4

Technology and Cost

In late 2001, Microsoft introduced its first *Xbox* video game console. The relatively simple model for which only a few games were available sold at a retail price of $300. By 2004, the price had fallen to $150 despite technical improvements, as Microsoft prepared to launch the new *Xbox 360* series. The base or *Core* model 360 emerged in 2005 at a price of $300—a price that fell to $200 in 2008 as the *Core* model was replaced by the *Arcade* series with initially a 60GB hard drive. That price fell further over the next two years, even as the hard drive capacity grew to 250GB and more features and games were added. As this text is going to press, an *Xbox 360S* series along with a *Kinect* scanner that allows full-body gaming is selling as a bundle for $250. In sum, the price of an *Xbox* game console has declined over the last decade—especially if the vastly improved technical features are considered—despite the fact that the prices of other consumer goods have risen over 25 percent (about 2.6 percent per year) in that same time period.

Clearly, Microsoft's pricing strategy reflects in part its rivalry with other video game platform producers such as Sony and Nintendo. Yet the above prices must also reflect *Xbox* production costs because the profitability of any price choice will depend critically on the cost of producing each unit of the total sold at that price. More generally, production costs affect both firm behavior and industrial structure. The four firms—General Mills, Kellogg, General Foods (Post), and Quaker Oats currently account for about 80 percent of sales in the US ready-to-eat breakfast cereal industry. By contrast, the largest four manufacturers of games and toys account for 35 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 such dramatic differences in 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 in which a firm is solely

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

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 change 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 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?

The questions raised by Coase (1973) cannot be answered within the simple neoclassical view of the firm. This is not to say though that economics cannot address these issues at all. Over the years, a large number of scholars have developed theories of the firm typically rooted in the recognition of two facts. First, it is typically not possible to write contracts that cover all possible contingencies. Contracts must therefore be incomplete. Second, information is often asymmetric. Managers have information that is not accessible to shareholders. Employees know things that managers do not.

The fact that contracts cannot be complete may require an authority relationship, typically within a firm as a means of dealing with all the contingencies not covered by the contract (Williamson 1975). This is particularly true when specialized investments are required that will only be profitable if the two contracting parties, say a firm and its supplier, continue working together. Otherwise, once such an investment is made, the party that paid for it may find itself at a disadvantage in setting contract terms, e.g., the supply price because it has little alternative use for these assets (Williamson 1975, Grossman and Hart 1986, Hart and Moore 1990, and Hart 1995). Likewise, the fact of asymmetric information can give rise to free-riding and moral hazard problems that can only be resolved by an organizational structure that includes someone who monitors the work efforts of others and, in return, receives a residual payment to insure that there is an incentive to do this job efficiently (Alchian and Demsetz 1972, Holmstrom 1982, and Holmstrom and Milgrom 1994). These and related issues concerning the nature of the firm continue to be important topics of economic research, as evidenced by Daniel Spulber's recent book (2009).

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. Therefore keep in mind throughout the following discussion that

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

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 $\overline{q}$ 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 $\overline{q}$. The firm obtains this solution by choosing the 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 $f(x_1,\ x_2,\ \ldots,\ x_k) = \overline{q}$.

If we solve this problem for different levels of output $\overline{q}$, 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 denote 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 represents 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 zero or a thousand 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.[3]

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 only incurred prior to a specific date—often prior to the entry date. 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 will be difficult to sell the license to another doctor and certainly 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 close to 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 US car manufacturing plants—production lines, robots, and other highly specialized machinery—likewise 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 firm chooses later 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

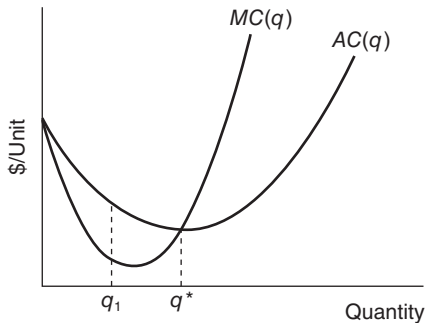

Figure 4.1 Typical average and marginal cost curves

[3] Formally, marginal cost is the slope of the total cost function and so is defined by the derivative term, $MC(q) = dC(q)/dq$. (See Appendix.) Note that in the case of a multiplant firm, shifting one unit of production to plant i from plant j while leaving total output unchanged, would change total cost by an amount: $MC_i - MC_j$. It follows that if $C(q)$ is a minimized, marginal cost is equalized across all plants.

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 the 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.

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

Closer consideration of Figure 4.1 reveals 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^*, the range through which average cost constantly falls. Conversely, marginal cost is greater than average cost for outputs greater than q^*, the range through which average cost constantly rises. 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. It follows immediately that marginal cost and average cost are equal when average cost is minimized.

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	1024	128	113	1.13
9	1143	127	123	1.03
10	1270	127	132	0.96
11	1408	128	151	0.85
12	1572	131	–	–

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.[4] The table documents the point made above that average cost falls when it lies above marginal cost, rises when it is below marginal cost, and (because the numbers are a discrete approximation) is essentially equal to marginal cost at the minimum average cost value. The Appendix to this chapter presents a formal derivation of cost measures and cost functions.

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.

When average cost falls as output increases, it means 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 sixty miles of track whether the number of trains per day is one or twenty. 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 gives 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 200 years ago.[5] This in turn permits specialization and more efficient production. Sometimes, the simple mathematics of the activity give 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 10x10x10 cube will hold 1,000 cubic feet, a 20x20x20 cube holds 8,000 cubic feet. Because the cost in terms of materials and labor depends on surface area but output depends on volume, it follows

[4] In Table 4.1, marginal cost is calculated as the average of the increase in cost associated with producing one unit more and the decrease in cost associated with producing one unit less.

[5] 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.

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.[6] 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.[7]

Whatever the source of the scale economies, the fact that scale economies are indicated 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 of average to marginal cost. That is:

$$S = \frac{AC(q)}{MC(q)} = \frac{[C(q) + F]/q}{MC(q)} = \frac{C(q) + F}{qMC(q)} \tag{4.2}$$

The more that S exceeds 1, the greater is the extent of scale economies. In such a setting, a 1 percent increase in output is associated with a less than 1 percent increase in costs. Conversely, when $S < 1$, diseconomies of scale are present. Increasing output by 1 percent now leads to more than a 1 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 the 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:

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

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 $16.67\%/12.37\% = 1.35\%$. This is not far from the measure of S ($= 1.42$) that we obtained using the ratio of average to marginal

[6] The classic study by Chenery (1949) on natural gas pipelines is an example of this technical relationship.

[7] 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).

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 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, S measures the proportionate increase in output one obtains for a given proportionate increase in costs.

4.1 Practice Problem

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 produces 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 an efficient 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 the extent of scale economies—the larger the output at which average cost is minimized—the fewer are the firms that can operate efficiently in the market. Thus, large scale economies will tend to result in concentrated markets.

4.2 Practice Problem

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, $\Delta q = -1$. Because 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 $S < 1$?

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 as $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 United States 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 a piece. 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 United States 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 just the cost of such phones that's fixed.

Source: C. Elliot, "Mystery of the Cell Phone that Doesn't Work at the Hotel," *New York Times*, 7 September, 2004, p. C8, and C. Page "Mobile Phones Jam Scam," *The Daily Record*, 26 August, 2003, p1.

4.2 SUNK COST AND MARKET STRUCTURE

Sunk costs also play a role in influencing market structure in a way 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. 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.

The foregoing logic permits us to see clearly the role of sunk cost in determining market structure. Imagine for example 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 because a 1 percent decrease in price is balanced by a 1 percent increase in quantity sold. Denote then 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, that is, $Q = Nq_i$. Putting these two relationships together we then obtain:

$$q_i = E/NP \tag{4.3}$$

Now recall the Lerner Index that was 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: $(P - c)/P$. Because 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:

$$(P - c)/P = A/N^{\alpha} \tag{4.4}$$

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.[8] Substituting this break-even requirement into equation (4.3) and combining that equation with equation (4.4) then yields that the equilibrium number of firms N^e at which each entrant just covers its sunk entry cost F, is given by:

$$N^e = \left[\frac{AE}{F}\right]^{\frac{1}{1+\alpha}} \tag{4.5}$$

Clearly, N^e declines as F rises. While the precise results of equation (4.5) reflect the particular assumptions made for our example, the intuition underlying those results is fairly general. Industry structure is likely to be more concentrated in markets where sunk entry costs are a high proportion of consumer expenditures.

4.3 COSTS AND MULTIPRODUCT FIRMS

Because scale economies are a description of the behavior of costs as output increases, investigating their existence in any industry requires that we measure the 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 US railroads carried passengers as well as freight.

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

Elsewhere in the world, this is still the case. Because 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 transport being provided.

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. Bernard, Redding, and Schott (2009), for example, find that 39 percent of US manufacturing firms produce more than one major product and that these firms account for 87 percent of total US production. Thus, 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 TVs, stereos, game consoles, and so on. Measuring the total output of these firms in a single index is clearly less than straightforward. 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.

4.3.1 Multiproduct Scale and Scope Economies

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 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.

To understand the basic idea of Ray Average Cost, consider a firm with two products, q_1 and q_2. Let us imagine that these two products may be combined to create a composite good called $q = q_1 + q_2$. However, this composite good will be different depending on how much of each good we have. A composite good made up of one unit of product 1 and five units of product 2 is very different from one comprised of three units of product 1 and one unit of product 2. Because we want to consider costs for a specific product, we choose a production ray along which the two goods are always produced in the same proportion. A possible production ray, for example, is one in which the production of good 1 is always twice that of good 2, that is, $q_1 = 2q_2$. Now we can consider the behavior of costs for all output levels in which this proportion is maintained—for $q_1 = 2$ and $q_2 = 1$; $q_1 = 20$ and $q_2 = 10$; and so on. By focusing on a specific production ray, we isolate the change in cost that is due only to changing the level of output and not that due to changes in the product mix.

Continuing with our example, suppose that $q_1 = 60$ so that $q_2 = 30$. Then it follows that $q = q_1 + q_2 = 90$. Note that we could write this alternatively as $q_1 = (60/90)q$ and $q_2 = (30/90)q$. More generally, for a given product mix or production ray, we can write $q_1 = \lambda q$ and $q_2 = (1 - \lambda)q$, where $\lambda/(1 - \lambda)$ is the fixed proportions of the product mix

along the production ray being considered. In our example, $\lambda = 2/3$ and $1 - \lambda = 1/3$, so that $\lambda/(1 - \lambda) = 2$.

We denote the total cost of producing both products at any level as $C(q_1, q_2) + F = C[\lambda q, (1 - \lambda)q] + F = C(q) + F$. We denote the marginal cost of goods 1 and 2, respectively, as MC_1 and MC_2. Recall that for the single product case, our scale economy measure was the ratio of average cost to marginal cost [equation (4.2)]. Analogously, we define a measure of multi-product (in our case, two-product) scale economies as:

$$S^M = \frac{C(q_1, q_2) + F}{q_1 MC_1 + q_2 MC_2} \tag{4.6}$$

Having an explicit definition for scale economies in the multiproduct case is, as we shall see, very helpful. Perhaps a more important insight of Baumol, Panzar, and Willig (1982), however, is their introduction of an equally critical measure called 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 $C(q_1, q_2) + F$. For the two-product case, scope economies exist if $C(q_1,\ 0) + F_1 + C(0,\ q_2) + F_2 - [C(q_1,\ q_2) + F] > 0$. The first two terms in this equation are the total costs of producing product 1—the variable costs associated with just this one good plus the fixed cost if only this good is produced. The next two terms are the analogous cost measures when only good 2 is produced. The last terms are of course the total cost of having these products produced by the same firm. If the 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:

$$S^C = \frac{C(q_1,\ 0) + F_1 + C(0,\ q_2) + F_2 - [C(q_1,\ q_2) + F]}{C(q_1,\ q_2) + F} \tag{4.7}$$

The concept of scope economies is a crucial one that provides the central technological reason for the existence of multiproduct firms. Such economies 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 where the common input is the track. Another example would include a firm's advertising expenditures that benefit all of the products carrying that same brand name.

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.

Scope economies have likely 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.

Reality Checkpoint

An Arm and a Leg . . . Scope Economies and Hospital Consolidation

The health care industry has been in transformation since at least the 1990s. As health care expenditures have risen steadily—from 6 or 7 percent of GDP thirty years ago to nearly 18 percent today—patients, insurance companies, and the doctors and hospital service providers themselves have all been forced to adjust to new market pressures. One manifestation of this transformation has been a continuing wave of hospital mergers and consolidation. Since 1997, the average Herfindahl Index for hospital services in metropolitan areas has increased from 4200 to nearly 4800 today. Because the DOJ/FTC Merger Guidelines define any market with an index over 2500 to be highly concentrated and any increase over 100 points in such markets to be significant, these data show that the typical urban hospital market is one in which hospitals may have considerable and growing market power.

Cost efficiency has been a major motivation behind the hospital merger wave and scope economies are likely an important source for any such cost savings. There may, for example, be important scope economies between serving patients needing surgery and follow-up in-hospital care such as amputees and those receiving hip or knee replacements, with serving patients needing continuing disease control and/or rehabilitation because many of the nursing and monitoring services required are the same. Similar scope economies may exist between inpatient and outpatient care. In addition, scope economies may be realized by combining in one hospital both those who specialize in various functional areas with generalists who treat diseases that affect those areas and others. For example, because prostate and colorectal are two of the most common kinds of cancers, it may be less costly to house proctology and urology specialists in the same hospital as oncologists.

Of course, acquiring market power may be another motivation behind the hospital merger wave. These issues came to a head recently when the Federal Trade Commission blocked a merger between Pro Medica Health System and St. Luke's Hospital in Toledo, Ohio. The merging hospitals argued that their joining was necessary to coordinate care and controlling costs especially in light of the changes instituted by the Affordable Care Act. The FTC provided data showing that this consolidation would substantially increase concentration in the Toledo area hospital market. It also gave evidence implying that Pro Medica's hospitalization fees were among the highest. FTC officials saw the merger as a classic case of an attempt to eliminate (buy out) a low-cost rival. In the FTC's view, even if the scope economies were realized there was no guarantee that these savings would be passed on to the public given the market power the newly merged firm would have. The FTC was not against saving lives and limbs more efficiently. It just wanted to make sure that it did not end up costing patients an arm and a leg.

Source: R. Pear, "Regulator Orders Hospitals to Undo a Merger in Ohio," *New York Times*, April 3, 2012, p. A11.

Consider, for example, the popular clothing manufacturer, Benetton. This Italian family firm has over 2,500 retail outlets around the globe. Each is equipped with special cash registers that collect and transmit information regarding which items are selling, and in what colors, sizes, and styles. This information is sent to the central manufacturing facilities. These, together with numerous small sub-contracting firms, then respond quickly

to these market signals. Benetton's famous coloring is the final step of the process, as the dyeing of the goods is done at the last moment just before shipment to the stores. All of this is made possible via Benetton's extensive use of computer-assisted-design/computer-assisted-manufacturing (CAD/CAM) technology, which allows it to produce a wide array of differentiated (by color) products.

4.3.2 Different Products versus Different Versions

So far, our discussion of multiproduct firms has not distinguished between situations in which the two outputs are somewhat related, as is the case with passenger and freight rail service, and those in which 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 because then we are more likely to find shared inputs and cost complementarities.

An issue that arises in this regard is the distinction between different goods and different versions of the same good. Campbell's produces both canned soup and canned beverages, principally, *V*8. While this may reflect some scope effects, the real source of scope economies at Campbell's is undoubtedly the fact that it is cheaper to manufacture of Campbell's thirty-plus types of canned soup at one firm rather than having over thirty separate firms each producing one of those varieties. Likewise, there are almost certainly some strong scope economies at work that lead L'Oreal to offer over more than fifty shades of its *Riche Lipstick*, as well as offering other cosmetics, including lipsticks, through its subsidiaries, Maybelline and Lancôme.

Because firms so often market multiple varieties within a product category as well as possibly marketing more than one distinct good, it is useful to have a way of modeling markets in which the basic good is sold in many differentiated versions. 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, and so on, 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 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 thus we want to construct an index of such 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 interpreted 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

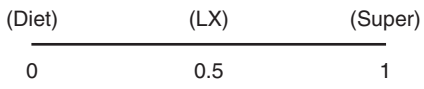

Figure 4.2 Location of cola products along the sugar content line

not quite so much as that contained in the Super variety will have addresses somewhere between the LX and Super points.

If scope economies exist, firms have a strong incentive to exploit them. If we extend this logic to the case of multiple varieties within a product category, then the implication is that firms in such markets ought to produce a wide range of product types. This logic has been formalized by Eaton and Schmitt (1994). These authors present a formal model of flexible manufacturing in which there are k possible versions of the good. They show that when scope economies are very strong, it will be natural for each firm in the industry to produce the entire range of k products.

The same incentive applies of course to scope economies across truly different product categories. Indeed, a critical insight of the analysis of multiproduct production is that it is the presence of *scope* economies as identified by equation (4.7) that makes it possible to exploit the multiproduct *scale* economies identified by equation (4.6). The airline industry offers another good example. For many years, major air carriers have organized their services around the hub-and-spoke system. Under this system, central large airports, designated as hubs, act as collector stations. Hub-to-hub flights are typically served by large aircraft for which the cost per passenger mile is low when they are filled to near capacity. Spoke airports, in contrast, serve as "feeders" that use relatively small aircraft to fly passengers from many different locations to a hub where they can all be collected and put on hub-to-hub flights to generate the passenger volume necessary to exploit the efficiencies of the larger aircraft. The multi-product scale economies can thus only be realized by operating in both the spoke and the hub markets.

Overall then, significant scope economies lead to more concentrated market structures in two ways. First, the scope effects by definition create incentives to consolidate the production of different products and different product versions within one firm rather than many separate firms. Second, strong scope economies also give rise to important multiproduct scale economies, leading firms to be large relative to market size.[9]

4.4 LEARNING-BY-DOING AND EXPERIENCE CURVES

Scale economies reflect the lower unit cost that comes from operating on a larger scale as defined by the volume of production per unit of time. If there are significant scale economies in automobile production, then General Motors will have a lower unit cost in 2013 if it produces say 3.8 million cars that year instead of 3.4 million. There is, however, another size factor that may influence a firm's cost. This is its total, cumulative output over its entire

[9] 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, Rhine, and Santos (1989) and DeGroot, McMahon, and Volkwein (1991) for evidence of scope economies in higher education; and Gilligan, Smirlock, and Marshall (1984) and Pulley and Braunstein (1992) for evidence of scope economies in finance.

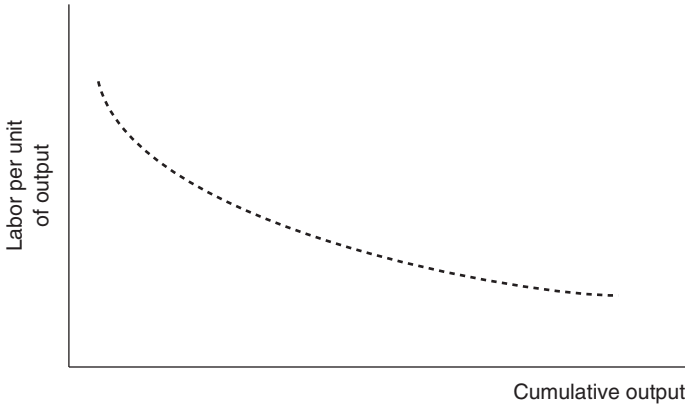

Figure 4.3 Hypothetical experience or learning curve

history. If this factor is important, then a firm that produces only 3.4 million cars might still have relatively low unit cost if it has produced a lot of cars in previous years.

Like any organization, firms and their employees can learn with experience. There are typically many ways of organizing design, production, and sales, and the best ways can only be learned by trial-and-error. As the cumulative experience of the firm grows, more such trials can be made and the insights from that experience can be used to raise productivity in the future. Based on a survey of the then-existing literature, Argote and Epple (1990) find that for the typical US manufacturing firm, labor costs fall by 20 percent as its cumulative output doubles. Ramanarayanan (2006) finds that for cardiac surgeons, each additional coronary artery bypass performed lowers the probability of death for subsequent patients by 0.14 percent. These results imply that cost reductions due to organizational learning are important, as illustrated by Figure 4.3.

Yet if organizations can learn, they can also forget. As older workers retire or quit and institutional memories are lost, firms may move backwards and unit costs may rise. Benkard (2000, 2004) finds evidence of both learning and forgetting in aircraft production. As firms learn and forget, the dynamics of industry structure become complicated.

However, the basic intuition regarding how important learning effects may affect market structure is relatively straightforward. Essentially, learning-by-doing leads to more concentrated market structures because it provides further advantage to first movers or early entrant firms, as these firms will typically have had the longest time to learn. Strategic considerations are likely to reinforce this result. Faced with possible new rivals, an incumbent firm will often find it optimal to price artificially low so as to sell more units. This not only helps the incumbent move down the learning curve itself but also makes it harder for rivals to do the same. Further, the prospect that whatever it learns may be forgotten gives a further incentive for the incumbent to price low and thereby acquire more experiences so that even if it loses some of its knowledge, it still has an advantage. Thus, both in and of itself and because of the strategic responses it induces, learning-by-doing will likely lead to more concentrated markets in which early incumbents maintain and even extend their dominance.[10]

[10] See Besanko, et al, (2010) for an extended examination of these issues.

4.5 NON-COST 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 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.

4.5.1 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 is: 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. 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 (4.5). 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:

$$F = K + \beta(AE) \tag{4.8}$$

where recall that A is a constant (K is also a constant) and E is aggregate consumer expenditure in the industry.

Using (4.8), equation (4.5) now may be written as:

$$N^e = \left[\frac{1}{\left(\frac{K}{AE}\right) + \beta}\right]^{\frac{1}{1+\alpha}} \tag{4.9}$$

Equation (4.9) 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.[11]

Somewhat similarly, our initial derivation of equation (4.5) assumed that price cost margins declined as a result of an increase in the number of firms as described by: $(P - c)/P = A/N^a$. 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.

4.5.2 Network Externalities and Market Structure

It is not news to anyone reading this book that Microsoft's *Windows* has an approximate 85 percent of the operating systems market for desktop and laptop computers with virtually all of the remaining 15 percent going to Apple's *Mac* products.[12] *Microsoft Word* and *Microsoft Excel* have similarly large, dominant market shares of the word-processing and spreadsheet markets, respectively. Scale and scope economies are undoubtedly part of the explanation for the highly concentrated nature of these markets. 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, an additional critical factor behind Microsoft's dominance is the presence of *network externalities*. Such network effects reflect the fact that for some products, a consumer's willingness to pay for a good or service increases as the number of other consumers buying the product rises.

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

[12] www.statowl.com/operating_system_market_share.php.

Network effects can be *direct* or *indirect*. Telephone service exhibits direct network externalities. The usefulness or value of a single consumer connecting to a telephone system is essentially nil. Yet, as more people sign on to the system, the number of potential calls and hence the utility of owning a phone increases as well. Operating systems, platforms such as Amazon's *Kindle*, and credit card services each exhibit more indirect network effects due to complementarities with other products. In the case of operating systems, for example, the significant scale economies that characterize the design and production of applications means that firms marketing computer apps want to make them compatible with as large a network as possible. Likewise, consumers want to use an operating system for which the number of available applications is largest. This establishes a positive feedback loop. As *Windows*' market share grows, more applications firms will want to design their applications to run on the *Windows* operating system. As more applications are written for *Windows*, demand for the *Windows* system rises and with it, *Windows*' market share. Likewise as more books are offered on *Kindle* and as more retailers accept *American Express* cards, the more consumers want to use *Kindle* readers and *Amex* credit cards, which in turn, increases the market shares of these firms.

We address the topic of network externalities more extensively in Chapter 22. 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.

4.5.3 The Role of Government Policy

From 1934 to the mid 1990s, 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 to the benefit of those lucky taxi owners who obtained the first batch of medallions. A court order in the mid-1990s led to an increase of nearly 300 medallions over the subsequent five years. Yet this was not enough to keep the price of medallion from rising dramatically.[13]

A similar phenomenon prevailed from the 1930s through the 1970s when the number of so-called trunk airlines flying interstate routes never exceeded sixteen and fell to ten by the end of the 70s. 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 US domestic airline industry. Indeed, this forty-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 airlinesto enter new city-pair markets. Virtually all

[13] C. Berdik, "Boston's Fare Game," *Boston Magazine*, September 2004: 19.

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, and whether such goals were appropriate, are questions to be answered elsewhere. The central point illustrated by both the taxicab and airline examples 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 that otherwise would have been driven out of the market by the large retail chains. Similarly, the decision of the US government after World War II 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.

4.6 EMPIRICAL APPLICATION: SCALE AND SCOPE ECONOMIES IN BANKING

Because 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. Thus estimating firm cost functions can be a tricky business. However, application of basic microeconomic theory can greatly facilitate the process.

The basic idea is clear enough. A profit-maximizing firm will choose inputs to produce a given output level at the lowest cost given the set of input prices. For single product firms, this will imply that cost is simply a function of its total output and the prices of its inputs (See the Appendix for a formal derivation of this relationship). With a data set in which these variables vary either across firms at a point in time or for the same firm at different times (or both), we can estimate this relationship to measure how total costs vary with output and, hence, to estimate the extent of any scale economies. For instance, with one output Q, and two inputs, capital and labor, whose input prices are r and w, respectively, we might estimate the simple logarithmic relationship below:

$$\ln C = \text{Constant} + \delta_1 \ln r + \delta_2 \ln w + \delta_3 \ln Q \tag{4.10}$$

In this case, the coefficient δ_3 is a direct measure of the elasticity of total cost with respect to output. Therefore, $1/\delta_3$ is a measure of single-product scale economies.

A very simple extension of equation (4.10) to handle a firm with two outputs, q_1 and q_2, we might instead assume a cost function of the form:

$$\ln C = \text{Constant} + \delta_1 \ln\ r + \delta_2 \ln\ w + \delta_3 \ln q_1 + \delta_4 \ln q_2 + \ \delta_5 (\ln q_1)(\ln q_2) \tag{4.11}$$

Here, the final term allows for interaction between the two outputs and thus allows for possible scope economies. A more complicated extension of equation (4.10) that is

potentially consistent with a very large range of underlying production functions is the translog cost function, which in this case is written as:

$$\begin{aligned}\ln C = {} & \text{Constant} + \delta_1 \ln\ r + \delta_2 \ln\ w + \delta_3 \ln q_1 + \delta_4 \ln q_2 + \delta_5 (\ln q_1)(\ln q_2) \\ & + \delta_6 (\ln r)(\ln w) + \delta_7 (\ln r)^2 + \delta_8 (\ln w)^2 + \delta_9 (\ln q_1)^2 + \delta_{10} (\ln q_2)^2 \\ & + \delta_{11} (\ln r)(\ln q_1) + \delta_{12} (\ln w)(\ln q_1) + \delta_{13} (\ln r)(\ln q_2) + \delta_{14} (\ln w)(\ln q_2) \quad (4.12)\end{aligned}$$

Note that equation (4.12) includes equations (4.11) and (4.10) as special cases.

Once the cost function parameters (the δs) have been estimated, it is straightforward, in principle, to derive the scale economy and scope economy measures described in the text. Christensen and Green (1976) use this approach to find important scale economies in electricity generation, while also finding that most power-generating firms are sufficiently large that these scale economies have been fully exploited. 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, no scope economies between graduate education and research.

Pulley and Braunstein (1992) apply this same approach to commercial banking firms with one important modification. Consider again the two-output case. Remember that the measure of scope economies in this case is $\{[C(q_1, 0) + F_1 + C(0, q_2) + F_2] - [C(q_1, q_2) + F]\}/[C(q_1, q_2) + F]$. It may turn out, however, that there are no firms in the sample that produce only one output. Thus, any estimates of scope economies will have to be based on hypothetical cases for which there are no actual counterparts. For this reason, Pulley and Braunstein (1992) introduce a measure of *quasi*-scope economies. To understand this concept, consider the two-variable case. While no firms may actually specialize completely in one product, some may come relatively close by devoting 90 percent of their effort to a single output and 10 percent to the remaining output. We might therefore consider the cost disadvantage that these firms face as: $\{[C(0.9q_1, 0.1q_2) + F_1 + C(0.1q_1, 0.9q_2) + F_2] - [C(q_1, q_2) + F]\}/[C(q_1, q_2) + F]$. More generally, if ε is the fraction devoted to the minor activity, then Pulley and Braunstein (1992) define *quasi*-scope economies in the two-output case as:

$$\begin{aligned}\text{quasi-}S^C = {} & \{[C((1-\varepsilon)q_1,\ \varepsilon q_2) + F_1 + C(\varepsilon q_1, (1-\varepsilon)q_2) + F_2] \\ & - [C(q_1, q_2) + F]\}/[C(q_1, q_2) + F] \quad (4.13)\end{aligned}$$

By using different values of ε, Pulley and Braunstein (1992) can infer a robust estimate of the importance of scope effects in banking.

For this purpose, Pulley and Braunstein (1992) define four banking outputs. These are: 1) demand deposits plus savings and small time deposits, 2) real estate loans, 3) commercial and industrial loans, and 4) installment and credit card loans. Table 4.2 below shows some of the (quasi) scope and scale estimates they obtain for different variations of the underlying cost equation and different values of ε.

Table 4.2 Estimates of (Quasi) scope and scale economies in US commercial banking (standard errors in parentheses)

	Composite Cost Function Quasi-Scope Economies	*Translog Cost Function Quasi-Scope Economies*
$\varepsilon = 0.10$	0.23 (0.06)	0.18 (0.28)
$\varepsilon = 0.15$	0.19 (0.05)	0.15 (0.13)
$\varepsilon = 0.20$	0.16 (0.05)	0.15 (0.05)
Scale Economy Estimate	0.04 (0.01)	0.06 (0.02)

The Pulley and Braunstein (1992) estimates indicate that there are sizable and statistically significant advantages of scope in US commercial banking. Combining all four activities in one firm reduces costs by 15 to 23 percent relative to the costs of producing these outputs at four separate and much more specialized banks. Statistically significant but relatively small scale effects are also found. Very large banks save only about 5 percent in costs relative to small ones—not trivial, but perhaps not enough to offset the bad effects of having large banks with a lot of market power—and that may need to be bailed out by the government if they are "too large to fail."

Summary

This chapter has focused on technology, 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. A further cost-related factor is captured by the experience or learning curve that characterizes some industries, such as aircraft manufacturing. In such markets, firms learn valuable, cost-saving techniques as their total production over time rises. This too will tend to result in somewhat concentrated markets as early entrants have a cost advantage over later rivals because these incumbents will be farther down the learning curve.

Other factors influence market structure as well. One of these is market size. Because a large market has room for a number of firms, larger markets tend to be less concentrated than smaller 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.

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.

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 analysis based on the theory of production costs has found significant scope economies between different banking activities but little evidence of significant scale economies. Other studies have found important scale effects in electric power generation and important economies of scope between graduate and undergraduate education.

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, the consultant advises that it would be better to 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 $q \leq 7$

   $C(q) = 7q$ for $q > 7$

   a. Derive average and marginal cost for all integer outputs less than or equal to 7.
   b. What are the 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 question 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 recent article (J. Peder Zane, "It Ain't for the Meat; It's for Lotion," *New York Times*, Sunday May 5, 1996: 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

Alchian, A. A., and Demsetz, H. 1972. "Production, Information Costs, and Economic Organization," *American Economic Review* 62(December): 772–95.

Argote, L., and D. Epple. 1990. "Learning Curves in Manufacturing," *Science* 247 (February 23, 1990): 920–24.

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.

Benkard, L. 2000. "Learning and Forgetting: The Dynamics of Aircraft Production," *American Economic Review* 90 (December): 1034–54.

_____. 2004. "A Dynamic Analysis of the Market for Wide-bodied Commercial Aircraft," *Review of Economic Studies* 71: 581–611.

Bernard, A., S. J. Redding, and P. K. Schott. 2009. "Products and Productivity," *Scandinavian Journal of Economics* 111 (December): 681–709.

Besanko, D., U. Doraszelski, Y. Kryukov, and M. Satterwaite. 2010. "Learning-By-Doing, Organizational Forgetting, and Industry Dynamics," (with Ulrich Doraszelski, Yaroslav Kryukov, and Mark Satterthwaite); *Econometrica*, 78 (March): 453–508.

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.

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):615–623.

Gale, I. 1994. "Price Competition in Noncooperative Joint Ventures," *International Journal of Industrial Organization* 12:53–69.

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.

Grossman, S. J., and , O. D. Hart. (1986), 'The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy* 94: 691–719.

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.

Holmström, B. 1982. "Moral Hazard in Teams," *Bell Journal of Economics* 13 (Autumn): 324–40.

Holmström, B., and P. Milgrom, 1994. "The Firm as an Incentive System," *American Economic Review* 84 (September): 972–91.

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.

Ramanarayanan S. 2006. "Does Practice Make Perfect?: An Empirical Analysis of Learning-By-Doing in Cardiac Surgery." Working Paper, UCLA Anderson School of Management.

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

Smith, Adam. 1776. *The Wealth of Nations*.

Spulber, Daniel. 2009. *The Theory of the Firm: Microeconomics with Endogenous Entrepreneurs, Firms, Markets, and Organizations*. Cambridge: Cambridge University Press.

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

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

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

Appendix

AVERAGE COST, MARGINAL COST, AND COST MINIMIZATION

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

$$\frac{dAC(q)}{dq} = \frac{q\frac{dC(q)}{dq} - [C(q) + F]}{q^2} = \frac{q\left(MC\,(q) - \frac{[C(q)+F]}{q}\right)}{q^2} = \frac{[MC(q) - AC(q)]}{q} \quad (4.A1)$$

The denominator is positive. So, 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. Minimizing average cost curve requires that the AC slope be zero, or that 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:

$$\frac{MP_i}{MP_j} = \frac{w_i}{w_j};\ \text{which is equivalent to}\ \frac{MP_i}{w_i} = \frac{MP_j}{w_j} \quad (4.A2)$$

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 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:

$$\eta_C = \frac{d(C(q) + F)}{C(q) + F} \Big/ \frac{dq}{q} = \left(\frac{dC\,(q) + F}{dq}\right)\left(\frac{q}{C\,(q) + F}\right) = \frac{MC(q)}{AC(q)} \quad (4.A3)$$