source of the scale economies, the fact that scale economies are measured by a falling average cost gives us a precise way to measure their presence. For we know that a declining average cost can only be observed if marginal cost is below average cost. Likewise, the presence of scale diseconomies or rising average cost requires that marginal cost be above average cost. Hence, we can construct a precise index of the extent of scale economies by defining the measure S to be the ratio $AC(q)/MC(q)$. That is, S is the ratio of average to marginal cost. S can also be shown (see the inset) to be the inverse of the elasticity of cost with respect to output. In other words, S measures the proportionate increase in output one obtains for a given proportionate increase in costs.

The more that S exceeds 1, the greater is the extent of scale economies. In such a setting, a one percent increase in output is associated with a less than one percent increase in costs. Conversely, when $S < 1$, diseconomies of scale are present. Increasing output by one percent now leads to more than a one percent increase in costs. Finally, when $S = 1$, neither economies nor diseconomies of scale are present. In this case, we say that the production technology exhibits constant returns to scale.

We define *minimum efficient scale* as the lowest level of output at which economies of scale are exhausted or, in other words, at which $S = 1$. In Figure 4.1 minimum efficient scale is q^*.

In Table 4.1, we can approximate the value of S at $q = 6$ as follows. The addition to total cost of increasing output from 6 to 7 is \$101. The reduction in total cost of *decreasing* q by one unit is \$91. So, an approximate measure of marginal cost at exactly $q = 6$ is the mean of these two numbers or \$96. Average cost at $q = 6$ is \$136. Accordingly, $S = 136/96 = 1.42$. S can also be estimated by dividing the percentage increase in total output by the percentage increase in total cost. For example, when output is increased from 6 to 7 the percentage increase is given by:

$$\frac{1}{6} \times 100\% = 16.67\%$$

[4] Adam Smith's classic, *The Wealth of Nations*, includes a famous chapter on the division of labor and the productivity enhancement that this yielded at a pin factory.

[5] The classic study by Chenery (1947) on natural-gas pipelines is an example of this technical relationship.

[6] The technical explanations given here reflect the shortcomings of the neoclassical approach in that they do not make clear why the scale economies associated with a specific production technology must be exploited within a single firm. For example, two or more firms can own pipelines jointly. Indeed, there is growing support for the use of co-ownership or cotenancy, as an alternative to direct regulation in the case of natural monopoly. See Gale (1994).

Derivation Checkpoint

The Scale Economy Index and the Elasticity of Total Cost

The standard definition of the elasticity η_C of costs with respect to output is the proportionate increase in total cost that results from a given proportionate increase in output. This can be written as:

$$\eta_C = \frac{dC(q)}{C(q)} \Big/ \frac{dq}{q} = \frac{dC(q)}{dq} \Big/ \frac{q}{C(q)} = \frac{MC(q)}{AC(q)}.$$

As the scale economy index S is defined as the ratio of average cost to marginal cost, it follows that: $S = 1/\eta_C$.

Meanwhile, this output rise induces a percentage increase in total cost of:

$$\frac{917 - 816}{816} \times 100\% = 12.37\%$$

The ratio of these two percentages is then 16.67 percent/12.37 percent = 1.35 percent. This is not far from the measure of S (= 1.42) that we obtained using the ratio of average to marginal cost. Indeed, if we could vary production more continuously and so consider the cost of producing 6.5 units, or 6.25 units and so on, the two measures would be virtually equal.

The ratio of the percentage change in total cost with respect to the percentage change in output is called the elasticity of cost with respect to output. What we have just shown is that the inverse of this ratio—the percentage change in output divided by the percentage change in cost—is a good indicator of scale economies. In other words, the inverse of the elasticity of cost with respect to output is a very good measure of S.

Practice Problem

4.1 Confirm that at an output of $q = 11$, the scale economy index in Table 4.1 is indeed 0.85.

How is the behavior of average cost or the extent of scale economies related to industry structure? Going back to Figure 4.1, we see that $S > 1$ for any level of output less than q^*. Scale economies are present at every output level in this range. By contrast, $S < 1$ for all outputs greater than q^*. Now suppose that we have other information indicating that demand conditions are such that the maximum extent of the market is less than q^* even if price falls to zero. We can then state that scale economies are present throughout the relevant range of production. Put another way, economies of scale are global in such a market.

If scale economies are global then the market is a natural monopoly. The term "natural" is meant to reflect the implication that monopoly is an (almost) inevitable outcome for this market because it is cheaper in such cases for a single firm to supply the entire market than

for two or more firms to do so. For example, the least expensive way to produce the quantity q^* in Figure 4.1 is to have one firm produce the entire amount. If instead, two firms divided this production equally, so that each produced an output $q_1 = q^*/2$, each of these two firms would have higher average costs than would the single firm producing q^*.

The role of scale economies in determining market structure should now be clear. If scale economies are global, there will be no more than one firm in the market. Even if they are not global but simply quite large, efficiency may still require that all the production be done in one firm. More generally, the greater is the extent of scale economies—the larger the output at which average cost is minimized—the fewer firms that can operate efficiently in the market. Thus, large-scale economies will tend to result in concentrated markets.

Reality Checkpoint

Hotel Phone Costs May Be Fixed

Business travelers stopping at the Hampton Inn in Salt Lake City often find themselves powering down their cell phones and just relying on their room phone even though this is far more expensive. At some hotels, the cost of using the in-room phone can run as high $2 per minute or even more for domestic calls and ten or more times that amount for international calls. This compares with a near zero charge for cell phone calls. So, why does anyone use a hotel room phone?

The main answer is that cell phones do not always work. Reception can be poor and getting cell phone service simply may not be possible. For many travelers, the need to be in phone contact with others is such that they are willing to pay the high prices of hotel room phones. In turn, those high prices are necessary in part because of the fixed costs the hotels incur whether the phone is used a lot, a little, or not at all. These costs include a fixed rental fee for each line, the expense of employing operators, and the cost of maintaining equipment all of which is incurred regardless of the intensity with which room phones are used. The hotels charge a hefty fee, well above marginal cost, to earn those fixed costs back.

Unfortunately for the hotels, the advent of cell phones has sharply cut into their room phone revenue. In fact, operating profits per room phone per year in the U.S. fell from $644 in 2000 to $152 in 2004. This loss in revenue and profit may have led some hotels to go to rather unusual lengths to beat the cell phone competition. In 2003, the Scottish newspaper, the *Daily Record*, reported evidence that a local firm, Electron Electrical Engineering Services, was selling cell phone jamming devices to hotels and bed-and-breakfast establishments for between $135 and $200 apiece. These devices have the ability to block cell phone reception without the cell phone customer realizing it. All the customer will see is a message that "service is unavailable" in the location from which they are calling. Loreen Haim-Cayzer, the director of marketing and sales for Netline Communications Technologies in Tel Aviv, also acknowledged that her company had sold hundreds of cell phone jammers to hotels around the world, though none in the U.S. as far as she knew.

Of course, savvy phone users have another option. They can carry a phone card for use whenever their cell phone cannot get a signal. Those who do not, however, will have to rely on the in-room phone . . . and pay the associated fees. These customers may perhaps be forgiven then if they suspect that it is more than the costs of such phones that's fixed.

Source: C. Elliot, "Mystery of the Cell Phone that Doesn't Work at the Hotel," *New York Times*, September 7, 2004, p. C8; and C. Page "Mobile Phones Jam Scam," *The Daily Record*, August 26, 2003, p. 1.

Practice Problem

4.2 Consider the following cost relationship: $C = 50 + 2q + 0.5q^2$.

a. Derive an expression for average cost. Plot the value of average cost for $q = 4$, $q = 8$, $q = 10$, $q = 12$, and $q = 15$.
b. Marginal cost can be approximated by the rise in cost, ΔC, that occurs when output increases by one unit, $\Delta q = 1$. However, it can also be approximated by the fall in cost that occurs when output is decreased by one unit, $\Delta q = -1$. Since these two measures will not be quite the same we often use their average. Show that for the above cost relation, this procedure produces an estimate of marginal cost equal to $MC = 2 + q$.
c. Compute the index of scale economies, S. For what values of q is it the case that $S > 1$, $S = 1$, and $S < 1$?

4.2 SUNK COST AND MARKET STRUCTURE

Sunk costs also play a role in influencing market structure and one that is conceptually similar to the role of scale economies. Again, firms only enter a market if they believe that they can at least break even. This means that if there are positive sunk costs associated with entry, then firms must earn positive profits in each subsequent period of actual operation to cover those entry costs. If this is the case, entry will occur. Indeed, this view leads naturally to a definition of long run equilibrium. Firms will stop entering the industry—and therefore the number of firms will be at its equilibrium level—when the profit from operating each period just covers the initial sunk cost that entry requires. Of course, the more firms that do enter an industry, the more competitive its pricing will be and the less profit a firm will make in any period of actual operation.

The foregoing logic permits us to see clearly the role of sunk cost in determining market structure. The higher the sunk cost, the fewer firms there will be in equilibrium. A high sunk entry cost requires that each firm that enters subsequently earns a fair bit of profit from its operations to repay the initial entry expense. This can only happen if the number of firms that enter is small so that competition is weak and price can rise above marginal (and average) cost.

To take a fairly simple example, imagine a market in which each firm produces an identical good and in which the elasticity of demand is exactly one, or $\eta = 1$, throughout the demand curve. This means that the total consumer expenditure for the product is constant. A 1 percent decrease in price is balanced by a 1 percent increase in quantity sold. Denote this constant total expenditure as E. If P is the market price and Q is total market output, we then have: $E = PQ$. However, total output Q is also equal to the output of each firm q_i times the number of firms, N, i.e., $Q = Nq_i$. Putting these two relationships together we then obtain:

$$q_i = E/NP \tag{4.2}$$

Now recall the Lerner Index that discussed in Chapter 3. If we assume that all firms are identical and that each has a constant marginal (and average) production cost c, then this index LI is given by: $(P - c)/P$. Since this index is a measure of the extent of monopoly

power in the industry, it is natural to assume that it declines as the number of firms N gets larger. We formalize this idea by assuming that the industry Lerner Index is negatively related to the number of firms N as follows:

$$(P - c)/P = A/N^{\alpha} \tag{4.3}$$

where A and α are both arbitrary positive constants. Finally, let's assume that firms only operate one period so that to break even requires that: $(P - c)q_i = F$, where F is the sunk entry cost.[7] Substituting this break-even requirement into equation (4.2) and combining that equation with equation (4.3) then yields that the equilibrium number of firms N^e at which each entrant just covers its sunk entry cost F, is given by:

$$N^e = \left[\frac{AE}{F}\right]^{\frac{1}{1+\alpha}} \tag{4.4}$$

The intuition underlying equation (4.4) is straightforward. Industry structure is likely to be more concentrated in markets where sunk entry costs are a high proportion of expected consumer expenditures

4.3 COSTS AND MULTIPRODUCT FIRMS

Since scale economies are a description of the behavior of costs as output increases, investigating their existence in any industry requires that we measure that output of the firms in that industry. This is not always so easy. Consider, for instance, the case of a railroad. One possible measure of output is the rail ton-mile, defined as the number of tons transported times the average number of miles each ton travels. However, not all railroads carry the same type of freight. Some carry mainly mining and forestry products, some carry manufactured goods, and some carry agricultural products. In addition, through the first half of this century, many private U.S. railroads carried passengers as well as freight. Elsewhere in the world, this is still the case. Since all of these different kinds of services have different carrying costs, aggregating each railroad's output into a simple measure such as total ton-miles will confuse any cost analysis. Such aggregation does not allow us to identify whether cost differences between railroads are due to differences in scale or to differences in the kinds of service being offered.

The railroad example points to a gap in our analysis of the firm. In particular, it implies the need to extend the analysis to cover firms producing more than one type of good, that is, to investigate costs for multiproduct firms. This need is perhaps more important today than ever before. Evidence provided by Dunne, Roberts, and Samuelson (1988) and others indicates that the great majority of business establishments produce more than one product—often many more. The major automobile firms also produce trucks and buses. *Microsoft* produces both the Windows operating systems and several applications written for that system. Consumer electronics firms produce TV's, stereos, CD players and so on. Measuring the output of these firms is clearly less than straightforward.

[7] Alternatively, we could assume that F is the annualized value of the sunk entry costs.

Even when firms produce what might be considered a single basic product, they typically offer several varieties of that good. In the ready-to-eat breakfast cereal industry, the top four firms market over eighty brands of cereal. If we are to use the technological approach to the firm to gain some understanding of industry structure, we clearly need to extend that approach to handle multiproduct companies. In other words, we need to develop an analysis of costs for the multiproduct firm. The question then becomes whether we can derive average cost and scale economy measures for multiproduct firms that are as precise and clear as the analogous concepts developed for the single product case.

Derivation Checkpoint

Ray Average Cost and Multiproduct Scale Economies

Scale economies are always indicated by declining average cost. The relevant concept of average cost for a multiproduct firm is ray average cost (*RAC*). If a firm has 2 products so that its cost function is $C(q_1, q_2)$ we may implicitly define total output q by the equations $q_1 = \lambda_1 q$, and $q_2 = \lambda_2 q$, where λ_1 and λ_2 sum to unity. Then ray average cost is:

$$RAC(q) = \frac{C(\lambda_1 q, \lambda_2 q)}{q}$$

In the single product case, the scale economy measure reflects the behavior of average cost as output expands. Similarly, for the two-product case, the issue is the behavior of *RAC* as output expands. Formally, this is given by derivative of *RAC* with respect to q. This is:

$$\frac{dRAC(q)}{dq} = \frac{(\lambda_1 MC_1 + \lambda_2 MC_2)q - C(\lambda_1 q, \lambda_2 q)}{q^2} = \frac{q_1 MC_1 + q_2 MC_2 - C(q_1, q_2)}{q^2}$$

where MC_i is the marginal cost of producing good i. It follows immediately that the sign of $dRAC(q)/dq$ is determined by the sign of the numerator of this expression. In other words, if $q_1MC_1 + q_2MC_2 > C(q_1, q_2)$ then $dRAC(q)/dq > 0$, while if $q_1MC_1 + q_2MC_2 < C(q_1, q_2)$ then $dRAC(q)/dq < 0$. Now define the ratio

$$S = \frac{C(q_1, q_2)}{q_1 MC_1 + q_2 MC_2}$$

The sign of the derivative above is then fully described by the value of S. If $S > 1$, this is equivalent to saying that ray average cost decreases with output and so exhibits multiproduct increasing returns to scale. If $S < 1$, ray average cost is increasing, and so exhibits multiproduct decreasing returns to scale. If $S = 1$, neither scale economies nor diseconomies exist for the multiproduct firm. Note the similarity of this measure with our single product scale economy index. In the single product case, we measured scale economies by the ratio of average to marginal cost. This is more or less what we are doing here except that average cost is now measured by total cost divided by a weighted average of marginal cost. This is why we continue to use S to indicate scale economies. Moreover, while we have worked out this case for just the two-product firm, it easily generalizes to the case in which there are more than two products.

The answer to the foregoing question is that, subject to some restrictions, yes, we can. This is one of the major contributions of Baumol, Panzar, and Willig (1982). These authors show that the restriction is simply that we measure average cost for a given mix of products, say two units of freight service for every one unit of passenger service in the railroad case. We can then measure average cost at any production level so long as we keep these proportions constant. This is what the Baumol, Panzar, and Willig (1982) call Ray Average Cost (RAC). They further show that we can derive a measure of scale economies based on the RAC measure that is conceptually quite similar to the scale economies measure for the single-product case. (See the Derivation Checkpoint: Ray Average Cost and Multiproduct Scale Economics.)

Perhaps the most important insight of Baumol, Panzar, and Willig (1982), however, is their introduction of the concept of economies of ***scope***. Economies of scope are said to be present whenever it is less costly to produce a set of goods in one firm than it is to produce that set in two or more firms. Let the total cost of producing two goods, q_1 and q_2, be given by $C(q_1, q_2)$. For the two-product case, scope economies exist if $C(q_1, 0) + C(0, q_2) - C(q_1, q_2) > 0$. The first two terms in this equation are the total costs of producing product 1, passenger services for example, in one firm and product 2, say freight services, in another. The third term is the total cost of having these products produced by the same firm. If this difference is positive, then scope economies exist. If it is negative, there are diseconomies of scope. If it is 0, then there are neither economies nor diseconomies of scope. The degree of such economies, S_C, is defined by the ratio:

$$S_C = \frac{C(q_1,0) + C(0,q_2) - C(q_1,q_2)}{C(q_1,q_2)} \tag{4.5}$$

The concept of scope economies is a crucial one that provides the central technological reason for the existence of multiproduct firms. Perhaps what is most important about scope economies, however, is that they give rise to multiproduct scale economies where we might not have expected any to exist. Looking at the production of only one product may not indicate any scale economy effects. However, if producing more of one product lowers the cost of producing another, then the firm may be able to lower its Ray Average Cost as it increases the production of both products.

Economies of scope can arise for two main reasons. The first of these is that particular outputs share common inputs. This is the source of economies of scope in the railroad example. There, the common factor is the track necessary to offer either passenger or freight rail service. Many other examples can be identified. For instance, a firm's advertising expenditures benefit all of its products to the extent that such advertising is intended to establish the firm's brand name. Similarly, if different products are manufactured with identical components —computer chips, for example—the manufacture of a whole range of such products allows the firm to take advantage of economies of scale in the manufacture of the components.

An alternative source of scope economies is the presence of cost complementarities. Cost complementarities occur when producing more of one good lowers the cost of producing a second good. There are numerous ways in which such interactions can take place. For example, the exploration and drilling of an oil well often yields not just oil but also natural gas. Hence, engaging in crude oil production will likely lower the cost of gas exploration. Similarly, a firm that manufactures computer software may also find it easy to provide computer consulting services.

Reality Checkpoint

Talk About Scope Economies, Holy Cow!

Economies of scope arise in many situations —including agricultural production. There is ample evidence that firms producing multiple crops and are more cost efficient than firms specializing in just one or two crops. For firms that specialize in particular livestock, the gains from adding other livestock or crops seem to be less clear. However, this may be because livestock production itself already embodies many scope economies even when totally specialized.

Consider a cattle ranch. Raising cattle not only produces beef, but also leather. So, it is clearly cheaper for one farm to produce both products rather than for two farms to do each separately. Yet the scope economies of cattle production do not stop there. Cattle carcasses are actually used in hundreds of processes. Glycerin and collagen are both cattle by-products. Other cattle body parts find their way into vaccines, animal feed, lubricants, asphalt, paper coatings, and fabric softeners. Imagine how much additional cost would be incurred if separate cattle stocks were maintained for the production of each of these goods.

Sources: V. Klinkenborg, "The Whole Cow and Nothing but the Whole Cow". *The New York Times*, January 20, 2004, p. 18; and C. Morrison Paul and R. Nehring, "Product Diversification, Production Systems, and Economic Performance in U.S. Agricultural Production," *Journal of Econometrics*, 126 (June, 2005), 525–48.

In our discussion of a multiproduct cost function such as $C(q_1, q_2)$, we did not distinguish between situations in which the two outputs are somewhat related, as is the case with passenger and freight rail service, and those where the two goods are substantially different products, say cologne and shirts. In the latter case, the two products use quite different production processes and the presence of scope economies seems less compelling. It seems more likely that scope economies will be found when the goods being produced use similar production techniques since then we are more likely to find shared inputs and cost complementarities.

We expect scope economies to be most prevalent in the joint production of different varieties of the same good, because in that case production similarities are strongest. For instance, the possibilities for cost savings due to sharing a common factor or due to cost complementarities seem clear in the case of a ready-to-eat cereal manufacturer producing many varieties of essentially the same wheat-based cereal product. It is probably also true for a firm such as Campbell's that produces a wide variety of prepared foods, most notably, soups. To consider these issues, we need to conceptualize more clearly the meaning of different varieties of the same good. For this purpose, we now introduce a model of product differentiation that will be used extensively in later chapters.

To speak about differentiated products in a rigorous way requires that we have some way to measure just how differentiated they are. One way to do this is to imagine that some particular characteristic is the critical distinguishing feature between different versions of the good. In the case of cars, this characteristic could be speed or acceleration. In the case of soft drinks, it could be sugar content. We can then construct an index to measure this feature. Each point on the index, ranging from low to high acceleration capacity, low to high sugar content, or whatever, represents a different product variety. Some consumers will prefer a car that accelerates rapidly or a very sweet beverage, while others will favor cars that

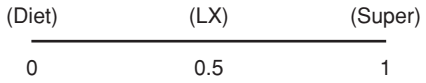

Figure 4.2 Location of cola products along the sugar content line

are easier to drive because they are capable of less acceleration or beverages with very low sugar content.

As an example, imagine a soft-drink company considering the marketing of three versions of its basic cola: (1) Diet or sugar free; (2) Super, with full sugar content; and (3) LX, an intermediate cola with just half the sugar content of Super. In this case, the distinctive feature separating each product type is sugar content, and so we want to construct an index of sugar content. It is customary to normalize such an index so that it ranges from 0 to 1. The spectrum of products for our imaginary company, therefore, ranges from Diet, located at point 0 on our index, to Super, located at point 1, with LX positioned, let's say squarely in the middle at point 0.5. This is illustrated in Figure 4.2.

The spectrum shown in Figure 4.2 may be alternatively regarded as a street. In turn, we may regard consumers as being located at different "addresses" on this street. Consumers who really like sugar will have addresses close to the Super product line. In contrast, consumers who really need to watch their calorie intake will have addresses near the Diet product line. Similarly, consumers who favor more than a medium amount of sugar but not quite so much as that contained in the Super variety, will have addresses somewhere between the LX and Super points.

It is reasonable to suppose that scope economies will exist for a firm producing different varieties of a common good, such as the various soft drink products just described. Such scope economies have become stronger in recent decades following the introduction of new manufacturing techniques, referred to as flexible manufacturing systems. They can be defined as "production unit(s) capable of producing a range of discrete products with a minimum of manual intervention" (U.S. Office of Technology Assessment, 1984, p. 60). The idea here is that production processes should be capable of switching easily from one variant of a product to another without a significant cost penalty.

A common example of a flexible manufacturing system is found in the popular clothing manufacturer, Benetton. Almost everyone is familiar with Benetton's advertisements and its array of brightly colored sweaters, T-shirts, and jeans. In fact, the coloring process is a distinctive feature of Benetton's manufacturing technology. The dyeing of the goods is done at the last moment just before shipment to the stores. Using computer-programmable equipment, Benetton is able to shift from one color-specific order to another with minimal adjustment costs. In other words, Benetton's extensive use of computer-assisted-design/computer-assisted-manufacturing (CAD/CAM) technology allows it to produce a wide array of differentiated (by color) products. In recent years, other firms have been similarly aided by CAD/CAM technology. Benjamin Moore paints and Toyota cars are just two of many companies that have used this technology to offer a wide range of choices within the same basic product line.

If scope economies exist, firms have a strong incentive to exploit them. Doing so will lower the firm's costs, possibly permit the firm to exploit multiproduct scale economies, and allow it to obtain a closer match between the products that offered and those desired by specific customers. Eaton and Schmitt (1994) show that this is exactly what happens in a formal model of flexible manufacturing in which there are k possible versions of the good. They show that

Reality Checkpoint

Flexible Manufacturing at Lands' End

In October of 2000, Lands' End started to offer custom-made pants on its website. Customers interested in buying shirts, blouses, chinos, or jeans can simply go to the firm's website and type in measurements such as weight and height, and the characterization of the proportions of their bust, hips, and general body shape. Customers can also choose the fabric and color, and stylistic features such as cuffs, hemming, pocket dimensions, and so forth. A computer program then analyzes the information, calculates the precise design, and sends the information to a manufacturing plant in Mexico. At the plant, a computerized cutting machine creates the pattern and the item is cut and sewn and shipped to customers two to four weeks later depending on the volume of orders.

The price in 2007 for a customized pair of traditional fit men's ringspun denim jeans was $70. This compared to a price of about $40 for a comparable non-customized pair of jeans at the same website. Shipping was $6 in both cases. Lands' End can charge so much more for the customized jeans because consumers are getting exactly what they want in these products. Indeed, within a year of launching the customized service, the percentage of jeans sold at the Lands' End website that were customized rose from 0 to 40. The custom service also helped Lands' End to reduce the amount of unwanted merchandize in its warehouse at the end of each season. In turn, this reduced carrying costs and further raised the profit per item, in part because fewer clothes were sold at clearance.

Source: B. Tedeschi, "E-Commerce Report: A Lands' End Experiment in Selling Custom-made Pants Is A Success, Leaving Its Rivals to Play Catch-Up," *The New York Times*, September 30, 2002, p. C3.

when scope economies are very strong, it will be natural for each firm in the industry to produce the entire range of k products. In addition, the presence of strong scope economies also tends to give rise to important multiproduct scale economies, and this suggests that the industry will be concentrated. Moreover, even weak scope economies may be sufficient to imply that it is less costly to organize production in a smaller number of firms. That is, it will be less costly to have fewer firms producing a range of products rather than to have a firm producing each product separately. In short, the presence of scope economies in the production of differentiated products tends to increase market concentration in such industries.[8]

4.4 NONCOST DETERMINANTS OF INDUSTRY STRUCTURE

So far, we have focused on the role of cost relationships, especially scale and scope economies, as being the main determinants of firm size and industry structure. There are, however, other factors that can play an important role. Here, we mention three factors

[8] See Panzar (1989) for a good discussion of cost issues in general. See Evans and Heckman (1986) and Roller (1990) for evidence of scope economies in the telephone industry; Cohn et al. (1989) and DeGroot et al. (1991) for evidence of scope economies in higher education; and Gilligan et al. (1984) and Pulley and Braunstein (1992) for evidence of scope economies in finance.

specifically. These are: (1) the size of the market; (2) the presence of network externalities on the demand side; and (3) the role of government policy.

4.4.1 Market Size and Competitive Industry

The influence of market size on industry structure has been extensively investigated by Sutton (1991, 2001). The fact that a firm must be large to reach the minimum efficient scale of operations does not necessarily imply a highly concentrated structure if the market in question is large enough to accommodate many such firms. Similarly, the fact that it is cheaper to produce many different products (or many versions of the same product) in one firm rather than in several firms does not necessarily imply a market dominated by a few firms. Most farms produce more than one crop. Yet farming is a very competitively structured industry in part because the market for agricultural products is so extensive.

Just how big does a market have to be in order to avoid domination by a few firms? The answer: it depends. When scale economies are extensive, for example when sunk or fixed costs associated with indivisible inputs are relatively large, the market will need to be greater to accommodate more firms. Thus, the relationship between market structure and market size will vary according to the specific market being examined.

If scale economies are exhausted at some point and if sunk entry costs do not rise with the size of the market then we ought to see that concentration declines as market size grows sufficiently large. Some direct evidence of this effect is provided by Bresnahan and Reiss (1991). They gathered data on a number of professions and services from over 200 towns scattered across the western United States. They find that a town of about 800 or 900 will support just one doctor. As the town grows to a population of roughly 3,500, a second doctor will typically enter. It takes a town of over 9,000 people to generate an industry of five doctors. The same positive relationship between market size and the number of firms is also found in other professions. For tire dealers, for example, Bresnahan and Reiss find that a town of only 500 people is needed to support one tire dealer and that five tire dealers will emerge when the town reaches a population of 6,000. The smaller market requirements needed to support a given number of tire dealers instead of doctors probably reflects, among other things, the fact that doctors have higher fixed/sunk costs than do the tire dealers.

Sutton (1991, 2001) however, provides an important qualification to the idea that concentration will decline with the size of the market, for example as implied by equation (4.4). He notes that such a relationship does not appear to hold in a number of industries, particularly in industries that compete heavily using either advertising, such as processed foods, or R&D, such as pharmaceuticals. Sutton argues that these expenditures are not only sunk but also endogenous. They are sunk in that once the expenditures for a promotional campaign or product design have been incurred, they cannot be recovered. They are endogenous in that in these kinds of industries, sunk cost F is not fixed but in fact increases as the market size grows.

The logic of the Sutton argument can be seen by focusing on the sunk entry cost term F in equation (4.4). Assume that this term reflects advertising and/or R&D expenditures. However, rather than simply assuming that such expenditures are equal to some exogenous level F, assume instead that they are related to market size. For example, we may assume a linear relationship of the form:

$$F = K + \beta(AE) \tag{4.6}$$

Where recall that A is a constant and E is aggregate consumer expenditure in the industry.

Using (4.6) equation (4.4) now may be written as:

$$N^e = \left[\frac{1}{\left(\frac{K}{AE} \right) + \beta} \right]^{\frac{1}{1+\alpha}} \tag{4.7}$$

Equation (4.7) says that the equilibrium number of firms in the industry will grow as market size AE grows but that this process has an asymptotic limit. Specifically, the number of firms will never exceed $(1/\beta)^{1/(1+\alpha)}$ no matter how large the market gets. For example, suppose that $\alpha = 1$ and $\beta = 0.0625$. If this is the case, then the equilibrium number of firms in the industry will never exceed four, regardless of market size.[9]

Somewhat similarly, our initial derivation of equation (4.4) assumed that price cost margins declined as a result of an increase in the number of firms as described by: $(P - c)/P = A/N^{\alpha}$. However, there may be systematic differences between industries in the relationship between the price cost margin and the population of firms. In particular, markets in which firms sell a homogenous product and in which they can quickly alter production to meet demand may have very small price cost margins. This is because in such homogenous good markets, the firm with the lowest price gets all the customers, especially if it can readily adjust output to meet that demand. Algebraically, this means that the parameter α above will differ across markets. It will be larger in those markets in which competition is naturally more intense. In such markets, the equilibrium number of firms will be correspondingly smaller.

4.4.2 Network Externalities and Market Structure

It's not news to anyone reading the recent press that there is basically only one firm producing operating systems for personal computers and that firm is *Microsoft*. For more than a decade the *Microsoft Corporation* has supplied about 95 percent of the market for operating systems for the personal computer market. Similarly, *Microsoft Word* and *Microsoft Excel* have nearly as great a share of the word-processing and spreadsheet software business. Scale and scope economies are undoubtedly part of the explanation for the highly concentrated nature of these markets. After all, once the costs have been sunk to design the basic program for the operating systems or application software, the cost of reproducing the product many times over is quite trivial. It is also highly likely that there will be a large common component to these design costs.

However, as many witnesses testified at the *Microsoft* antitrust case of 1999–2000, scale and scope economies are not the only reasons behind the dominance of this high-technology firm. A particularly important factor explaining the high concentration in this market is the presence of a demand factor known as *network externalities*. Network externalities refer to the phenomenon by which a consumer's willingness to pay for a good or service increases as the number of other consumers buying the product rises.

Telecommunications is an area in which network externalities are particularly strong. Consider the telephone for example. The usefulness or value of a single consumer connecting to a

[9] See Baldwin (1995) for some evidence on this point.

telephone system is essentially nil. If no one else is connected, the telephone cannot be used to make even one call. However, as more people sign on to the system, the number of potential calls and hence the utility of owning a phone increases as well. That is, each customer's individual decision to join the system confers benefits to the other customers—benefits that are therefore external to the consumer who is signing on. This is what we mean by a network externality. When market demand exhibits such an externality, there is a strong incentive for a firm to try to get a large number of consumers signed on to its network. To put it another way, any telephone system without a large number of customers would not be able to survive because it would not be very valuable to the few customers it does have.

We address the topic of network externalities more extensively in Chapter 24. However, from the brief discussion above, it should be relatively easy to see that markets with important network externalities are likely to be ones populated by a few very large firms. In other words, they are likely to have a highly concentrated structure—even if scale economies are not present on the cost side. Indeed, many analysts view network externalities as a case of scale economies that exist on the demand side of the market.

4.4.3 The Role of Government Policy

From 1934 to 1988—a period of 54 years—the number of medallions authorizing legal ownership of a taxicab in Boston was fixed at 1,525. Not a single additional medallion was issued in all that time despite the fact that the regional population increased by over 50 percent and the level of income and economic activity doubled several times over. Costs and technology were not the source of this fixed industrial structure. The primary reason for the limited entry into the Boston taxi industry was government policy. City and state officials deliberately limited the number of taxi medallions, largely at the request of those lucky taxi owners who obtained the first batch of medallions. Even in recent years with a court order to issue 300 new medallions outstanding for nearly five years, only a few additional ones have actually been issued as officials have again tried to slow the creation of additional legal taxi operators as much as possible.

A similar phenomenon prevailed from the 1930s through the 1970s when the number of so-called trunk airlines flying interstate routes never exceeded 16 and fell to 10 by the end of the 1970s. Not only was the total number of airlines small on a national scale, it was even smaller for individual city-pair markets. Many of these were often served by only one or two carriers. Here again, the primary cause was government policy. In this case, that policy was implemented by the Civil Aeronautics Board (CAB), the federal agency established in 1938 as the economic regulator of the airline industry. Throughout its existence, the CAB deliberately limited entry and sustained a high concentration level in the U.S. domestic airline industry. Indeed, this 40-year period witnessed numerous applications by freight and charter airlines to be granted the right to offer scheduled passenger services, as well as frequent applications of existing passengers to enter new city-pair markets. Virtually all of these requests were turned down. The CAB argued that this policy was necessary to promote the stability and healthy development of the airline industry. Whether it achieved its perceived goals, or whether such goals were appropriate, is a question to be answered elsewhere. The central point illustrated by both the taxicab and airline example is that explicit government policies often play an important role in determining market structure.

More often than not, the role of government policy has been to increase market concentration as both of the examples above illustrate. However, some government policies do work to increase the number of firms in an industry. The Robinson–Patman Act that prohibits price

discounts to large firms if such discounts are deemed anticompetitive reflects a conscious effort to keep independent retailers in business. These are typically small firms who otherwise would have been driven out of the market by the large retail chains. Similarly, the decision of the U.S. government after the Second World War to force the Alcoa Company to sell some of its wartime aluminum plants to the Kaiser and Reynolds corporations was clearly an effort to promote a more competitive structure. Perhaps most obviously, antitrust policies that lead either the Federal Trade Commission or the Justice Department to block mergers also increase the equilibrium number of firms in an industry.

4.5 EMPIRICAL APPLICATION
Cost Function Estimation—Scale, and Scope Economies

Since the underlying technology and associated cost implications are central determinants of industrial structure, economists have been interested in getting evidence on cost relationships for a long time. Unfortunately, we rarely have direct evidence on the production technology. Hence, estimating firm cost functions can be a tricky business. However, application of basic microeconomic theory can greatly facilitate the process.

To see this let us suppose that production is generated from capital K and labor L inputs using a Cobb-Douglas production function of the form:

$$Q = K^{\alpha}L^{\beta} \tag{4.8}$$

Total cost is simply the cost of the inputs. If r is the rental price of capital and w is the wage rate of labor, then total cost $C = rK + wL$. Cost minimization requires choosing the capital and labor inputs that minimize cost for achieving a given level of output. If we denote the target output level as $\bar{Q}$, the firm's problem then becomes:

$$\text{Minimize } C = rK + wL \text{ subject to } \bar{Q} = K^{\alpha}L^{\beta} \tag{4.9}$$

While there are many ways to solve this minimization problem, one way is to use the production requirement to substitute out the labor input. That is, the production constraint implies: $L = \bar{Q}^{\frac{1}{\beta}}K^{-\frac{\alpha}{\beta}}$. The cost function then becomes $C = rK + w\bar{Q}^{\frac{1}{\beta}}K^{-\frac{\alpha}{\beta}}$. If we hold output $\bar{Q}$ constant at the target level, we can minimize this expression with respect to the capital input by setting its derivative with respect to K equal to zero. Solving for K yields:

$$K = \left(\frac{\alpha}{\beta}\frac{w}{r}\right)^{\frac{\beta}{\alpha+\beta}}\bar{Q}^{\frac{1}{\alpha+\beta}} \tag{4.10}$$

If we now substitute the above value for K into the labor requirement implied by the production constraint, we may then solve for the optimal or cost minimizing labor input and we obtain:

$$L = \left(\frac{\beta}{\alpha}\frac{r}{w}\right)^{\frac{\alpha}{\alpha+\beta}}\bar{Q}^{\frac{1}{\alpha+\beta}} \tag{4.11}$$

Together, expressions (4.10) and (4.11) imply that the minimal cost for producing any given level of output $\bar{Q}$ is:

$$C = w\left(\frac{\beta}{\alpha}\frac{r}{w}\right)^{\frac{\alpha}{\alpha+\beta}}\bar{Q}^{\frac{1}{\alpha+\beta}} + r\left(\frac{\alpha}{\beta}\frac{w}{r}\right)^{\frac{\beta}{\alpha+\beta}}\bar{Q}^{\frac{1}{\alpha+\beta}} = \left[\left(\frac{\alpha}{\beta}\right)^{\frac{\beta}{\alpha+\beta}} + \left(\frac{\beta}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}}\right] r^{\frac{\alpha}{\alpha+\beta}} w^{\frac{\beta}{\alpha+\beta}} \bar{Q}^{\frac{1}{\alpha+\beta}} \quad (4.12)$$

Apart from the fact that we have now expressed total costs as a function of input prices and the level of output, certain features of the final expression above are important to recognize. First, note that the partial derivative of the cost function with respect to either the rental cost r or the wage w is positive. In other words, if input prices rise, it becomes more costly to produce a given output. Second, observe that the exponents for each factor price, r and w, in fact sum to unity. This means that if all factor prices increase by, say, 10 percent, the minimal cost of producing a given output will also increase by 10 percent. In economics jargon, the cost function is homogeneous of degree one in factor prices. Finally, the exponent on the output level is $1/(\alpha + \beta)$. If $\alpha + \beta = 1$, then total costs will rise proportionately with output, i.e., there will be constant returns to scale. If, however, $\alpha + \beta > 1$, then costs will rise less rapidly than output so that there will be scale economies. It follows that if $\alpha + \beta < 1$, there will be scale diseconomies. Finally, note that while costs are homogeneous of degree one in both input prices, they rise less than proportionately with any one input price. That is, the coefficient on either r or w alone is less than one. This reflects the fact that as one factor's price rises, the firm will substitute out of that input and into the less expensive one.

It is convenient to write the cost equation above in logarithmic form. Hence, we have:

$$\ln C = \ln\left[\left(\frac{\alpha}{\beta}\right)^{\frac{\beta}{\alpha+\beta}} + \left(\frac{\beta}{\alpha}\right)^{\frac{\alpha}{\alpha+\beta}}\right] + \left(\frac{\alpha}{\alpha+\beta}\right)\ln r + \left(\frac{\beta}{\alpha+\beta}\right)\ln w + \left(\frac{1}{\alpha+\beta}\right)\ln \bar{Q} \quad (4.13)$$

For estimation purposes, this can easily be translated into:

$$\ln C = \text{Constant} + \delta_1 \ln r + \delta_2 \ln w + \delta_3 \ln Q \quad (4.14)$$

With observations on input prices and output levels, we may estimate the above equation and then use the estimated coefficients δ_i to recover the underlying production parameters, α and β. The properties of a well-behaved cost function discussed above suggest that both δ_1 and δ_2 should be positive, less than one separately, but sum to one together. Further, since changes in logarithms translate into proportional changes in the underlying variable, e.g., $\Delta \ln C \approx \Delta C/C$, a measure of the elasticity of costs with respect to output or η_C is provided by $\partial \ln C/\partial \ln Q$, *which* in this case, is reflected in the estimate of δ_3. Thus, our measure of scale economies derived earlier is: $S = 1\Big/\dfrac{\partial \ln C}{\partial \ln Q} = 1/\delta_3$.

However, the above derivation is based on the assumption that the underlying production technology is of the Cobb–Douglas type. Because that may be a rather strong assumption, empirical analyses often use a more flexible cost specification that permits a much wider

array of underlying technologies including, as a special case, the Cobb-Douglas specification above. One such flexible specification is the translog cost function. For the basic, two-input case above, this function has the form:

$$\begin{aligned}\ln C = \text{Constant} &+ \delta_1 \ln r + \delta_2 \ln w + 0.5[\delta_{11}(\ln r)^2 + \delta_{12}(\ln w)(\ln r) \\ &+ \delta_{21}(\ln w)(\ln r) + \delta_{22}(\ln w)^2] + \delta_3 \ln Q + \delta_{31}(\ln Q)(\ln r) \\ &+ \delta_{32}(\ln Q)(\ln w) + 0.5\delta_{33}(\ln Q)^2 \end{aligned} \tag{4.15}$$

As before, we expect δ_1 and δ_2 to be positive fractions that sum to unity. However, our measure of scale economies $S = 1\Big/\dfrac{\partial \ln C}{\partial \ln Q}$ is now no longer necessarily constant but instead can depend on the level of output. That is we now have: $S = 1\Big/\dfrac{\partial \ln C}{\partial \ln Q} = 1/(\delta_3 + \delta_{33} \ln Q + \delta_{31} \ln r + \delta_{32} \ln w)$. Only if $\delta_{31} = \delta_{32} = \delta_{33} = 0$, will the index of scale economies be independent of the level of Q. This is a restriction that one can test. Indeed, use of a translog function such as (4.15) above allows the researcher to test numerous restrictions. For example, if $\delta_{31} = \delta_{32} = 0$ but $\delta_{33} \neq 0$, then while the technology is not homogeneous, the mix of capital and labor inputs does not change as output expands and the production technology exhibits what economists call homotheticity.

One of the earliest papers estimating a translog cost function is also one of the most illustrative. Christensen and Greene (1976) applied the translog approach to the electric power-generating industry adding fuel as a basic input along with labor and capital. Denoting the price of fuel as F, their estimation led to an equation with a constant and nine input price terms and five output terms. The five output variables are the pure output term $\ln Q$, the three interaction terms between the output term and each of the input price terms, and the $(\ln Q)^2$ term. Here we focus on these five terms because these are the ones that will indicate the extent of any economies of scale. The Christensen and Greene (1976) estimates for these terms are shown in Table 4.2.

Note that the interaction terms are virtually all statistically significant indicating that these interaction effects belong in the equation. Indeed, when Christensen and Greene (1976) estimate the cost equation with only the first term $\ln Q$ included, as would be implied by a Cobb-Douglas technology, they obtain a coefficient estimate on the order of 0.8. This would yield a scale economy index of $S = 1/0.8 = 1.25$ and indicate substantial unexploited scale economies for all firms. Including the additional terms permits the scale effect to vary by firm size as measured by the volume of output. On this basis, they then find that few firms operate with $S > 1.17$. Further, fully half of the electrical power was generated by firms that were sufficiently large that no further scale economies were present.

Table 4.2 Christensen and Greene (1976) cost function estimates and (85) scale economies in electric power generation

Variable	*Coefficient*	*t-statistic*
$(\ln Q)$	0.587	20.87
$(\ln Q)(\ln r)$	–0.003	–1.23
$(\ln Q)(\ln w)$	–0.018	–8.25
$(\ln Q)(\ln F)$	0.021	6.64
$(\ln Q)^2$	0.049	12.94

It is relatively straightforward to adapt the translog cost function to the case of multiproduct firms. If, for example, we continue to assume just two inputs with prices r and w, but now instead assume two outputs, Q_1 and Q_2, the function has the form:

$$\begin{aligned}\ln C = \text{Constant} &+ \delta_1 \ln r + \delta_2 \ln w + 0.5[\delta_{11}(\ln r)^2 + \delta_{12}(\ln w)(\ln r) \\ &+ \delta_{21}(\ln w)(\ln r) + \delta_{22}(\ln w)^2] + \delta_3 \ln Q_1 + \delta_{31}(\ln Q_1)(\ln r) \\ &+ \delta_{32}(\ln Q_1)(\ln w) + 0.5\delta_{33}(\ln Q_1)^2 + \delta_4 \ln Q_2 + \delta_{41}(\ln Q_2)(\ln r) \\ &+ \delta_{42}(\ln Q_2)(\ln w) + 0.5\delta_{44}(\ln Q_2)^2 + \delta_5(\ln Q_1)(\ln Q_2) \qquad (4.16)\end{aligned}$$

However, in applying equation (4.16) to data gathered from many firms, there is a real possibility that some of the firms will be single-product enterprises for which either Q_1 or Q_2 is zero. This creates a serious problem in estimating the cost relationships because the logarithm of zero is not well defined. Hence, when multiple outputs are considered, many researchers follow the suggestion of Caves, Christensen, and Tretheway (1980) and use what is known as a Box-Cox transformation of the output variables. Under this transformation, the log of the level of good i production is replaced with the term $(Q_i - 1)/\theta$, where θ is estimated along with other parameters. Note that the limit of $(Q_i - 1)/\theta$ as θ approaches zero is in fact $\ln Q_i$.

Once the multiproduct cost function has been estimated, it is straightforward to derive the scale economy and scope economy measures described in the text. DeGroot, McMahon, and Volkwein (1991) use this translog approach to model the cost structure of American research universities, assuming three university outputs: (1) undergraduate education; (2) graduate education; and (3) research. They find that for the product mix of the typical university there were significant unexploited scale economies (declining ray average cost). However, this was not true for the less student-intensive product mix of the top private schools for which they found little if any scale economies. They also found significant scope economies between graduate and undergraduate education but, somewhat surprisingly, little scope economies between graduate education and research.

Summary

This chapter has focused on technology and key cost concepts and the implications they have for industrial structure. Scale economies tend to increase market concentration. Economies of scope have a similar effect of concentrating the production of different products within a single firm. Scope economies also typically give rise to important multiproduct scale economies. This is particularly the case when the various products are not truly different goods but, instead, different versions of the same goods. In such product-differentiated markets, the presence of scope and scale economies will again imply a more concentrated structure.

Other factors influence market structure as well. One of these is market size. Because a large market has room for a number of firms, even if each firm is of considerable size, larger markets tend to be less concentrated than small ones. However, increasing market size does not lead to less concentration in markets in which sunk costs also increase with size. These are typically markets in which advertising or research and development costs play a major role.

Another important determinant of market structure comes from the demand side of the market in the form of network externalities. Network externalities imply that the value of a product to any one consumer increases as other consumers use it. Such externalities act much like scale economies on the demand side and they foster increased market concentration. Careful application of economic theory can generate clear implications for the statistical measurement of cost

relationships. Such work has been extremely useful in identifying scale and scope economies. For example, regression analyses based on the theory of production costs have found significant scale economies in electric power generation and important economies of scope between graduate and undergraduate education.

Finally, government policy is also a very important determinant of market structure. Regulations such as those long applied to local taxi markets and the airline industry typically reduce the ability of new firms to enter the market. Antitrust policy can raise the number of firms in a market by blocking a proposed merger.

Problems

1. Let the cost function be $C = 100 + 4q + 4q^2$. Derive an expression for average cost. Derive an expression for marginal cost. Is there any range of production characterized by scale economies? At what production level are scale economies exhausted?

2. An urban rapid-transit line runs crowded trains (200 passengers per car) at rush hours, but nearly empty trains (10 passengers per car) at off-peak hours. A management consultant argues that the cost of running a car for one trip on this line is about $50 regardless of the number of passengers. Hence, the consultant concludes that the per passenger cost is about 25 cents at rush hour but rises to $5 in off-peak hours. Consequently, we had better discourage the off-peak business. Is the consultant a good economist? Why or why not?

3. Consider the following cost relationships for a single-product firm:

   $C(q) = 50 + 0.5q$ for $q < 7$
   $C(q) = 7q$ for $q > 7$

   a. Derive average and marginal cost for all integer outputs less than or equal to 7.
   b. What are average and marginal cost for all outputs above 7?

4. In the problem above is there a minimum efficient scale of plant implied by these cost relationships? If so, what is it?

5. Let P be industry price and Q be total industry output. If the industry demand curve is $P = 84 - 0.5Q$ use the data in question 3 to determine what is the maximum number of efficient-sized firms that the industry can sustain?

6. How would your answer to 5 be changed if industry demand were instead $P = 14 - 0.5Q$? Explain.

7. Some estimates for the cement industry suggest the following relationship between capacity and average cost:

Capacity (thousands of tons)	*Average cost*
250	28.78
500	25.73
750	23.63
1,000	21.63
1,250	21.00
1,500	20.75
1,750	20.95
2,000	21.50

   a. At what production level are scale economies exhausted?
   b. Calculate the scale economy index for the production levels 500, 750, 1,000, 1,500, and 1,750.

8. A newspaper article (J. Peder Zane, "It Ain't for the Meat; It's for Lotion," *New York Times*, Sunday, May 5, 1996, p. E5) presented the following data for a cow brought to market:

Part	*Use*	*Price/lb ($)*
Horns	Gelatin Collagen	0.42
Cheek	Sausage Baloney	0.55
Adrenal gland	Steroids	2.85
Meat	Beef	1.05
Lips	Taco filling	0.19
Hide	Footwear Clothing	0.75

   Comment on the scope economies illustrated by this example. What is the source of such economies? What does the existence of such economies imply about the supply of such products as leather skins, beef, and gelatin powder?

References

Baldwin, John R. 1995. *The Dynamics of Industrial Competition: A North American Perspective*. Cambridge: Cambridge University Press.

Baumol, W. J., J. C. Panzar, and R. D. Willig. 1982. *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt, Brace, Jovanovich.

Bolton, P. and A. Scharfstein. 1998. "Corporate Finance, the Theory of the Firm, and Organizations." *Journal of Economic Perspectives* 12 (Autumn): 95–114.

Bresnahan, T., and P. Reiss. 1991. "Entry and Competition in Concentrated Markets." *Journal of Political Economy* 99 (October): 977–1009.

Caves, D., L. Christensen, and M. W. Tretheway. 1980. "Flexible Cost Functions for Multiproduct Firms." *Review of Economics and Statistics* 62 (August): 477–81.

Chenery, H. 1949. "The Engineering Production Function." *Quarterly Journal of Economics* 63 (May): 507–31.

Christensen, L. and W. Greene. 1976. "Economies of Scale in U.S. Electric Power Generation." *Journal of Political Economy* 84 (August): 655–76.

Coase, R. H. 1937. "The Nature of the Firm." *Economica* 4 (March): 386–405.

Cohn, E., S. L. Rhine, and M. C. Santos. 1989. "Institutions of Higher Education as Multi-Product Firms: Economies of Scale and Scope." *Review of Economics and Statistics* 71 (May): 284–90.

De Groot, H., W. McMahon, and J. F. Volkwein. 1991. "The Cost Structure of American Research Universities." *Review of Economics and Statistics* 73 (August): 424–31.

Dunne, T., M. J. Roberts, and L. Samuelson. 1988. "Patterns of Firm Entry in U.S. Manufacturing Industries." *Rand Journal of Economics* 19 (Winter): 495–515.

Eaton, B. C., and N. Schmitt. 1994. "Flexible Manufacturing and Market Structure." *American Economic Review* 84 (September): 875–88.

Evans, D. and J. Heckman. 1986. "A Test for Subadditivity of the Cost Function with Application to the Bell System." *American Economic Review* 74 (September): 615–23.

Gale, I. 1994. "Price Competition in Non-cooperative Joint Ventures." *International Journal of Industrial Organization* 12 (March): 53–69.

Gilligan, T., M. Smirlock, and W. Marshall. 1984. "Scale and Scope Economies in the Multi-Product Banking Firm." *Journal of Monetary Economics* 13 (May): 393–405.

Hart, O. 1995. *Firms, Contracts, and Financial Structure*. New York: Oxford University Press.

Hart, O. and J. Moore. 1990. "Property Rights and the Nature of the Firm." *Journal of Political Economy* 98 (December): 1119–58.

Milgrom, P. and J. Roberts. 1992. *Economics, Organization, and Management*. Upper Saddle River, NJ: Prentice Hall.

Panzar, J. C. 1989. "Technological Determinants of Firm and Industry Structure." In R. Schmalensee and R. Willig, eds., *Handbook of Industrial Organization*. Vol. 1. Amsterdam: North-Holland, 3–60.

Pulley, L. B. and Y. M. Braunstein. 1992. "A Composite Cost Function for Multiproduct Firms with an Application to Economies of Scope in Banking." *Review of Economics and Statistics* 74 (May): 221–30.

Roller, L. 1990. "Proper Quadratic Cost Functions with Application to the Bell System." *Review of Economics and Statistics* 72 (May): 202–10.

Sutton, John. 1991. *Sunk Costs and Market Structure*. Cambridge, MA: MIT Press.

—— . 2001. *Technology and Market Structure*. Cambridge, MA: MIT Press.

Williamson, O. E. 1975. *Markets and Hierarchies: Analysis and Antitrust Implications*. New York: Free Press.