Chapter 1

TECHNOLOGICAL DETERMINANTS OF FIRM AND INDUSTRY STRUCTURE

JOHN C. PANZAR*

Northwestern University

Contents

I would like to thank Avner Greif for his research assistance, Bobby Willig for helpful comments, and the National Science Foundation, SES 8409171, for partial research support. Any errors are, of course, solely my responsibility.

Handbook of Industrial Organization, Volume I, Edited by R. Schmalensee and R.D. Willig

1. Introduction

The title of this volume is the *Handbook of Industrial Organization*. The literal interpretation of the term "industrial organization" has, in large part, receded from the surface when the noun or adjective "IO" is used. As many of the subsequent chapters in this volume indicate, the field has moved far beyond the mere description of how industries are organized. Yet it is at this basic level that the discussion must begin. For the very name of the field alerts one to the fact that we are dealing with questions that do not even arise in the traditional Marshallian framework. There the industry, itself, was the unit of analysis. Its internal organization, while perhaps of anecdotal interest, was not viewed as being at all important for answering the important positive or normative questions of value theory. Thus, the distinguishing feature of research in industrial organization is that, for some reason or other, it is not fruitful to employ the classical perfectly competitive model to analyze the problems of interest. This chapter explores the technological conditions that may make it necessary to abandon the competitive model: there simply may not be "enough room" in the market for a sufficiently large number of firms to give credence to the assumption of price-taking behavior.

The chapter is organized in the following manner. Section 2 introduces the cost concepts required for analyzing the role of technology in the determination of firm and industry structure. The emphasis is on the general multiproduct case, although important single product aspects of the problem are also discussed. Section 3 presents an analysis of the role these cost concepts play in determining efficient industry structure. Section 4 addresses some issues that must be dealt with in any empirical study of technology and industry structure, as well as presenting selective surveys of such studies of the telecommunications and electric power industries. Section 5 ends the chapter with some concluding observations.

2. The multiproduct cost function[1]

The most basic concept with which to characterize the productive technology available to the firm is the technology set T, a list of the combinations of inputs and outputs that are available to the firm. Thus, let x denote a vector of r inputs

[1]The material in this section is based upon the discussion in Baumol, Panzar and Willig (1982). Most formal proofs have been included in order to make this discussion of important multiproduct cost concepts as self-contained as possible.

available to the firm and y a vector of possible outputs that may be selected from the set $N = \{1, 2, \ldots, n\}$. Then the technology set is formally defined as

Definition 1. The technology set

$T = \{(x, y)$: y can be produced from $x\}$.

In the familiar single output case, T can be directly related to the simple production function $y = f(x)$. Assuming free disposal, the technology set can be characterized as $T = \{(x, y)$: $y \leq f(x)\}$. While this definition of T is intuitively quite clear, more structure must be assumed in order to facilitate mathematical analysis. The following weak regularity condition is commonly employed:

Regularity condition R1

Input vectors x are elements of the compact set $X \subset R^r_+$ and output vectors y are elements of the compact set $Y \subset R^n_+$. The technology set T is a nonempty closed subset of $X \times Y$, with the additional properties that (i) $(\mathbf{0}, y) \in T$ iff $y = \mathbf{0}$, and (ii) If $(x, y) \in T$, $(x^1, y^1) \in X \times Y$, $x^1 \geq x$, and $y^1 \leq y$, then $(x^1, y^1) \in T$.

*R*1(i) states that positive inputs are required to produce positive outputs. *R*1(ii) is a "free disposal" axiom that assures that the production process is at least weakly monotonic, i.e. an increase in input use makes possible at least a weak increase in output.

Given *R*1, there exists a continuous production transformation function $\varphi(x, y)$ that is nondecreasing in x and nonincreasing in y such that $\varphi(x, y) \geq 0$ iff $(x, y) \in T$.[2] The production transformation function provides a convenient functional representation of the set of feasible input/output combinations. It is directly related to the familiar single output production function. For example, if $y = f(x)$ is the production function, then $\varphi(x, y) = f(x) - y$ is a well-defined production transformation function.

Since most of the analysis of this chapter will be carried out under the assumption that the firms in the industry are price takers in input markets, it is more convenient to work with the cost function representation of the technology. Therefore define the multiproduct minimum cost function:

$$C(y, w) = \min_x \{w \cdot x: (x, y) \in T\} = w \cdot x^*(y, w),$$

[2]See McFadden (1978).

where $x^*(y, w)$ is an efficient, cost-minimizing input vector for producing the output vector y when factor prices are given by w.

It will be convenient (and sometimes essential) to assume that this central analytic construct has the following smoothness property:

Regularity condition R2

For all $i \in N$, if $y_i > 0$, then $C_i \equiv \partial C/\partial y_i$ exists.

This simply assumes that marginal cost is well defined for any output that is produced in strictly positive quantity. It is not desirable to assume that the cost function is globally differentiable, because that would rule out the possibility that additional fixed or startup costs may occur when production of another output begins. (Mathematically, such possibilities would require the presence of jump discontinuities along the various axes.) At this point it is also appropriate to introduce a regularity condition defined on the transformation function $\varphi(x, y)$ that suffices for $R2$:

Regularity condition R3

T can be characterized by a transformation function, $\varphi(x, y)$, that is continuously differentiable in x and in y_i, for $y_i > 0$, at points (x, y) where x is cost-efficient for y.

A particularly convenient and reasonably general[3] specification of a cost function satisfying $R2$ is as follows. Let $C(y) = F\{S\} + c(y)$, where c is continuously differentiable, $c(0) = 0$, $S = \{i \in N: y_i > 0\}$ and $F\{\varnothing\} = 0$.[4] A simple two-product example will serve to illustrate the usefulness of this construction:

$$C(y_1, y_2) = \begin{cases} F^{12} + c_1 y_1 + c_2 y_2, & \text{for } y_1 > 0,\ y_2 > 0, \\ F^{1} + c_1 y_1, & \text{for } y_1 > 0,\ y_2 = 0, \\ F^{2} + c_2 y_2, & \text{for } y_1 = 0,\ y_2 > 0, \end{cases} \tag{1}$$

[3]This formulation is not completely general, however. In particular, it does not allow for the possibility that the magnitude of the jump discontinuity that result when a new product is introduced may depend upon the quantities of the outputs in the existing product mix as well as the composition of that mix.

[4]Here, and wherever it will not lead to confusion, the vector w will be suppressed as an argument of the cost function.

with $0 < F^1, F^2 < F^{12}$ and $C(0,0) = 0$. Thus, in this example, starting up production of y_1 only requires incurring the fixed cost F^1. If the firm then begins production of y_2 as well, additional fixed costs of $F^{12} - F^1$ are incurred. The important role in the determination of industry structure played by such product specific fixed costs will be discussed below.

2.1. Economies of scale

The technological limits to competitive market structures have long been attributed to the presence of economies of scale.[5] Later we shall discuss its importance in the determination of firm and industry structure in great detail. But what, precisely, is meant by the term "economies of scale"? The most natural intuitive characterization in the single product case is: Given a proportional increase in all input levels, does output increase more or less than proportionately? While this technological definition is often used in undergraduate textbooks,[6] it is not useful for current purposes, because it does not bear the desired relationship to the properties of the firm's cost curves.

To understand why, suppose that in a small neighborhood of the output level $y = f(\boldsymbol{x})$, it is the case that $f(k\boldsymbol{x}) = k'y$. Then the above definition would say that returns to scale were increasing, constant, or decreasing as k' is greater than, equal to, or less than k. It is easy to see that increasing returns to scale, by this definition, implies that average costs are lower at $k'y$ than they are at y.[7] However, the converse is not necessarily true. The reason is that if the firm wishes to increase output by the factor k', the cheapest way to do so is not necessarily a proportionate increase in all input levels. Thus, even if per unit expenditure does not fall when output is increased by expanding all inputs proportionately (i.e. $k > k'$), it may decrease when inputs are chosen in a cost-minimizing manner. Part of the conceptual difficulty is due to the need to relate economies of scale concepts, defined in terms of properties of the productive technology without reference to factor prices, to the cost conditions facing the firm. For, as we shall see, it is the latter that play a key role in determining firm and industry structure.

Fortunately, assuming that regularity condition $R3$ holds, it is possible to define a technologically based measure of the degree of scale economies that also serves to characterize important properties of firms' cost functions.

[5]See Scherer (1980, ch. 4) for an extended factual and intuitive discussion of the sources of and limits to economies of scale in a manufacturing setting.

[6]See, for example, Hirshleifer (1984, p. 329).

[7]*Proof*: $AC(k'y, \boldsymbol{w}) \leq k\boldsymbol{w} \cdot \boldsymbol{x}/k'y = (k/k')\boldsymbol{w} \cdot \boldsymbol{x}/y = (k/k')AC(y, \boldsymbol{w}) < AC(y, \boldsymbol{w})$.

Definition 2. Technological economies of scale

The degree of technological scale economies at $(x, y) \in T$ is defined as $\tilde{S}(x, y) = -\{\sum x_i(\partial\varphi/\partial x_i)\}/\{\sum y_i(\partial\varphi/\partial y_i\}$. Returns to scale are said to be (locally) increasing, constant or decreasing as $\tilde{S}$ is greater than, equal to or less than 1.

For the single output case, this definition reduces to the familiar concept known as the elasticity of scale.[8] That concept is defined as the elasticity with respect to t of $f(tx)$, evaluated at $t = 1$. That is,

$$e(tx) \equiv t[\mathrm{d}f(tx)/\mathrm{d}t]/f(tx) = t\left[\sum x_i f_i(tx)\right]/f(tx),$$

where $f_i \equiv \partial f/\partial x_i$. Thus, $e(x) = \sum x_i f_i(x)/f(x)$. To see that this is exactly equal to $\tilde{S}$, note that in the single output case, $\varphi(x, y) = f(x) - y$. Thus, $\partial\varphi/\partial x_i = \partial f/\partial x_i$ and $\partial\varphi/\partial y = -1$. Substituting into Definition 1 yields the result that $\tilde{S} = \sum x_i f_i/y = \sum x_i f_i/f = e(x)$.

Now consider an alternative measure of economies of scale that is defined in terms of the firm's cost function:

Definition 3. Cost function economies of scale

The degree of cost function scale economies enjoyed by the firm at any output vector y when facing factor prices w is defined as $S(y, w) = C(y, w)/[\sum y_i C_i(y, w)]$. Again, returns to scale are said to be (locally) increasing, constant or decreasing as S is greater than, equal to or less than 1.

Of course, for the single product case, S reduces to $C/yC' = AC/MC$, the ratio of average cost to marginal cost. Since $\mathrm{d}AC/\mathrm{d}y = (MC - AC)/y$, this means that the firm enjoys increasing, constant or decreasing returns to scale as the derivative of average cost with respect to output is negative, zero or positive.

Note that it is not quite correct to replace this characterization with one that determines the presence or absence of scale economies based upon whether AC is decreasing or increasing. Consider, for example, the cost function $C(y) = y[2 - (y - 1)^3]$. At $y = 1$, $MC = AC = 2$ and $S(1) = 1$. However, $AC = [2 - (y - 1)^3]$ is clearly decreasing at $y = 1$, since $AC(1 + \varepsilon) = 2 - \varepsilon^3 < AC(1) = 2 < 2 + \varepsilon^3 = AC(1 - \varepsilon)$ for any small, positive ε. This establishes

Proposition 1

Locally, economies of scale are sufficient but not necessary for the firm's average cost curve to be declining in the single output case.

[8]See, for example, Varian (1984, ch. 1) and Ferguson (1969).

At first, it is difficult to see the connection between the technology based definition of scale economies $\tilde{S}(x, y)$ and the cost based definition $S(y, w)$. Both are local concepts, possibly taking on different values at every point in their domain. Because $\tilde{S}$ is a property of a point in input/output space, while S is a property of the cost function, it is not obvious that they are closely related. In fact, however, they are equivalent!

Proposition 2

Given $R1$ and $R3$, $\tilde{S}(x^*(y, w), y) = S(y, w)$, i.e. when outputs are produced in a cost efficient manner, the degree of scale economies is the same, whether it is measured using the transformation function or the cost function.

Proof

The cost function results from minimizing $w \cdot x$ subject to $x \geq \mathbf{0}$ and $\varphi(x, y) \geq 0$. Letting λ denote the value of the Lagrange multiplier at the optimum, the Kuhn–Tucker necessary conditions for this problem are

$$w_i - \lambda \partial \varphi / \partial x_i \geq 0,$$

$$x_i^* [w_i - \lambda \partial \varphi / \partial x_i] = 0, \tag{2}$$

$$\varphi(x^*, y) \geq 0, \quad \lambda \geq 0 \quad \text{and} \quad \lambda \varphi(x^*, y) = 0. \tag{3}$$

Summing (2) over all inputs and using the fact that $C(y, w) \equiv w \cdot x^*(y, w)$, yields:

$$C(y, w) = \lambda \sum x_i^* \partial \varphi / \partial x_i. \tag{4}$$

The Lagrangian expression for this problem evaluated at the optimum is given by

$$C(y, w) \equiv w \cdot x^* - \lambda \varphi(x^*, y).$$

Thus, from the Envelope Theorem, we have:

$$C_j = -\lambda \partial \varphi / \partial y_j. \tag{5}$$

Multiplying (5) through by y_j and summing over all outputs, using (4), yields:

$$S(y, w) = -\left[\lambda \sum x_i^* \partial \varphi / \partial x_i\right] / \left[\lambda \sum y_j \partial \varphi / \partial y_j\right] = \tilde{S}(x^*(y, w), y),$$

as long as $\lambda \neq 0$. But $\lambda = 0$ and (2) would imply $x^* = \mathbf{0}$, which, given (3), would violate $R1$. Q.E.D.

Thus, we have succeeded in developing a technologically based measure of the degree of scale economies that can be directly related to properties of the firm's multiproduct cost function.

In the subsequent discussion, I shall often adopt the assumption that the cost function exhibits increasing returns to scale at small output levels that are eventually exhausted, followed by a region of decreasing returns as output levels become ever larger. In the single product case, this just the traditional assumption that average cost curves are U-shaped, although the flat and rising portions of the U may lie beyond the range of experience. In this case, it is also well understood that the output level at which average cost attains its minimum plays a particularly important role in the determination of industry structure. However, in the multiproduct case, it is not clear what will fulfill this role, for average cost is not a clearly defined concept when the firm produces more than one product.

Fortunately, it is not average cost itself that plays this crucial role, but rather that the shape of the AC curve indicates the output level at which economies of scale become exhausted. This concept does translate directly to the multiproduct world, and the easiest way to make this clear is to reduce n dimensions down to one by fixing the proportions in which the various outputs of the firm are produced. It is then possible to study the behavior of costs as a function of the (scalar) number of such fixed proportion bundles. Geometrically, this is equivalent to studying the behavior of total costs as production is varied along a ray through the origin in output space. Therefore, consider

Definition 4. Ray average cost

The ray average cost of producing the output vector $y \neq \mathbf{0}$, $RAC(y)$ is defined to be $C(y)/a \cdot y$, $a > \mathbf{0}$. Ray average cost is said to be increasing (decreasing) at y if $RAC(ty)$ is an increasing (decreasing) function of the scalar t, at $t = 1$. Ray average cost is said to be minimized at y if $RAC(y) < RAC(ty)$ for all positive $t \neq 1$.

The vector of positive weights a in this definition is, of course, completely arbitrary, as are the units in which each output level is measured. However, this somewhat artificial construction of a denominator for this multiproduct average cost construct makes it possible to formally relate the slope of the RAC curve at a point in output space to the degree of scale economies experienced by the firm, just as in the single product case.

Proposition 3

The derivative with respect to t of $RAC(ty)$, evaluated at $t = 1$, is negative, zero or positive as the degree of scale economies $S(y)$ is greater than, equal to, or less than 1.

Proof

Since $RAC(ty) = C(ty)/a \cdot (ty)$, $\mathrm{d}[RAC(ty)]/\mathrm{d}t = [t(a \cdot y)(y \cdot \nabla C(ty) - (a \cdot y)C(ty)]/t^2(a \cdot y)^2 = [1 - S(ty)]/[t^2(a \cdot y)(y \cdot \nabla C(ty)]$. Therefore, when $t = 1$, $\mathrm{sign}\{\mathrm{d}RAC/\mathrm{d}t\} = -\mathrm{sign}\{S(y) - 1\}$. Q.E.D.

Hence, just as in the single product case, the firm enjoys increasing, constant or decreasing returns to scale depending upon whether the derivative of ray average cost with respect to the level of (a fixed bundle of) output is negative, zero or positive.

It is now possible to make precise the above presumption that returns to scale are first increasing, then constant and, eventually, decreasing, i.e. RAC curves are U-shaped. The only complication in the multiproduct case is that the size of the output bundle at which economies of scale are exhausted will tend to vary with the composition of the bundle. Thus, instead of a single point of minimum efficient scale at which scale economies are first exhausted, as in the scalar output case, in higher dimensions there will be a locus (surface, hypersurface) of such points: the M-locus. As depicted in Figure 1.1, the M-locus connects all the

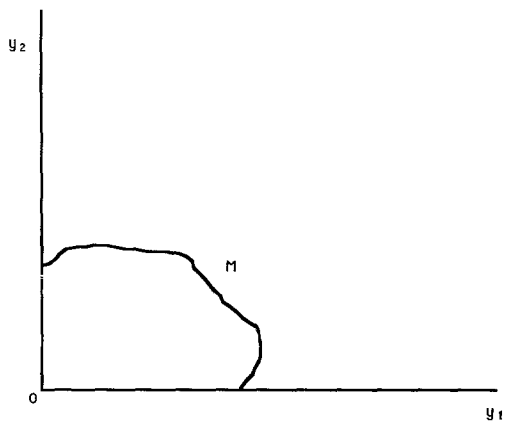

Figure 1.1

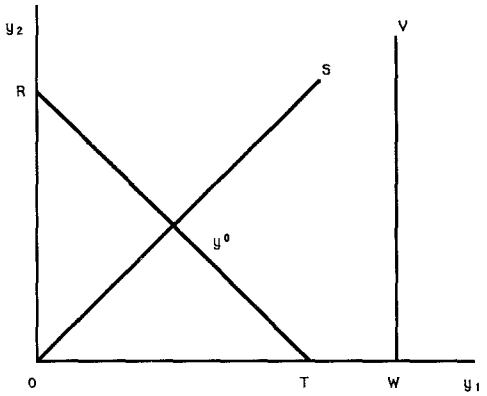

Figure 1.2

minima of the RAC curves corresponding to different output proportions.[9] The points of the M-locus on the axes represent the minimum points of the average cost curves in stand alone production of the various products.

2.2. *Product specific economies of scale*[10]

Our discussion of multiproduct economies of scale revealed that that important property of the technology pertains to the change in costs resulting from proportional variations in output, i.e. as output moves along a ray through the origin. In terms of Figure 1.2, if one envisions the cost surface plotted in the vertical dimension, then economies of scale are characterized by the behavior of RAC as output varies along a ray such as OS. However, when one refers to a firm "increasing" its scale of operations, one might just as easily have in mind an upward movement along WV as an outward movement along OS. That is, the change in costs resulting from a proportional increase in one product (or a subset of products) holding other output levels constant, also has important implications for firm and industry structure.

To begin to discuss this matter requires us to precisely define the incremental cost of product i as the change in the firm's total cost caused by its introduction

[9] If the cost function is twice continuously differentiable, except at the axes, then M will be smooth, if irregular, in the interior of output space.

[10] This term seems to have first been used in Scherer et al. (1975) and Beckenstein (1975) to refer to a concept similar in spirit, but less useful for the present analysis. The discussion here closely follows that in Baumol, Panzar and Willig (1982, ch. 4).

at the level y_i, or, equivalently, the firm's total cost of producing $\mathbf{y}$ minus what that cost would be if the production of good i were discontinued, leaving all other output levels unchanged. More formally, we have

Definition 5. Incremental cost of a single product

The incremental cost of the product $i \in N$ at $\mathbf{y}$ is $IC_i(\mathbf{y}) \equiv C(\mathbf{y}) - C(\mathbf{y}_{\hat{i}})$, where $\hat{i} = \{j \in N: j \neq i\}$, the complement of i in N, and $\mathbf{y}_{\hat{i}}$ is a vector with a zero component in place of y_i and components equal to those of $\mathbf{y}$ for the remaining products. The average incremental cost of product i is defined as $AIC_i(\mathbf{y}) \equiv IC_i(\mathbf{y})/y_i$.

For example, for $y_2 > 0$, the incremental cost of y_1 in the generalized affine cost function of equation (1) above is given by

$$IC_1 = F^{12} - F^2 + c_1 y_1,$$

and the average incremental cost of y_1 by

$$AIC_1 = c_1 + (F^{12} - F^2)/y_1.$$

Contrast these formulae with those that would result if the cost function were given by the simple affine function $C = F + c_1 y_1 + c_2 y_2$ for $(y_1, y_2) \neq (0, 0)$. In that case, for $y_2 > 0$, $IC_1 = c_1 y_1$ and $AIC_1 = c_1$. The difference stems from the fact that there are product specific fixed costs of $F^{12} - F^2$ in the first case and none in the second. (All the fixed costs are incurred as soon as any positive amount of either product is produced.) These product specific fixed costs give rise to decreasing average incremental costs in the first case and constant average incremental costs in the second. By analogy to the single product case, it is natural to describe the former example as one that exhibits increasing product specific returns to scale. More precisely, we have

Definition 6. Scale economies specific to a single product

The degree of scale economies specific to product i at output level $\mathbf{y}$ is given by $S_i(\mathbf{y}) = IC_1(\mathbf{y})/y_i C_i(\mathbf{y}) = AIC_i/MC_i$. Returns to the scale of product i at $\mathbf{y}$ are said to be increasing, decreasing or constant as $S_i(\mathbf{y})$ is greater than, less than, or equal to 1.

Since it is quite possible to envision a proportional expansion of a proper subset of the firm's products (i.e. more than 1 but less than n), it is useful to

generalize Definitions 5 and 6 to describe the properties of the cost function in such cases.

Definition 7. Incremental cost

The incremental cost of the product set $T \subset N$ at y is given by $IC_T(y) = C(y) - C(y_{\hat{T}})$. Again, $\hat{T}$ is the complement of T in N and $y_{\hat{T}}$ is that vector with components equal to y for products in the set $\hat{T}$ and zero for products in the set T.

Using the same technique as for $RAC(y)$, it is possible to unambiguously define the average incremental costs of a product set:

Definition 8. Average incremental cost

The average incremental cost of the product set T at y is $AIC_T(y) \equiv IC_T(y)/a \cdot y_T$. The average incremental cost of the product set T is said to be decreasing (increasing) at y if $AIC_T(ty_T + y_{\hat{T}})$ is a decreasing (increasing) function of t at $t = 1$.

We can now define a measure of the degree of product set specific scale economies that is consistent with both the scalar and multiproduct measures developed so far.

Definition 9. Product specific economies of scale

The degree of scale economies specific to the product set $T \subset N$ at y is given by $S_T(y) \equiv IC_T(y)/y_T \cdot \nabla C(y)$. Product set specific economies of scale are said to be increasing, decreasing or constant at y as $S_T(y)$ is greater than, less than, or equal to 1.

This definition is identical to $S(y)$ when $T = N$ and equals the product specific measure of Definition 6 when $T = \{i\}$. Also, the same arguments employed to establish Proposition 3, can be used to establish that the sign of $\mathrm{d}AIC_T(ty)/\mathrm{d}t$, evaluated at $t = 1$, is the same as the sign of $1 - S_T(y)$. Thus, just as in the scalar and n product cases, the degree of economies of scale can be defined in terms of the derivative of the (appropriately defined) average cost curve. Note also that $S_T(y) > 1$ implies $DAIC_T(y)$ (decreasing average incremental costs of the product set T at y), but, as in the scalar case, not conversely.

Having explored the concept of product specific economies of scale it is interesting to examine the relationship between the overall degree of scale economies $S(y)$ and the degree of scale economies that pertain to a subset of

products T and its complement $\hat{T}$. Using Definitions 7 and 9 yields:

$$S = \{\alpha_T S_T + (1 - \alpha_T) S_{\hat{T}}\} / \{(IC_T + IC_{\hat{T}})/C\}, \tag{6}$$

where $\alpha_T = y_T \cdot \nabla C / y \cdot \nabla C$. If the denominator of this expression were 1, then the overall degree of scale economies would be a simple weighted average of that of any subset of products and its complement. Indeed, if the production processes used in producing T and $\hat{T}$ were completely separable, that denominator would be 1. Substituting using the definition of incremental cost allows us to write the denominator of (6) as

$$[C(y) - C(y_{\hat{T}}) + C(y) - C(y_T)]/C(y)$$

or

$$1 + [C(y) - C(y_{\hat{T}}) - C(y_T)]/C(y). \tag{7}$$

If the production processes for product sets T and $\hat{T}$ were truly independent, then the total costs of producing all n products would be exactly equal to the sum of the stand-alone costs of the subsets T and $\hat{T}$ [i.e. $C(y_T)$ and $C(y_{\hat{T}})$]. However, if economies of joint production are present, total costs will be less than the sum of the stand-alone costs. Then (7) will be less than 1, and the overall degree of scale economies will exceed the weighted sum of the two product specific measures. The next section discusses these *economies of scope* in detail.

2.3. *Economies of scope*

The multiproduct cost constructs discussed in the previous sections have described the behavior of the cost surface over conveniently chosen cross sections of output space. This section discusses a cost concept that is crucial to our understanding of firm and industry structure, yet cannot be characterized directly in terms of such a "slice" of the cost surface.

In addition to the intuitively familiar economies deriving from the shear size or scale of a firm's operations, cost savings may also result from the production of several different outputs in one firm rather than each being produced in its own specialized firm. That is, the *scope* of the firm's operations may give rise to economies as well. More formally, consider

Definition 10. Economies of scope

Let $P = \{T_1, \ldots, T_m\}$ denote a nontrivial partition of $S \subseteq N$. That is, $\cup T_i = S$, $T_i \cap T_j = \varnothing$ for $i \neq j$, $T_i \neq \varnothing$, and $m > 1$. Then there are economies of scope at

y_S with respect to the partition P if $\sum_i [C(y_{Ti})] > C(y_S)$. There are said to be weak economies of scope if this inequality is weak rather than strict, and diseconomies of scope if the inequality is reversed.

For example, in the simplest two product case, $N = \{1, 2\}$ and $P = \{1, 2\}$. Then economies of scope are present at the output vector (y_1, y_2) if $C(y_1, y_2) < C(y_1, 0) + C(0, y_2)$. In the generalized affine example of equation (1), there are economies of scope if and only if $F^{12} < F^1 + F^2$.

In order to study the relationship between economies of scope and the measures of economies of scale derived above, the following quantitative description is useful:

Definition 11. Degree of scope economies

The degree of economies of scope at y relative to the product set T and the partition $P = \{T, \hat{T}\}$ is defined as $SC_T(y) \equiv [C(y_T) + C(y_{\hat{T}}) - C(y)]/C(y)$.

The degree of economies of scope measures the percentage increase in cost that would result from dividing the production of y into product lines T and $\hat{T}$. Breaking up the firm along these lines increase, decreases, or leaves total costs unchanged as SC_T is greater than, less than, or equal to zero.

If all products have positive incremental costs, it is easy to show that the degree of economies of scope must be less than 1 for any binary partition. Rearranging terms, Definition 11 can be rewritten as

$$SC_T(y) = 1 - [IC_T(y) + IC_{\hat{T}}(y)]/C(y) < 1. \tag{8}$$

Equation (8) allows us to examine the role of economies of scope in relating the degrees of product specific and overall scale economies. Using (8), equation (6) can be rewritten as

$$S(y) = [\alpha_T S_T + (1 - \alpha_T) S_{\hat{T}}]/[1 - SC_T(y)].$$

Thus, it is the presence of economies of scope that "magnifies" the extent of overall economies of scale beyond what would result from a simple weighted sum of product specific levels.

As mentioned briefly above, the literature abounds with discussions of the technological, "engineering" sources of economies of scale. Since economies of scope has a much briefer life as a precise analytic construct,[11] it is desirable to

[11]The term "economies of scope" was introduced and precisely defined in Panzar and Willig (1975).

spend some time describing, in intuitive terms, the properties of the productive technology that give rise to its presence as a property of the multiproduct cost function. The natural place to begin the search for sources of economies of scope is the Marshallian notion of joint production. Intuitively, it must clearly be cheaper to produce pairs of items such as wheat and straw, wool and mutton, and beef and hides in one firm than in two specialized firms. Therefore, I shall construct a formal model of technological joint production and derive its relationship between that concept and economies of scope.

Joint production, in the Marshallian sense, arises because one or more factors of production are public inputs. That is, once acquired for use in producing one good, they are costlessly available for use in the production of others.[12] Assume that there are n production processes:

$$y_i = f_i(z^i, K), \quad i = 1, \ldots, n,$$

where z^i is a vector of inputs that are directly attributable to the production of products i and K is the amount available of the pure public input. It is more convenient to work with the variable cost representation of the productive technology which expresses the minimized level of attributable costs, V^i, of producing product i as a function of y_i, K and the vector w of the prices of the private inputs. That is,

$$V^i(y_i, K, w) = \min\{z^i \cdot w: f(z^i, K) \leq y_i\}, \quad i = 1, \ldots, n.$$

Assuming that the public input is at least weakly productive, it must be the case that

$$V^i(y_i, K_1) \leq V^i(y_i, K_2), \quad \text{for } K_2 \leq K_1, \; i = 1, \ldots, n. \tag{9}$$

If, in addition, the public input is strictly productive in the weak sense that any positive amount of the public input is better than none at all, then it is also true that

$$V^i(y_i, K) < V^i(y_i, 0), \quad \text{for all } y_i, \; K > 0, \; i = 1, \ldots, n. \tag{10}$$

Finally, assume for simplicity that units of the public input are available at the constant price β. Then we can state the following result:

[12] The clearest examples are to be found in the peak load pricing literature: see, for example, Clark (1923), Demsetz (1973) and Panzar (1976). In Marshall's agricultural examples, the plant or animal in question can be viewed as the public input.

Proposition 4

The multiproduct minimum cost function $C(y, w, \beta)$ that is dual to a set of multiproduct production techniques employing a public input (as described above) exhibits economies of scope.

Proof

A firm that produces at minimum cost any subset of products Tj at output levels y_{Tj} solves the program: $\min_k\{\sum_{i\in Tj}[V^i(y_i, K) + \beta K]\}$. Let $\tilde{K}_{Tj}$ solve this program. Then the multiproduct minimum cost function has the property that

$$C(y_{Tj}, w, \beta) = \sum_{i\in Tj}\left[V^i\left(y_i, w, \tilde{K}_{Tj}\right)\right] + \beta\tilde{K}_{Tj}.$$

Now let $\{T_1, \ldots, T_k\}$ constitute a nontrivial partition of N and define the feasible cost function:

$$\bar{C}(y, w, \beta) = \sum_{j=1}^{k}\sum_{i\in Tj} V^i\left(y_i, w, \bar{K}\right) + \beta\bar{K},$$

where $\bar{K} = \max_j\{\tilde{K}_{Tj}\}$, $j = 1, \ldots, k$. Then

$$\bar{C}(y) - \sum_j C(y_{Tj}) = \sum_j\sum_{i\in Tj}\left[V^i\left(y_i, \bar{K}\right) - V^i\left(y_i, \tilde{K}_{Tj}\right)\right] + \beta\left[\bar{K} - \sum_j\tilde{K}_{Tj}\right].$$

Given (9) and (10), both terms on the right-hand side of this expression are nonpositive, with at least one strictly negative. Because $C(y)$ is defined to be the minimum cost function, we know that $C(y) \leq \bar{C}(y)$. Therefore,

$$C(y) - \sum_j C(y_{Tj}) \leq \bar{C}(y) - \sum_j C(y_{Tj}) < 0. \tag{11}$$

But for $y > \mathbf{0}$, the inequality in (11) is precisely the definition of economies of scope.

Q.E.D.

The public input model analyzed above illustrates one technological source of economies of scope. Proposition 4 has demonstrated that the presence of a public

input is sufficient for the existence of economies of scope. However, it is far from necessary. This is fortunate, since cases of joint production involving pure public inputs do not seem numerous enough to account for the ubiquity of multiproduct firms that presumably enjoy economies of scope. There is another tradition in the literature[13] that explains the existence of economies of scope as a result of the presence of inputs that, perhaps because of indivisibilities, are easily shared by the production processes of several different outputs.

In order to investigate this type of phenomenon more precisely, consider a micro model of the sharing of "overhead" between n otherwise independent production processes. For ease of exposition, assume that there is only one such input, called "capital". Let $\psi(\boldsymbol{k}, \boldsymbol{\beta})$ denote the cost of acquiring the vector $\boldsymbol{k} = (k_1, k_2, \ldots, k_n)$ of capital services used in production processes 1 through n when the relevant input prices are $\boldsymbol{\beta}$. If ψ is strictly subadditive in $\boldsymbol{k}$ (i.e. $\psi(\boldsymbol{k}^0 + \boldsymbol{k}^1, \boldsymbol{\beta}) < \psi(\boldsymbol{k}^0, \boldsymbol{\beta}) + \psi(\boldsymbol{k}^1, \boldsymbol{\beta})$), then it is natural to describe k as a quasipublic input, since its services can be shared by two or more product lines at a lower total cost than would be incurred if each obtained its capital services independently. An extreme example is the pure public input case considered above, in which $\psi(\boldsymbol{k}, \beta) = \beta \max_i[k_i]$. Another benchmark case is when capital is a pure private input obtainable at a constant price per unit, so that $\psi(\boldsymbol{k}, \beta) = \beta \sum k_i$ is only weakly subadditive in k. It is also possible to envision situations in which the production processes impede one another, so that ψ is actually superadditive in $\boldsymbol{k}$, e.g. $\psi(k_1, k_2, \beta) = a(\beta)(k_1 + k_2)^2$.

Perhaps more common might be a situation in which capital services are private inputs in the sense that a given total capacity K can be exhaustively allocated across product lines (e.g. $k_1 + k_2 \leq K$), but there are regions of increasing and decreasing returns to scale in the installation of K. For example, consider the case in which $\psi(k_1, k_2, \boldsymbol{\beta}) = \Phi(k_1 + k_2, \boldsymbol{\beta}) = a_0(\boldsymbol{\beta}) + a_1(\boldsymbol{\beta})(k_1 + k_2)^2 = a_0(\boldsymbol{\beta}) + a_1(\boldsymbol{\beta})K^2$. Here, it can be shown that Φ and ψ are subadditive for $k_1 + k_2 = K \leq \sqrt{2a_0/a_1}$.

Thus, economies of shared inputs (subadditivity of ψ) may arise either because the input in question is public or quasipublic or because there are economies of scale in its acquistion. In any event, there is an intimate connection between the acquisition cost properties of shared inputs and the presence or absence of economies of scope at the final output level:

[13]Consider, for example, Hicks (1935/1952, p. 372): "... almost every firm does produce a considerable range of different products. It does so largely because there are economies to be got from producing them together, and these economies consist largely in the fact that the different products require much the same overhead." See also, Clemens (1950/1958, p. 263): "It is a commonplace of business practice that the production and sales managers work hand in hand to devise new products that can be produced with the company's idle capacity... . What the firm has to sell is not a product, or even a line of products, but rather its capacity to produce."

Proposition 5

For any nontrivial partition of N, there are economies (diseconomies) of scope if and only if ψ is strictly subadditive (superadditive) in the relevant range.

Proof

The multiproduct minimum cost function associated with the above micro model of the technology is given by

$$C(y_S, w, \beta) = \min_{k} \left\{ \sum_{i \in S} V^i(y_i, w, k_i) + \psi(k, \beta) \right\}. \tag{12}$$

Let the vector $\tilde{k}(y, w, \beta)$ denote the argmin of program (12) for $S = N$, i.e. the cost minimizing vector of capital services for the production of the output vector y. Assuming that capital services are an essential input into each production process implies that $\tilde{k}_i > 0$ for $y_i > 0$, while, if ψ is nondecreasing, $\tilde{k}_i = 0$ if $y_i = 0$. Now let $\{T_1, \ldots, T_l\}$ be a nontrivial partition of N and let $\tilde{k} = \sum_j [\tilde{k}(y_{Tj})]$, the sum of the optimal capital services vector for each product subgroup if it were produced in isolation. Then from (12) and the definition of $\tilde{k}$, it follows that

$$\sum_{i \in N} V^i\big(y_i, \tilde{k}_i(y)\big) + \psi\big(\tilde{k}(y)\big) = C(y) \leq \sum_{i \in N} V\big(y_i, \hat{k}_i\big) + \psi(\hat{k}) \tag{13}$$

and

$$\sum_{i \in Tj} V^i\big(y_i, \tilde{k}_i(y)\big) + \psi\big(\tilde{k}_{Tj}(y)\big) \geq C(y_{Tj}) = \sum_{i \in Tj} V^i\big(y_i, \tilde{k}_i(y_{Tj})\big) + \psi\big(\tilde{k}(y_{Tj})\big). \tag{14}$$

Summing (14) over $j = 1, \ldots, l$ and subtracting (13) yields:

$$\psi\big(\tilde{k}(y)\big) - \sum_{j=1}^{l} \psi\big(\tilde{k}_{Tj}(y)\big) \leq C(y) - \sum_{j=1}^{l} C(y_{Tj}) \leq \psi(\hat{k}) - \sum_{j=1}^{l} \psi\big(\tilde{k}(y_{Tj})\big). \tag{15}$$

The conclusions follow since the leftmost (rightmost) term in equation (15) is positive (negative) if and only if ψ is strictly superadditive (subadditive) over the relevant range. Q.E.D.

This micro model of the firm's production process establishes the intimate connection between the existence of economies of scope and the presence of inputs that may be effectively shared among production processes. While the focus of the above discussion may have seemed to have been directed exclusively toward technological, engineering considerations, a broader interpretation is certainly possible. For example, the shareable inputs might include managerial expertise, a good financial rating, a sales staff, and so forth.[14]

The foregoing discussion has analyzed the sources of economies of scope at the micro level, in effect, deriving that property of the multiproduct cost function on the basis of assumptions about the way that the firm's production processes interact with one another. While this may provide an intuitive understanding of the factors responsible for economies of scope, it is not terribly useful for empirically testing for their presence. It is difficult to envision obtaining the data that would be required to estimate ψ and evaluating whether or not it is subadditive. Therefore, it is useful to have available a condition defined in terms of properties of the multiproduct cost function that can be used to infer the presence of economies of scope. The following multiproduct cost concept will prove useful in this quest:

Definition 12. Weak cost complementarities

A twice-differentiable multiproduct cost function exhibits weak cost complementarities over the product set N, up to the output level y, if $\partial^2 C(\hat{y})/\partial y_i \partial y_j \equiv C_{ij}(\hat{y}) \leq 0$, $i \neq j$, for all $\mathbf{0} \leq \hat{y} \leq y$, with the inequality strict over a set of output levels of nonzero measure.

The presence of weak cost complementarities implies that the marginal cost of producing any one product does not increase with increases in the quantity of any other product. According to Sakai (1974), this is a normal property of joint production. Note that, because C_{ii} is allowed to be positive, Definition 12 does not impose the strong condition that all of the individual product marginal cost curves C_i are decreasing. The following result is true:

[14]See Teece (1980) for a discussion of such less easily quantifiable sources of economies of scope.

Proposition 6

A twice-differentiable multiproduct cost function that exhibits weak cost complementarities over N up to output level y exhibits economies of scope at y with respect to all partitions of N.

Proof

Since any partition of N can be obtained by a sequence of binary partitions, it suffices to demonstrate the result for the partition $T, \hat{T}$, where $N \neq T = \varnothing$. Rearrange terms so that the condition to be demonstrated is

$$[C(y_T + y_{\hat{T}}) - C(y_T)] - [C(y_{\hat{T}}) - C(\mathbf{0})] < 0.$$

The first term in brackets can be rewritten as $\int_\Gamma \sum_{i \in \hat{T}} [C_i(y_T + x_{\hat{T}})\, dx_i]$ and the second bracketed term as $\int_\Gamma \sum_{i \in \hat{T}} [C_i(x_{\hat{T}})\, dx_i]$, where Γ is any smooth monotonic arc from $\mathbf{0}$ to $y_{\hat{T}}$. Since these are line integrals along the common path Γ, their difference can be written as

$$\int_\Gamma \sum_{i \in \hat{T}} [C_i(y_T + x_{\hat{T}}) - C_i(x_{\hat{T}})]\, dx_i$$

$$= \int_\Gamma \sum_{i \in \hat{T}} \int_\Lambda \sum_{j \in T} C_{ij}(z_T + x_{\hat{T}})\, dz j x_i < 0,$$

where Λ is a smooth monotonic arc from $\mathbf{0}$ to y_T. Q.E.D.

The only problem with this sufficient condition for economies of scope is that it requires that the cost function be twice differentiable everywhere, even at the origin and along the axes. As discussed earlier, this is overly restrictive, since it rules out the presence of overall and product specific fixed costs. Fortunately, the result can be easily extended to deal with this important complication. Without loss of generality, any multiproduct cost function can be expressed as $C(y) = F\{S\} + c(y)$, where $S = \{i \in N: y_i > 0\}$. This formulation allows for $C(\cdot)$ to exhibit discontinuities along the axes even if $c(\cdot)$ is smooth. Thus, Proposition 6 can be generalized to:

Proposition 7

If $c(\cdot)$ is a twice-differentiable function that exhibits weak complementarities over N up to output level y, and if F is not superadditive – i.e. $F\{S\} + F\{T\} \geq F\{S \cup T\}$ for all $S, T \subseteq N$ – then the cost function exhibits economies of scope at $y > \mathbf{0}$ with respect to all partitions of N.

Proof

The proof is the same as that of Proposition 6, with $c(\cdot)$ replacing $C(\cdot)$, so that the above equations also contain the expression $F\{N\} - F\{T\} - F\{\hat{T}\}$. This term is nonpositive by hypothesis. Q.E.D.

Proposition 7 reveals that a multiproduct cost function may exhibit economies of scope because of complementarities in either "fixed" or "variable" components. Clearly, economies of scope may occur even in cases in which $c_{ij} > 0$, as long as $F\{\cdot\}$ is sufficiently subadditive. For example, the cost function introduced in Section 2 exhibits global economies of scope even though $c_{12} = 0$ at all output levels. This follows from the fact that $C(y_1, y_2) - C(y_1, 0) - C(0, y_2) = F^{12} - F^1 - F^2 < 0$ for all $\boldsymbol{y} \neq \boldsymbol{0}$. Similarly, suppose $C(y_1, y_2) = F + a(y_1 + y_2)^2$ for $\boldsymbol{y} \neq \boldsymbol{0}$. This cost function never exhibits cost complementarities, as $C_{12} = 2a(y_1 + y_2) > 0$. Yet it exhibits economies of scope for all output vectors such that $y_1 y_2 < F/2a$.[15]

2.4. Cost subadditivity and natural monopoly

There has been a long tradition of government regulation of "monopolies" in the United States, and, recently, a wave of deregulation in industries that were once thought to be characterized by substantial monopoly attributes. Thus, much of the empirical work on firm and industry structure to be discussed below has focused upon trying to determine the extent of "natural monopoly" in various regulated industries. Somewhat surprisingly, until fairly recently there was considerable confusion as to what, precisely, is meant by the term "natural monopoly".[16] Therefore this subsection will provide a precise definition of this important concept and a discussion of the properties of the cost function that ensure its presence.

Definition 13. Strict subadditivity

A cost function $C(\boldsymbol{y})$ is strictly subadditive at $\boldsymbol{y}$ if for any and all output vectors $\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots, \boldsymbol{y}^k$, $\boldsymbol{y}^i \neq \boldsymbol{y}$, $i = 1, \ldots, k$, such that $\sum \boldsymbol{y}^i = \boldsymbol{y}$, it is the case that $C(\boldsymbol{y}) < \sum C(\boldsymbol{y}^i)$.

[15] *Proof*: Economies of scope are present whenever the cost of producing both outputs together, $F + a(y_1 + y_2)^2$, is less than the total cost of producing each product in a separate firm, $[Fay_1^2] + [F + ay_2^2]$, i.e. when $2ay_1 y_2 < F$.

[16] Baumol (1977) provided the first rigorous discussion of this issue in the multiproduct setting. The discussion that follows is based primarily on that in Baumol, Panzar and Willig (1982, ch. 7). See also Sharkey (1982, ch. 4).

Intuitively, then, subadditivity of the cost function at y ensures that that output vector can be produced more cheaply by a single firm than by any group of two or more firms. Thus, subadditivity of the cost function can be taken as the obvious criterion for natural monopoly.

Definition 14. Natural monopoly

An industry is said to be a natural monopoly if the cost function is strictly subadditive over the entire relevant range of outputs. An industry is said to be a natural monopoly through output level y if $C(y')$ is strictly subadditive at all $y' \leq y$.

It is important to note that subadditivity is a local concept in that it refers to a particular point on the cost surface. However, determining whether or not costs are subadditive at any such point requires knowledge of the cost function at all smaller output levels. That is to say, in order to know whether single-firm production of y is or is not cheaper than its production by any combination of smaller firms, one must know the level of cost that would be incurred by any smaller firm, i.e. one must know $C(y^*)$ for every $y^* \leq y$.

In the "familiar" single product case, natural monopoly has been associated with the presence of increasing returns to scale and falling average costs. However, this characterization is imprecise at best and can be seriously misleading. In order to examine this issue, we need

Definition 15. Declining average costs

Average costs are strictly declining at y if there exists a $\delta > 0$ such that $C(y')/y' < C(y'')/y''$ for all y' and y'' with $y - \delta < y'' < y' < y + \delta$. Average costs are said to decline through output y if $C(y')/y' < C(y'')/y''$ for all y' and y'' such that $0 < y'' < y' \leq y$.

In nontechnical discussions, the notion of falling average costs and natural monopoly are often confused. The following result clarifies the relationship between the two for the single output case.

Proposition 8

Decreasing average cost through y implies that the cost function is subadditive at y, but not conversely.

Proof

Let $y^1, \ldots, y^k$ be any nontrivial way of dividing y among two or more firms, so that $\sum y^i = y$ and $y > y^i > 0$. Because average cost is declining and because $y^i < y$, $C(y)/y < C(y^i)/y^i$, so that $(y^i/y)C(y) < C(y^i)$. Summing over i yields $\sum C(y^i) > \sum (y^i/y)C(y) \equiv C(y)$, which is the definition of subadditivity. To prove that the converse is not true requires only a counterexample. Consider a cost function such that $C(y) = a + cy$ for $0 < y < y^0$ and $C(y) = a + b + cy$ for $y \geq y^0$, with $a > b > 0$. This cost function is clearly globally subadditive, since $C(y) \leq a + b + cy < 2a + cy \leq C(y^1) + \cdots + C(y^k)$ for all $k > 1$ and $y^i > 0$ such that $\sum y^i = y$. Yet there is a region in which average costs are increasing, since $AC(y^0 - \delta) = a/(y^0 - \delta) < (a + b)/(y^0 + \delta) = AC(y^0 + \delta)$ for $0 < \delta < y^0$. Q.E.D.

If one wishes to maintain the presumption that all AC curves are, ultimately, U-shaped, then another counterexample is required. (The AC curve in the counterexample in the above proof is falling almost everywhere, with a discontinuous upward jump at y^0.) Consider the cost function given by $C(y) = F + ay^2$ for $y > 0$. It is easy to see that average costs are U-shaped: falling for $0 < y < \sqrt{F/a} \equiv y_m$ and increasing for $y > \sqrt{F/a}$. Yet this cost function remains subadditive through $y = \sqrt{2F/a} \equiv y_s$. To see this, first note that, when there are rising marginal costs, any industry output y is divided in positive portions most cheaply among k different firms if each firm produces the same amount, y/k. Then (minimized) total industry costs for a k firm industry are $kC(y/k) = kF + ay^2/k > F + ay^2$ for all $y < y_s$.

The foregoing discussion has revealed that the relationship between economies of scale and natural monopoly is nontrivial, even in the single product case. In the multiproduct world things are even more complicated. In fact there exists no logical connection between the two concepts in the multiproduct world!

Proposition 9

Economies of scale is neither necessary or sufficient for natural monopoly.

Proof

Non-necessity was proven as part of Proposition 8. To see that economies of scale is not sufficient for natural monopoly, consider the cost function $C(y_1, y_2) = \sqrt{y_1 + y_2}$. This function exhibits economies of scale everywhere, as $S(y_1, y_2) = 2$. Yet it is *super*additive for all $(y_1, y_2) > (0, 0)$, since $C(y_1, 0) + C(0, y_2) = \sqrt{y_1} + \sqrt{y_2} < \sqrt{y_1 + y_2} = C(y_1, y_2)$. Q.E.D.

The proof of Proposition 9 has revealed that one reason that the presence of economies of scale does not suffice for natural monopoly is that economies of scale do *not* imply economies of scope. Economies of scope is clearly necessary for natural monopoly, since one way of viewing its definition is as a requirement that the cost function be subadditive for all *orthogonal* divisions of the output vector y. Clearly, this requirement is subsumed in those of Definition 13. More simply, if single firm production is to be less costly than *any* multifirm alternative, it must involve less costs than those that would result if the firm were split up along product lines.

Therefore it should not be surprising that economies of scope must always be assumed as part of (or be implied by) any set of conditions that are sufficient for subadditivity. What is somewhat surprising is that economies of scale and economies of scope, together, do not imply subadditivity!

Proposition 10

Economies of scale and economies of scope do not suffice for subadditivity.

Proof[17]

Consider the cost function given by

$$C(y_1, y_2) = 10v + 6(x - v) + z + \varepsilon,$$

$$\text{for } (y_1, y_2) \neq (0,0) \text{ and } C(0,0) = 0, \quad (16)$$

where $x \equiv \max[y_1, y_2]$, $v \equiv \min[y_1, y_2]$, $z \equiv \min[v, x - v]$, and ε is an arbitrarily small positive number.[18] This function would be linearly homogeneous (exhibiting globally constant returns to scale) were it not for the presence of the fixed cost $\varepsilon > 0$. Therefore it exhibits increasing returns to scale everywhere. For the case of stand-alone production, $C(y_i) = 6y_i + \varepsilon$, so that

$$C(y_1, 0) + C(0, y_2) = 6y_1 + \varepsilon + 6y_2 + \varepsilon = 6(x + v) + 2\varepsilon.$$

[17]The following counterexample is from Baumol, Panzar and Willig (1982, ch. 7, pp. 173–74 and ch. 9, pp. 249–251). Another can be found in Sharkey (1982, ch. 4, pp. 68–69).

[18]An intuitive interpretation might go as follows. A farmer is in the business of producing "meat" and "fiber". The technologies available to him include raising sheep, raising chickens and growing flax. Raising sheep costs \$10 per animal and yields one unit of meat and one unit of fiber. Raising chickens (woolless sheep) costs \$6 per unit of meat obtained and growing flax (meatless sheep) cost \$6 per unit of fiber obtained. However, since sheep will destroy the flax crop, the farmer must fence in the *smaller* of these operations at a cost of \$1 per unit. When combined with a setup cost of ε, these options give rise to the stated *minimized* cost function.

Subtraction of equation (16) yields:

$$C(y_1, 0) + C(0, y_2) - C(y_1, y_2) = 2v - z + \varepsilon \geq v + \varepsilon > 0,$$

which demonstrates that this cost function exhibits economies of scope everywhere. Without loss of generality, assume that $x = y_2 > y_1 = v$ and consider dividing the production of (y_1, y_2) between two firms with output levels (y_1, y_1) and $(0, y_2 - y_1)$. This division results in total costs of

$$C(y_1, y_1) + C(0, y_2 - y_1) = 10v + \varepsilon + 6(x - v) + \varepsilon.$$

Subtracting this from (16) yields:

$$C(y_1, y_2) - C(y_1, y_1) + C(0, y_2 - y_1) = z - \varepsilon = \min[y_1, y_2 - y_1] - \varepsilon.$$

Since ε can be chosen as small as desired without violating the properties of global economies of scale and scope, it is always possible to choose a positive $\varepsilon < y_2 - y_1$ so that the above expression is positive. Thus, the cost function is not subadditive for $y_1 \neq y_2$.[19] Q.E.D.

In view of this result, it is clear that a stronger set of sufficient conditions is required to guarantee subadditivity of costs. Intuitively, this strengthening can be accomplished in one of two ways. We can strengthen the assumptions concerning the savings achieved as the scale of the firm's operations increases, or we can strengthen the assumptions about the extent of production cost complementarities. First, we consider the former option. The discussion in Subsection 2.3 suggests how to proceed. Instead of requiring only that the cost function exhibit economies of scale with respect to the firm's entire product line, we assume that the cost function exhibit decreasing average incremental costs with respect to *each* product line. That this will, in general be a more stringent requirement can be seen from equation (6), which reveals that it is quite possible for the cost function to exhibit overall economies of scale at y even though there may be decreasing product specific returns to scale for one (or both) of the product lines involved.

Proposition 11

Decreasing average incremental costs through y for each product $i \in N$ and (weak) economies of scope at y imply that the cost function is subadditive at y.

[19] This "vanishing ε" can easily be dispensed with. Without it, the cost function exhibits globally constant returns to scale and economies of scope but is strictly superadditive everywhere except on the diagonal $(y_1 = y_2)$, where it is additive.

Proof

For clarity and notational convenience, the proof presented here will be for the two output case.[20] The key to the argument is the fact that *DAIC* in a product line implies that *that product line* must be monopolized if industry costs are to be minimized. Consider an output vector $\boldsymbol{y} = (y_1, y_2)$ and divide it into two batches, $\hat{\boldsymbol{y}} + \tilde{\boldsymbol{y}} = \boldsymbol{y}$, with both $\hat{y}_1$ and $\tilde{y}_1 > 0$. Then the following lemma is true:

Lemma

If $DAIC_1(\boldsymbol{y})$ holds, then either

$$C(\hat{y}_1 + \tilde{y}_1, \tilde{y}_2) + C(0, \tilde{y}_2) < C(\hat{\boldsymbol{y}}) + C(\tilde{\boldsymbol{y}}) \tag{17}$$

or

$$C(\hat{y}_1 + \tilde{y}_1, \tilde{y}_2) + C(0, \hat{y}_2) < C(\hat{\boldsymbol{y}}) + C(\tilde{\boldsymbol{y}}). \tag{18}$$

To establish the lemma, assume without loss of generality that the average incremental cost of shifting the production of $\tilde{y}_1$ from one firm to the other is no greater than the cost of shifting the production of $\hat{y}_1$, i.e.

$$[C(\hat{y}_1 + \tilde{y}_1, \hat{y}_2) - C(\hat{\boldsymbol{y}})]/\tilde{y}_1 \leq [C(\hat{y}_1 + \tilde{y}_1, \hat{y}_2) - C(\tilde{\boldsymbol{y}})]/\hat{y}_1. \tag{19}$$

From the *DAIC* assumption we have

$$[C(\hat{y}_1 + \tilde{y}_1, \tilde{y}_2) - C(0, \tilde{y}_2)]/(\hat{y}_1 + \tilde{y}_1) < [C(\tilde{\boldsymbol{y}}) - C(0, \tilde{y}_2)]/\tilde{y}_1.$$

Cross-multiplying and adding and subtracting $\tilde{y}_1 C(\tilde{\boldsymbol{y}})$ on the left-hand side yields:

$$[C(\hat{y}_1 + \tilde{y}_1, \tilde{y}_2) - C(0, \tilde{y}_2)]/\hat{y}_1 < [C(\tilde{\boldsymbol{y}}) - C(0, \tilde{y}_2)]/\tilde{y}_1.$$

Along with (19) this implies:

$$C(\hat{y}_1 + \tilde{y}_1, \hat{y}_2) < C(\hat{\boldsymbol{y}}) + C(\tilde{\boldsymbol{y}}) - C(0, \tilde{y}_2),$$

which completes the proof of the lemma. Now to complete the proof of the proposition suppose, without loss of generality, that (17) holds. Now applying the lemma again tells us that consolidating the production of product 2 will also

[20] The proof for the n output case can be found in Baumol, Panzar and Willig (1982, pp. 176–77, 186).

reduce industry costs, i.e. either

$$C(\hat{y}_1 + \tilde{y}_1, \hat{y}_2 + \tilde{y}_2) + C(0,0) < C(\hat{y}) + C(\tilde{y}) \tag{20}$$

or

$$C(\hat{y}_1 + \tilde{y}_1, 0) + C(0, \hat{y}_2 + \tilde{y}_2) < C(\hat{y}) + C(\tilde{y}). \tag{21}$$

If (20) holds, subadditivity is established immediately. If (8) holds, then (weak) economies of scope establishes the result, since, then

$$C(y) \leq C(\hat{y}_1 + \tilde{y}_1, 0) + C(0, \hat{y}_2 + \tilde{y}_2) < C(\hat{y}) + C(\tilde{y}). \quad \text{Q.E.D.}$$

Not only does this result establish sufficient conditions for an industry to be a natural monopoly, the lemma itself provides important information for understanding industry structure. This important result bears restating.

Proposition 12

If the cost function exhibits Declining Average Incremental Costs for product i ($DAIC_i$) through y, then industry cost minimization requires that production of good i be consolidated in a single firm.

It is clear why this result was important in establishing the sufficient conditions for natural monopoly set forth in Proposition 11, since if $DAIC_i$ holds across all products, the addition of economies of scope implies that all product lines must be monopolized together. However, Proposition 12 is more generally applicable, since it establishes a condition that suffices for any single product to be efficiently monopolized, regardless of the overall presence or absence of natural monopoly. As we shall see, this has important implications for public policy toward industry structure in cases in which there economies of scale in one product that shares scope economies with another for which economies of scale are exhausted at relatively small output levels.

Proposition 11 has set forth sufficient conditions for subadditivity of costs based upon economies of scope and a strengthened version of multiproduct economies of scale. Next, consider the alternative response to the problem posed by Proposition 10: maintaining the assumption of multiproduct economies of scale, while strengthening the accompanying cost complementarity condition. To do this requires the following multiproduct cost concept:

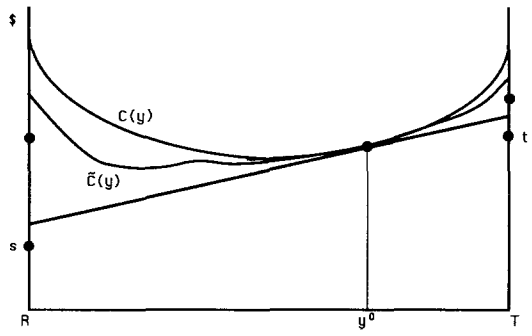

Figure 1.3

Definition 16. Trans-ray supportability

A cost function $C(\boldsymbol{y})$ is trans-ray supportable at $\boldsymbol{y}^0$ if there exists at least one trans-ray direction above which the cost surface is supportable. That is, there is a trans-ray hyperplane $H \equiv \{\boldsymbol{y} \geq \mathbf{0}: \boldsymbol{a} \cdot \boldsymbol{y} = \boldsymbol{a} \cdot \boldsymbol{y}^0\}$, $\boldsymbol{a} > \mathbf{0}$, for which there exists a constant v_0 and a vector $\boldsymbol{v}$ such that $C(\boldsymbol{y}) \geq v_0 + \boldsymbol{v} \cdot \boldsymbol{y}$ for all $\boldsymbol{y} \in H$.

This powerful condition is difficult to interpret intuitively. It can be made clearer with the aid of Figure 1.3, in which the vertical axis measures total cost and the horizontal axis coincides with the base RT of the trans-ray slice in Figure 1.2. Consider a point $\boldsymbol{y}^0$ on this ray. If a straight line can be drawn through $C(\boldsymbol{y}^0)$ that nowhere rises above $C(y)$ over RT, then the cost function C has a support at $\boldsymbol{y}^0$ over the trans-ray hyperplane $H = RT$. Now consider all possible trans-rays through $\boldsymbol{y}^0$ in the y_1, y_2 plane. If the cost function C has such a support above any *one of them*, then it is said to be trans-ray supportable at $\boldsymbol{y}^0$. This brings us to a basic set of sufficient conditions for subadditivity:

Proposition 13

If $C(\boldsymbol{y})$ is trans-ray supportable at $\boldsymbol{y}^0$ and exhibits decreasing ray average costs along all rays through the origin *up to H*, the hyperplane of trans-ray supportability for $\boldsymbol{y}^0$, then C is strictly subadditive at $\boldsymbol{y}^0$.

Proof

Let $\boldsymbol{y}^1 + \cdots + \boldsymbol{y}^k = \boldsymbol{y}^0$, with $\boldsymbol{y}^i \neq \mathbf{0}$ and $0 < \boldsymbol{a} \cdot \boldsymbol{y}^i < \boldsymbol{a} \cdot \boldsymbol{y}^0$, where $\boldsymbol{a} > \mathbf{0}$ is the vector of coefficients that defines H. Then the vector $(\boldsymbol{a} \cdot \boldsymbol{y}^0 / \boldsymbol{a} \cdot \boldsymbol{y}^i)\boldsymbol{y}^i \equiv \alpha^i \boldsymbol{y}^i \in H$

is well defined. Letting the vector $\boldsymbol{v}$ contain the coefficients of the hyperplane that supports C at $\boldsymbol{y}^0$, by hypothesis, $C(\alpha^i \boldsymbol{y}^i) \geq \boldsymbol{v} \cdot (\alpha^i \boldsymbol{y}^i) + v_0$. Dividing by α^i yields:

$$C(\alpha^i \boldsymbol{y}^i)/\alpha^i \geq \boldsymbol{v} \cdot \boldsymbol{y}^i + v_0/\alpha^i. \tag{22}$$

Since $\alpha^i > 1$, declining ray average costs ensure that

$$C(\boldsymbol{y}^i) > C(\alpha^i \boldsymbol{y}^i)/\alpha^i. \tag{23}$$

Putting (22) and (23) together yields $C(\boldsymbol{y}^i) > \boldsymbol{v} \cdot \boldsymbol{y}^i + v_0/\alpha^i$. Summing over all i yields

$$\sum_i C(\boldsymbol{y}^i) > \boldsymbol{v} \cdot \sum_i \boldsymbol{y}^i + \left[\boldsymbol{a} \cdot \sum_i \boldsymbol{y}^i/(\boldsymbol{a} \cdot \boldsymbol{y}^0)\right] v_0 = \boldsymbol{v} \cdot \boldsymbol{y}^0 + v_0 = C(\boldsymbol{y}^0).$$

Q.E.D.

The logic behind this proof is as follows. Let $k = 2$, so that $\boldsymbol{y}^1 + \boldsymbol{y}^2 = \boldsymbol{y}^0$ in Figure 1.2. Now extend rays from the origin through $\boldsymbol{y}^1$ and $\boldsymbol{y}^2$ to H, the hyperplane along which the cost function is trans-ray supportable, i.e. to the points $\alpha_1 \boldsymbol{y}^1$ and $\alpha_2 \boldsymbol{y}^2$. By declining ray average costs, the unit cost of each of these commodity bundles is thereby reduced. Now, since $\boldsymbol{y}^0$ can be expressed as a weighted sum of $\alpha_1 \boldsymbol{y}^1$ and $\alpha_2 \boldsymbol{y}^2$, trans-ray supportability ensures that the cost of producing it is less than or equal to a similarly weighted sum of the costs of those output vectors. Thus, both steps in the procedure which makes it possible to compare $C(\boldsymbol{y}^0)$ to $C(\boldsymbol{y}^1) + C(\boldsymbol{y}^2)$ serve to reduce the former relative to the latter.

This combination of declining ray average cost and trans-ray supportability is another set of sufficient conditions for natural monopoly. A whole class of stronger sufficient conditions is immediately available, since any cost complementarity condition that implies trans-ray supportability will, when combined with *DRAC*, yield subadditivity. These stronger conditions may prove easier to verify on the basis of the parameter values of empirically estimated cost functions. Two conditions that guarantee that the cost function has a support in at least one trans-ray direction are *trans-ray convexity* of the cost function and quasiconvexity of the cost function.

The concept of trans-ray convexity, developed in Baumol (1977), requires that the cost function be convex on the trans-ray hyperplane in question, e.g. line RT in Figure 1.2. Since a convex function can be supported at any point in its domain, trans-ray convexity of the cost function with respect to any hyperplane immediately implies trans-ray supportability. Quasiconvexity can be shown to

imply that the cost function has a support over the trans-ray hyperplane defined by the gradient of the cost function at the point in question.[21] There is one more issue that must be discussed in connection with this set of sufficient conditions for subadditivity and natural monopoly. Trans-ray supportability (unlike trans-ray convexity) does not rule out the presence of product specific fixed costs. In Figure 1.3, these would show up as jump discontinuities above R and T. As with the cost function $\tilde{C}$, it may still be possible to support the cost function above a point such as y^0. If, however, product specific fixed costs were greater, so that the single product cost levels dropped to s and t, then none of the cost functions depicted could be supported over the trans-ray in question. Yet it seems intuitively clear that the degree of economies of scale and the extent of natural monopoly can only be enhanced by increases in fixed costs, since they would seem to increase the advantage of single firm production over that of any multifirm alternative. Therefore, let us again use the formulation introduced above, writing $C(y) = F\{S\} + c(y)$, where S is the set of outputs produced in strictly positive quantities. Then we can state the following result:

Proposition 14

If $c(y)$ is strictly (weakly) subadditive at y^0 and $F\{S\}$ is weakly (strictly) subadditive in the sense that $F\{S \cup T\} \leq (<) F\{S\} + F\{T\}$, $\forall S, T \subset N$, then $C(y)$ is strictly subadditive at y^0.

Proof

Consider the nonzero output vectors $y^1, y^2, \ldots, y^k$ s.t. $\sum y^s = y^0$. Then

$$\sum_{s=1}^{k} C(y^s) = \sum_{s=1}^{k} F\{S^s\} + \sum_{s=1}^{k} c(y^s),$$

where $S^s = \{i \in N: y_i^s > 0\}$. Using the weak (strict) subadditivity of F, we have

$$\sum_{s=1}^{k} C(y^s) \geq (>) F\{S\} + \sum_{s=1}^{k} c(y^s),$$

where $N \supseteq S = \{i \in N: \sum y_i^s \equiv y_i^0 > 0\}$. The result then follows from the strict (weak) subadditivity of $c(\cdot)$. Q.E.D.

[21] The proof of both of these assertions can be found in Baumol, Panzar and Willig (1982, ch. 4, appendix I, p. 91).

This completes the present discussion of sufficient conditions for subadditivity. Additional sets of sufficient conditions can be found in Baumol, Panzar and Willig (1982, ch. 7) and Sharkey (1982, ch. 4).

3. Industry configurations[22]

Thus far we have examined the properties of the cost function of the firm that are important determinants of firm and industry structure. However, in order to gain a complete understanding of market structure, it is necessary to understand the interactions between the determinants of firm size and the size of the market. The former is, as I have argued, determined in large part by the position of the cost function. The latter is determined by the position of the market demand curve. The interaction between these two exogenously given constructs places bounds on the structure of the industry, i.e. limits on the number and size distribution of firms that can be present in equilibrium.

In a private enterprise economy any industry structure that persists in the long run must yield the firms in the industry at least zero economic profits. This places certain restrictions on the relative locations of the cost and demand curves. Thus, at a minimum, it must *not* be the case that the market demand curve lies entirely to the left of the firm average cost curve. For in such a circumstance the firm and industry could not break even unless it had recourse to some form of discriminatory pricing policies or a subsidy. Note that this is true even though it may well be the case that the industry in question *ought* to produce because the total benefits to consumers, as approximated by the area under the market demand curve, exceed the total cost of providing, say, W units of output in Figure 1.4.

Thus, a minimal requirement for inclusion in the set of industries relevant to the student of Industrial Organization, is that there exist some industry configurations that are *feasible* in the sense that the firms involved in the industry at least break even. It will prove useful to precisely define the terms to be used in this discussion:

Definition 17. Industry configuration

An industry configuration is a number of firms, m, and associated output vectors $\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^m$ such that $\sum \mathbf{y}^i = \mathbf{Q}(\mathbf{p})$. Here, $\mathbf{p}$ is the vector of market prices and $\mathbf{Q}(\mathbf{p})$ is the system of market demand equations.

[22] The discussion in this section is based upon that in chapter 5 of Baumol, Panzar and Willig (1982), which, in turn, built upon Baumol and Fischer (1978).

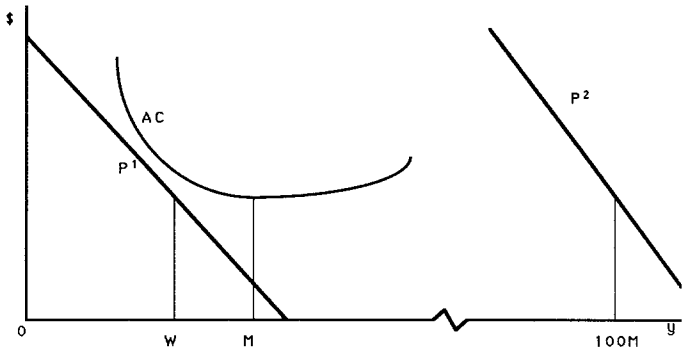

Figure 1.4

Definition 18. Feasible industry configuration

An industry configuration is said to be feasible if, in addition, it is the case that $\boldsymbol{p} \cdot \boldsymbol{y}^i \geq C(\boldsymbol{y}^i)\ \forall i$. If, alternatively, we use the system of market inverse demand relationships, $\boldsymbol{P}(\boldsymbol{y}^I)$, this condition becomes $\boldsymbol{P}(\boldsymbol{y}^I) \cdot \boldsymbol{y}^i \geq C(\boldsymbol{y}^i)$, where $\boldsymbol{y}^I = \sum \boldsymbol{y}^i$ is the industry output level.

Since the primary focus of this analysis is on long-run industry structure, the first definition limits attention to industry situations in which supply equals demand and the second requires that *each firm* earns non-negative profits from its market activities. One must go further than this, of course. For there are situations in which there may exist feasible industry configurations containing one, four, or a hundred firms. The industry demand curve $P^2(y^1)$ in Figure 1.4 illustrates such a situation, if one imagines that the average cost curve rises, but only imperceptibly, beyond M. While competitive, monopoly or oligopoly market structures may all be *feasible* for this industry, common sense and standard practice suggest that this industry be classified as naturally competitive. Thus, another important characteristic of an industry configuration is its *efficiency*.

Definition 19. Efficient industry configuration

$\{y^1, y^2, \ldots, y^m\}$ is an efficient industry configuration if and only if

$$\sum_{j=1}^{m} C(\boldsymbol{y}^j) = \min_{m,\, y^1, \ldots, y^m} \sum_{j=1}^{m} C(\boldsymbol{y}^j) \equiv C^I(\boldsymbol{y}^I),$$

where $y^I \equiv \sum y^j$ is total industry output and $C^I(y^I)$ is the *industry cost function*. Thus, an industry configuration is efficient if and only if it consists of a number of firms and a division of output that yield the lowest possible total industry costs of producing the industry output vector in question.

The analysis that follows will focus on the determination of the number of firms that can constitute a feasible and efficient industry configuration for the relevant set of industry output levels.[23] This focus does not mean that it is logical to presume that only efficient industry configurations can be observed in real world industries. Rather, it is an attempt to determine an unambiguous standard for determining the maximum amount of concentration that is *required* by considerations of productive efficiency. Thus, as the analyses of later chapters in this Handbook indicate, there may be strategic considerations that cause an industry to remain a monopoly even if it is structurally competitive. However, it is important to recognize that this type of "market failure" argument can go only one way. If, for example, only one or few firms can be part of feasible and efficient industry configurations, that industry simply *cannot* be structurally competitive.

3.1. Feasible and efficient single product industry configurations

I shall now relate the above constructs to the standard textbook practice of making inferences about market structure from the relative positions of the market demand curve and the average cost function of the firm. Suppose the firm's average cost curve is as depicted in Figure 1.5. If the market inverse demand curve is given by $P^1(y)$, then the industry has been traditionally classified as a "natural monopoly".[24] Alternatively, if the market inverse demand curve is given by $P^2(y)$, so that it intersects the competitive price level p_c at an output level, C, that is a large multiple of M, the industry is classified as structurally competitive. Finally, if C is a small multiple, then the industry is

[23] The determination of the relevant set of industry output levels can be a nontrivial exercise, especially in the multiproduct case. In the single product examples depicted in Figure 1.5, the relevant set of industry outputs for the industry whose inverse demand curve is given by $P^2(y)$ is the compact interval $[W, 100M]$. For output levels smaller than W, there exists no price at which even a monopolist could break even. For outputs greater than $100M$, consumers' willingness to pay is less than the lowest possible unit cost achievable by the industry. For the industry facing the inverse demand curve given by $P^1(y)$, the set of relevant output levels is empty.

[24] Of course the subadditivity analysis of the previous section has revealed that the natural monopoly region will typically also include some output levels to the right of the minimum point of the average cost curve.

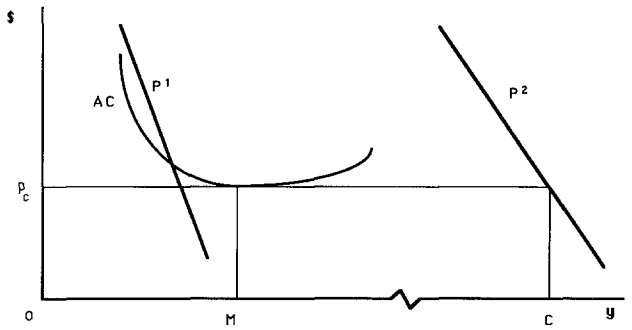

Figure 1.5

traditionally considered as likely to be an oligopoly. The following propositions make this standard practice precise.[25]

Proposition 15

Assume that the average cost function $C(y)/y$ has a unique minimum at y^M, is strictly decreasing for $0 < y < y^M$, and is strictly increasing for $y > y^m$. Then the cost-minimizing number of firms for the production of the industry output y^I is exactly y^I/y^M if that number is an integer. In this case, $C^I(y^I) = (y^I/y^M)C(y^M) \equiv y^I \cdot AC^M$. If y^I/y^M is not an integer, then the cost-minimizing number of firms is either the integer just smaller or the integer just larger than y^I/y^M.

This result formalizes the intuitive notion that, with U-shaped average cost curves, the most efficient way to produce any given industry output is to divide it up equally among the required number of minimum efficient scale firms. It goes beyond this, however, in that it addresses the case in which the required number is not an integer. This turns out to be a nontrivial extension that requires the hypothesis that the average cost curve be monotonically decreasing (increasing) for outputs smaller (greater) than y^M.

Similarly, it is possible to rigorously justify the standard practice of determining industry structure using the relative positions of the market demand and average cost curves.

[25]These results, which are stated here without proof, are from Baumol, Panzar and Willig (1982). Proposition 15 has also been proved by Ginsberg (1974) under more restrictive conditions.

Proposition 16

Assume that the average cost function has a unique minimum at y^M, is strictly decreasing (increasing) for $0 < y < (>) y^M$. Let $[x]$ denote the smallest integer at least as large as x. Then no more than $[Q[AC^M]/y^M]$ firms can participate in a feasible and efficient industry configuration and there always exists a feasible and efficient industry configuration containing $[Q[AC^M]/y^M] - 1$ firms.

This result instructs the analyst to find the quantity demanded at the price equal to unit cost at y^M. Next, divide that quantity by y^M, the quantity that minimizes average cost, to determine the critical number $Q[AC^M]/y^M \equiv m^*$. The proposition establishes that no feasible and efficient industry configuration has more than $[m^*]$ firms and that there *does* exist a feasible and efficient industry configuration of $[m^*] - 1$ firms. Note that this test requires quantitative information about the cost function only at y^M.

Thus, if m^* is large, the industry is structurally competitive, because there are feasible and efficient industry configurations with as many as $[m^*] - 1$ firms. If $0 < m^* < 1$, the industry is a natural monopoly, since there can exist no feasible and efficient configurations of more than one firm. However, in this case, one cannot be sure that there exist *any* feasible configurations. That depends upon whether the demand curve intersects the average cost curve or not, as in the case of $P^1(y)$ in Figure 1.4. That cannot be determined from cost information only at y^M.

Before leaving the single product world, it is important to modify the above results to deal with an important departure from the assumption that average cost curves are strictly U-shaped. Conventional wisdom holds that average costs in many industries decline for a range of outputs, attain their minimum at

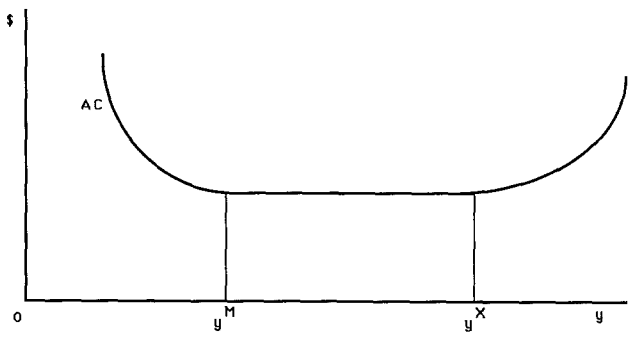

Figure 1.6

minimum efficient scale, and then remain constant for a considerable range of output levels.[26] Figure 1.6 depicts such a situation, with minimum efficient scale being achieved at y^M and average cost remaining constant through output level y^X, the point of *maximum efficient scale*, and rising thereafter.[27]

If average costs are constant up to an output level at least twice as large as minimum efficient scale (i.e. $y^X \geq 2y^M$), the set of market structures that are consistent with feasible and efficient industry configurations is greatly expanded. This means that Propositions 15 must be modified:

Proposition 17

When $y^X \geq 2y^M$, an efficient industry configuration for industry output levels $0 < y^I < 2y^M$ can involve only one firm, and for larger industry outputs at least $\underline{m}$ firms and at most $\overline{m}$ firms are required, where $\underline{m}$ is the smallest integer greater than or equal to y^I/y^X and $\overline{m}$ is the largest integer less than or equal to y^I/y^M.

Thus, if y^M is large relative to $Q(AC^M)$, a competitive industry structure is inconsistent with industry cost minimization. Similarly, if $Q(AC^M)/y^X$ is small, a concentrated industry does not result in any loss of *productive efficiency*, though, of course, there may be welfare losses from oligopolistic pricing.

3.2. *Efficient multiproduct industry configurations*

The problem of establishing bounds on the number of firms that can participate in an efficient industry configuration is considerably more complicated in the multiproduct case. First, when there are two or more products, any given industry output vector $\mathbf{y}^I$ can be apportioned among firms in ways that may involve some or all of the firms producing different output mixes, i.e. operating on different output rays. Second, as noted earlier, the size of firm at which economies of scale are first exhausted may differ across output rays, so that the set of outputs at which there are locally constant returns to scale will be a locus rather than a single point. Therefore, in order to get any results at all, it is necessary to assume that all ray average cost curves are strictly U-shaped, so that all points inside (outside) the M-locus depicted in Figure 1.7 exhibit increasing (decreasing) returns to scale. Also, since economies of scale do not ensure

[26] See Bain (1954) and the discussion in chapter 4 of Scherer (1980).

[27] Of course this latter region may never be observed, since no firm would be operating there under most reasonable notions of industry equilibria.

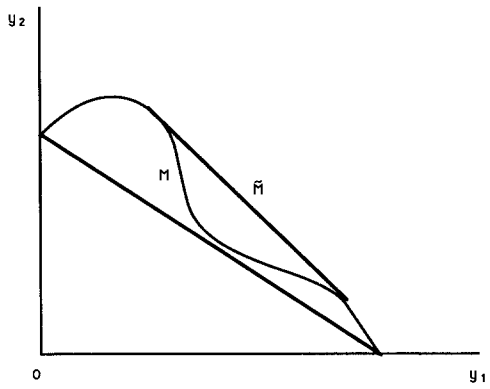

Figure 1.7

subadditivity in the multiproduct case, it is necessary to *assume* that the firm cost function $C(y)$ is strictly subadditive "inside" the M-locus.[28]

Next, let $\tilde{M}$ denote the convex hull of the M-locus, as illustrated in Figure 1.7. Then, for any given industry output vector y^I, let $\bar{t}(y^I) = \max\{t: ty^I \in \tilde{M}\}$ and $\underline{t}(y^I) = \min\{t: ty^I \in \tilde{M}\}$. Then the following result is true:[29]

Proposition 18

The cost-minimizing number of firms for the production of industry output vector y^I, $m(y^1)$, satisfys the following conditions: (i) $m(y^I) > 1/2\bar{t}(y^I)$ and $m(y^I) \geq [1/2\bar{t}(y^I)]$; (ii) $m(y^I) = 1$ if $\underline{t}(y^I) \geq 1$, otherwise, $1 \leq m(y^I) < 2/\underline{t}(y^I)$ and $1 \leq m(y^I) \leq [2/\underline{t}(y^I)] - 1$.

These upper and lower bounds on the cost-minimizing number of firms tell us, in effect, that the "average-sized" firm in the industry must be sufficiently close to the M-locus. Also, it is clear that these bounds are not nearly as "tight" as those in the single product case. Nevertheless, they are the tightest bounds available, as it is possible to construct examples in which they are exactly satisfied.[30]

[28] In the single product case subadditivity was implied by the assumption that average cost was decreasing up to y^M. Actually, this assumption is required only for output vectors inside the convex closure of M, a concept to be defined below.

[29] Baumol, Panzar and Willig (1982, proposition 5F1). The proof is given in appendix III to chapter 5.

[30] Baumol, Panzar and Willig (1982, pp. 119–120).

There is another issue in the multiproduct case. Even if the lower bound on $m(\mathbf{y}^I)$ is large, that would not be sufficient to conclude that the industry is likely to be competitive. For such a finding would indicate only that one could expect a large number of firms involved in producing *all* the industry's products, not a large number of firms producing *each* product, which is required for the *industry* to be competitive. Thus, it is necessary to modify Proposition 18 in order to calculate bounds on the number of firms producing any particular product or subset of products in an efficient industry configuration:[31]

Proposition 19

Let $m_S(\mathbf{y}^I)$ denote the number of firms producing products in the subset S in an industry configuration efficient for the production of $\mathbf{y}^I$, and let $M_S \equiv \{\mathbf{y}_S\colon\ \mathbf{y} \in M\}$ denote the projection of the M-locus in the subspace corresponding to the product subset S, with $\tilde{M}_S$ its convex hull. Then $m_S(\mathbf{y}^I) > 1/2t^S(\mathbf{y}^I)$ and $m_S(\mathbf{y}^I) \geq [1/2t^S(\mathbf{y}^I)]$, where $t^S(\mathbf{y}^I) \equiv \max\{t\colon\ t\mathbf{y}_S^I \in \tilde{M}_S\}$.

The implications of Proposition 19 are illustrated in Figure 1.8 for the case of a two-product industry. If the object of the investigation is to determine whether or not the industry is a candidate for pure competition, then it is necessary to place large lower bounds on the number of producers of *both* products required in an efficient industry configuration producing $\mathbf{y}^I > \mathbf{0}$. Specializing Proposition 19 to the case of $S = \{i\}$, the relevant lower bounds are $m_i(\mathbf{y}^I) > y_i^I/2\tilde{y}_i$, and $m_i(\mathbf{y}^I) \geq [y_i^I/2\tilde{y}_i]$, where $\tilde{y}_i = \max\{y_i\colon\ \mathbf{y} \in M\}$. Note that it is the maximum value of y_i over the entire projection of the M-locus, rather than $\hat{y}_i$, the output that achieves minimum efficient scale in stand-alone production, that is used to calculate the lower bounds. When, as drawn, the M-locus is "bowed out" from the axes, $\tilde{y}_i$ may be considerably larger than $\hat{y}_i$. Also note that these are the relevant bounds even if the ray through $\mathbf{y}^I$ intersects the M-locus at output levels for each product that are small relative to $\tilde{y}_i$, i.e. even when an industry of "average" size firms would be relatively unconcentrated.

The above discussion indicates the important role played by the shape of the M-locus in determining the lower bounds on the number of firms producing in an efficient industry configuration. Given this, it is important to know what shape the M-locus is "likely" to be, based upon properties of the multiproduct cost function. Unfortunately, the answer here is discouraging. Under most plausible scenarios, the M-locus will tend to have the "bowed out" shape shown in Figure 1.8. This is true even if the two production processes are completely independent! To see this, consider the rectangle formed by $\hat{y}_1$, $\hat{y}_2$, $\hat{\mathbf{y}}$, and the origin. At any point on, say, the right border of this rectangle, $S_1 = 1$ and $S_2 > 1$. But, from

[31] Baumol, Panzar and Willig (1982, proposition 5G1, p. 123).

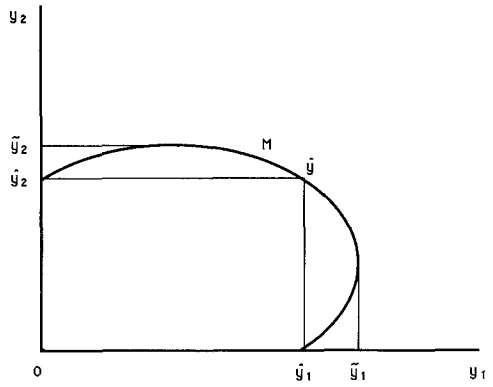

Figure 1.8

equation (6), we know that S is simply a weighted sum of S_1 and S_2 when the products are produced independently. Therefore S must be greater than 1. When ray average costs are U-shaped, this means that one must proceed outward along the ray from the origin before S falls to 1. This effect is increased when there are economies of scope, for then S *exceeds* the weighted average of S_1 and S_2. Finally, when there are product specific fixed costs, the M-locus is discontinuous at the axes, so that $\hat{y}_1$ and $\hat{y}_2$ lie *inside* the points where the M-locus reaches the axes.[32] Without assuming diseconomies of scope, only the limiting case in which (at least some) inputs are perfectly transferable between outputs can one expect the M-locus not to be concave to the origin. In that case, the cost function can be written as $C(y) = \sum c_i y_i + \Phi[\sum a_i y_i]$ and the M-locus is a hyperplane (e.g. the outer border of the rectangle in Figure 1.8) characterized by $\{ y: \sum a_i y_i = k \}$, where $k\Phi'(k) = \Phi(k)$.[33]

This completes the discussion of the theoretical results underlying the role played by technology in the determination of industry structure. The next section examines the extent to which practice has kept up with theory in this area.[34]

4. Empirical issues

The remainder of this chapter is devoted to a discussion of issues that arise in attempting to give empirical content to the theory developed in the previous two sections. The first two subsections discuss general methodological problems that

[32] See Baumol, Panzar and Willig (1982, figure 9C3, p. 255).

[33] Baumol, Panzar and Willig (1982, pp. 129–130).

[34] As is usual, practice tends to follow theory only with a lag. This is especially true with respect to the material of this section. For an exception, see Wang Chiang and Friedlaender (1985).

often arise in empirical applications. The final two subsections discuss empirical cost function studies in electric power and telecommunications. I will not be concerned with econometric methodology, but rather with the extent to which the cost functions estimated are useful for addressing the questions of industry structure that, presumably, provided the initial motivation for such empirical studies. For as Nerlove (1963) remarked (p. 409) in his pioneering study of electricity supply, "the first question one must ask is 'To what use are the results to be put?' ". The bulk of the discussion of empirical studies will be devoted to those that have been published in the last few years, since surveys by Gold (1981) and Bailey and Friedlaender (1982) in the *Journal of Economic Literature* cover the earlier literature in some detail.

4.1. Aggregation and the hedonic approach[35]

In most real world industries, firms are likely to be producing a large number of distinct products. Thus, in attempting to estimate cost functions econometrically, the analyst will usually be forced to aggregate the output data in some way in order to reduce the parameters to be estimated to a manageable number. Until fairly recently, the typical approach was to construct a single, scalar measure of output, $Y \equiv \sum a_i y_i$, were $\boldsymbol{a} > \mathbf{0}$ is some vector of weights, often based on output prices. However, this procedure imposes the implicit restriction that the multiproduct cost function can be written as $C(\boldsymbol{y}) = \tilde{C}(Y)$. Unfortunately, this imposes severe restrictions on the important multiproduct cost constructs developed earlier in this chapter. Of course, in principle, it is possible to test the validity of such restrictions, but such tests usually require that sufficient data is available to render such restrictions unnecessary in the first place.[36]

Of course, if there is reason to believe that the true multiproduct cost function can be written in this simple form, all of the important multiproduct cost constructs discussed above can be calculated from the estimated parameters of $\tilde{C}$. However, there is also a situation in which it is possible to reliably infer something about the properties of C from estimates of $\tilde{C}$. Suppose all of the output vectors in the sample lie on or close to the same ray, i.e. firms in the sample always produced essentially the same output mix. Then the single product measure of economies of scale calculated from the parameters of $\tilde{C}$ will correctly measure the degree of multiproduct scale economies *along the ray in question*. In particular, the intersection of that ray with the M-locus, at which returns to scale are locally constant, can be correctly identified. It is important to note, however,

[35]Most of this section is drawn from the discussion in Baumol, Panzar and Willig (1982, pp. 446–448).

[36]See Blackorby, Primont and Russell (1977) or Denny and Fuss (1977) for discussions of econometric aggregation tests.

that it is *not* generally valid to extrapolate those measures to output mixes outside the sample. As the discussion of the previous section indicates, knowledge of the *entire* M-locus is generally required when calculating bounds on the number of firms in an efficient industry configuration for *any* output mix.

During the last decade, estimates of *hedonic cost functions* have become common in the literature. These represent a compromise between the estimation of scalar aggregate and multiproduct product cost functions. This approach was pioneered by Spady and Friedlaender (1978) in their analysis of trucking firms. Rather than attempt to estimate costs as a function of the (large) number of different types of freight carried over a (very large) number of origin and destination pairs, they specified costs as a function of aggregate ton-miles and hedonic variables such as the average length of haul. Formally, a hedonic cost function can be expressed as $\tilde{C}(Y, Z_1, \ldots, Z_k)$, where Y is, again, a scalar measure of aggregate output and $\boldsymbol{Z}$ are hedonic measures of the output mix. Of course, if enough such hedonic measures are included, then the output vector $\boldsymbol{y}$ could be reconstructed from Y and $\boldsymbol{Z}$, so that there would exist a hedonic cost function representation equivalent to the true multiproduct cost function. However its estimation would, in general, require estimating the same number of parameters.

Use of hedonic cost functions enables the investigator to, in effect, perform often unavoidable aggregation based upon informed judgements about characteristics that are likely to have important impacts upon the costs associated with producing a given aggregate output vector. If the resulting hedonic cost function is judged to be good approximation of the true multiproduct cost function, the multiproduct cost characteristics developed above can be computed from its parameter estimates and employed in an analysis of industry structure. This last step requires some care in interpretation, however, even if the specified hedonic cost function is assumed to reflect the true multiproduct cost structure.

For example, consider the recent study of airline cost by Caves, Christensen and Tretheway (1984). They estimate a cost function for airlines as a function of, inter alia, aggregate output (e.g. revenue passenger miles, RPM) and P, the number of points served by the firm. Therefore, assume that the true cost function can be written as $C(\boldsymbol{y}) = H\{G(Y), P\}$, where Y is aggregate output and P is the number of points served. Caves et al. define two cost concepts to be used in describing the cost characteristics of airline networks, returns to density, $RTD = H/YG'$ and "returns to scale", $RTS = H/[PHp + YG']$. RTD measures the proportional increase in output made possible by a proprotional increase in all inputs, with points served (and other hedonic measures) held constant. RTS measures the proportional increase in output and points served made possible by a proportional increase in all inputs, ceteris paribus. At the sample mean, they found significant returns to density ($RTD \approx 1.2$), but essentially constant "returns to scale" ($RTS \approx 1.0$).

How do these cost concepts relate to those that have been shown to be important for the determination of industry structure earlier in this chapter? In order to examine this question, it is necessary to relate Caves et al.'s hedonic cost function more directly to an underlying multiproduct cost function. For expository purposes, it is convenient to assume a simple structure for H, i.e. $H = cY + rP$, where c and r are positive constants. Then the underlying multiproduct cost function can be written as

$$C(y) = c \cdot \sum_{i \in T} y_i + rP = c \cdot Y + rP = H\{G(Y), P\},$$

where T is the set of markets served and P is the cardinality of T.[37] Now it is easy to see that returns to density are precisely equal to (what has been previously defined to be) the degree of multiproduct economies of scale! That is, $RTD = [(c \cdot Y + rP)/c \cdot Y] = C(y)/\nabla C(y) \cdot y \equiv S$. Also, in this example it is easy to see that Caves et al.'s measure of "returns to scale" is always equal to 1. Unfortunately, it is also easy to see that that fact is not particularly relevant for the analysis of industry structure described in Sections 2 and 3.

Examining this example in terms of the cost concepts developed above, reveals globally increasing returns to scale, both with respect to the entire product set and any subset of products. That is, S and S_T both exceed 1 at all output levels and for all subsets of markets T. There are no economies of scope in this example. However, the more general hedonic specification is consistent with either economies or diseconomies of scope. These would be determined by the returns to scale properties of G and of H with respect to P. The extent of economies of scope would, of course be useful in determining efficient firm size. However, measures of economies of scope were not computed.

Industries characterized by network technologies are disproportionately represented in econometric cost studies. There are two related reasons for this. First, network technologies are usually thought to be characterized by economies of scale. This has resulted in most of them being regulated over the years, which, in turn, has meant much better than average data availability for cost function estimation. Furthermore, the opening of such industries to competitive entry has often focused important policy debates on the extent of scale economies that may or may not be present. Unfortunately, their network structure makes the aggregation problem under discussion particularly severe. If point-to-point transportation (or transmission) movements are viewed as the true cost-causitive outputs of the firm, a firm operating even a relatively small network must be viewed as producing an astronomical number of products. The hedonic approach has, in large part, arisen as an attempt to deal with the problem of networks.

[37]For ease of exposition, I am ignoring the network aspects of airline costs, so that the number of economically distinct outputs is the same as the number of points served, as in a simple star-shaped network.

In an important recent paper, Spady (1985) proposes an innovative solution to this problem. By assuming that the cost of production on each link of the network are quadratic, he is able to construct a multiproduct cost function that is econometrically parsimonious, yet a true aggregate of the underlying production processes. All that is required, in addition to aggregate data, are estimates of the first and second moments of the distribution of the links' traffic and technological characteristics. This exciting approach has, to my knowledge, yet to be empirically implemented.

4.2. Long-run and short-run measures of returns to scale

The discussion in this chapter has focused on the properties of the *long-run* cost function and the role that they play in determining equilibrium industry structure. However, in an empirical application that attempts to estimate the cost function of an individual firm, the data available may be better suited for estimating a *short-run* cost function: a specification that assumes that only some of the inputs available to the firm are set at their cost-minimizing levels. For example, if a cost function were to be estimated using monthly data, it would be unrealistic to assume that the firms capital inputs were adjusted to the cost-minimizing level associated with each month's rate of output. In that case, it might be appropriate to estimate a *variable cost* function representation of the technology. Conceptually, this is done by dividing the vector of inputs available to the firm into two categories: $\boldsymbol{x} = (\boldsymbol{v}, \boldsymbol{k})$. The *variable inputs* $\boldsymbol{v}$ are assumed to be observed at their cost-minimizing levels but the *fixed inputs* $\boldsymbol{k}$ may or may not be. Then $V(\boldsymbol{y}, \boldsymbol{k}, \boldsymbol{s})$ is defined as the minimum expenditure on variable inputs $\boldsymbol{v}$ required to produce the output vector $\boldsymbol{y}$, given the availability of the fixed inputs at level $\boldsymbol{k}$, provided that $\boldsymbol{y}$ can be produced from $\boldsymbol{k}$. That is, $V(\boldsymbol{y}, \boldsymbol{k}, \boldsymbol{s}) = \min_{\boldsymbol{v}}\{\boldsymbol{s} \cdot \boldsymbol{v}\colon\ (\boldsymbol{v}, \boldsymbol{k}, \boldsymbol{y}) \in T\}$, where $\boldsymbol{s}$ is the vector of variable input prices. Examples of empirical cost studies that estimate variable cost functions include the telecommunications study by Christensen, Cummings and Schoech (1983) and the railroad study of Friedlaender and Spady (1981).

Once estimates of the parameters of a variable cost function are obtained, however, how does one calculate the degree of economies of scale to be used for policy purposes? For example, Braeutigam and Daughety (1983) present the following measure of economies of scale:

$$\hat{S} = \left\{V - \sum_i (\partial V/\partial k_i)k_i\right\} \Big/ \left\{\sum_j (\partial V/\partial y_j)y_j\right\}.$$

This measure of economies of scale is defined as a function of variable input prices, output levels and fixed input levels, i.e. $\hat{S} = \hat{S}(\boldsymbol{y}, \boldsymbol{k}, \boldsymbol{s})$. Clearly, this measure differs from both $S(\boldsymbol{y}, \boldsymbol{w})$ and $\tilde{S}(\boldsymbol{y}, \boldsymbol{x})$, the measures discussed in

Subsection 2.1. However, the relationships between these measures should be clear. $\hat{S}$ is a hybrid of the technological and cost function measures of economies of scale. $\tilde{S}$ reflects a fundamental property of the productive technology and can be calculated at every point in input/output space, $X \times Y$. S measures the elasticity of costs with respect to output(s). By construction, it pertains only to cost-efficient input/output combinations.

If one assumes, instead, that only a subset of inputs is selected optimally, $\hat{S}$ is the resulting measure of economies of scale. In fact, using the same argument that established Proposition 2, it is possible to show that $\tilde{S}(y, \hat{v}(y, k, s), k) = \hat{S}(y, k, s)$, where $\hat{v}$ is the argmin of the above variable cost-minimization problem. Thus, the technological and variable cost measures of the degree of scale economies coincide at those input/output points at which the variable inputs are chosen optimally. Similarly, it is also the case that $\hat{S}(y, k^*(y, r, s), s) = S(y, w)$, where r is the vector of fixed input prices, so that $w = (r, s)$.

Given that it is sometimes necessary and/or desirable to estimate a variable cost function, there remains the question of how to use the estimates to provide the most appropriate information regarding economies of scale. As Braeutigam and Daughety (1983) point out, one would not expect $\hat{S}$ and S to be equal unless the fixed inputs happen to be at their cost-minimizing levels or the technology is homothetic.[38] Furthermore, they show that it is generally not possible to make any inferences about the relative magnitudes of S and $\hat{S}$ for any given y and s. Thus, if one wishes, for policy purposes, to determine the extent of economies of scale *in the long run*, there is no alternative but to use the variable cost function estimates and the vector of fixed input factor prices to derive the long-run cost function $C(y, s, r)$, which can be used to calculate S.[39] Of course, this eliminates one of the perceived advantages of the variable cost function approach, the ability to obtain estimates without observations on the prices of fixed factors. This issue is particularly important for studies in which the "fixed factor" is actually some accounting measure of assets for which it is difficult to impute a price.

4.3. Empirical studies of electric power

Nerlove (1963) provided a pioneering study of economies of scale in electricity generation based upon modern duality theory. Using data from 1955, he estimated a cost function that included factor prices as arguments. His basic

[38] In this context the appropriate multiproduct version of homotheticity requires that the transformation function can be written as $\varphi(v, k, y) = F(h(v, k), y)$, with h linearly homogeneous. That is, there exists a natural aggregate of inputs that can be used to produce any desired output mix. In that case, $\tilde{S}$, the technological measure of economies of scale, is constant along any isoquant.

[39] This is the method used by Friedlaender and Spady (1981) in their study of rail and trucking costs.

estimation equation was a Cobb–Douglas log-linear specification of the cost function:

$$\ln C = K + (1/S)\ln y + (1/S)\sum a_i[\ln p_i].$$

Here S is the (single product) degree of scale economies, the p_i refer to the prices of labor, fuel and capital, and $\sum a_i$ is constrained to equal unity. Two variants of this model yielded estimates of $S \approx 1.4$ [Nerlove (1963, tables 3 and 4)]. To put this in perspective, this would mean that the utility's costs would exceed its revenues by forty percent if its output were priced at marginal cost.

As Nerlove recognized, this functional form has the disadvantage of imposing the condition that the degree of returns to scale is the same for all output levels and factor prices. And, indeed, examination of the residuals from this basic regression equation revealed that the true relationship of costs to output could not be log-linear. Therefore he tried various techniques to allow for the intuitively plausible possibility that the degree of scale economies decreases as output increases. Dividing the sample into five output categories yielded estimates (depending on the treatment of the price of capital) ranging from $S > 2.5$ at small output levels to $S \approx 0.95$ for the largest output category. This suggested that economies of scale were exhausted at large plant sizes.

In order to examine the robustness of these results, Nerlove then estimated equations based upon what he referred to as the hypothesis of "continuous neutral variations in returns to scale". That is, he assumed that the degree of economies of scale depended upon the output level of the firm but not upon the factor prices it faced. (It is easy to show that this implies that the underlying production function must be homothetic.) The estimated equations were of the form:

$$\ln C = K + \alpha[\ln y] + \beta[\ln y]^2 + \sum a_i[\ln p_i].$$

These estimates also yielded the result that the degree of economies of scale declined with output. However, in this case, economies of scale persisted throughout, with estimates ranging from $S \approx 3.0$ at low output levels to $S \approx 1.7$ for the largest plants in the sample. [However, Christensen and Greene (1976) point out an error in Nerlove's calculation of the degree of scale economies from his estimated parameter values. When corrected, the results of this regression equation also show that economies of scale are exhausted by firms with the largest output levels.]

Nerlove's pioneering study was important because it clearly established the cost function, with its factor price arguments, as the proper framework in which to study the returns to scale experience of public utilities and because it demonstrated that economies of scale tend to decline with output, thereby setting

the stage for the application of more flexible function forms better suited to capture this effect empirically.

Christensen and Greene (1976) studied economies of scale in the electric power industry using data from 1970. They employed the translog cost function:

$$[\ln C] = K + \alpha Y + \beta Y^2 + \sum a_i P_i + \sum \sum b_{ij} P_i P_j + \sum \gamma_i Y P_i, \tag{24}$$

where capital letters denote the natural logarithms of the independent variables y and the p_i. In order for (24) to represent a proper cost function, it is required that $\sum a_i = 1$, $\sum \gamma_i = 0$ and $\sum_i b_{ij} = \sum_j b_{ij} = \sum\sum b_{ij} = 0$. This functional form allowed them to encompass all of the equations estimated by Nerlove as special cases and statistically test the validity of the implied restrictions. For example, if one restricts $\beta = b_{ij} = \gamma_i = 0$, equation (24) reduces to the homogeneous Cobb–Douglas function, with a degree of scale economies $(1/\alpha)$ that is constant overall output ranges and unaffected by changes in factor prices. Allowing for $\beta \neq 0$ yields Nerlove's homothetic model in which the degree of scale economies varies with output $(S = 1/(\alpha + 2\beta Y))$ but not factor prices, and retains the property that the elasticity of substitution between factors of production is fixed at unity.

Christensen and Greene found it possible to reject the hypotheses of homogeneity, homotheticity and unitary elasticities of substitution in the generation of electric power, both for their own 1970 data and Nerlove's 1955 data. They also found that maintaining the hypothesis of homogeneity results in estimates of global economies of scale, whereas more flexible representations again reveal that economies of scale are exhausted at very large scales of operation. One implication of their estimates that they do not discuss is the effects of factor price changes on the degree of scale economies. Based on the parameter values reported in their table 4, an increase in the price of labor (fuel) results in a small (≈ 2 percent), but statistically significant, increase (decrease) in the degree of scale economies enjoyed by firms in both 1955 and 1970. Changes in the price of capital appear to have no effect on the degree of scale economies. (The coefficients are extremely small and insignificant for both years.)

Christensen and Greene conclude their article with an illuminating presentation of the average cost curves based on the estimates of the translog model. One interesting finding was that, while the representative average cost curve had shifted downward considerably, the shape of the average cost curve did not change. Thus, since firms had expanded their output levels considerably, far fewer firms were operating with substantial unexploited economies of scale: the figures for steam-generated electric power were 48.7 percent in 1970 versus 74.1 percent in 1955. Finally, while the estimated translog average cost curve was U-shaped, there was a large segment that was approximately flat. This range of

output levels – from 19 to 67 billion kWh – was produced by firms producing 44.6 percent of total output.

The papers by Nerlove, and Christensen and Greene, established standards for rigor, thoroughness and precision in the empirical study of economies of scale. However, they did not even begin to deal with the multiproduct nature of the world inhabited by the firms that they study. Electric utilities often sell natural gas as well. Electricity produced at peak times of the day or seasons of the year may cause the firm to incur much greater costs than electricity generated off peak. If the peak/off peak mix varies across firms (i.e. they are not on the same output ray), estimates of single output cost functions will be biased. And what of the costs of transmission and distribution, and the interrelationship between those costs and the costs of generation? Empirical work addressing such *multiproduct* issues did not really begin until the theoretical constructs discussed in previous sections had been developed.

Mayo (1984) attempted to extend the Christensen and Greene (1976) type of analysis of the efficiency of industry structure in the case of regulated public utilities to the case of multiple outputs, electricity (kilowatt hours) and natural gas (cubic feet). He estimated two forms of the quadratic (in outputs) cost function for electric and gas utilities using data from 1979. In his single intercept equation, Mayo found that the estimated coefficient of the electricity–gas interaction term was positive. This guaranteed that diseconomies of scope would eventually set in, despite the positive estimate of the intercept (fixed cost) term. What is somewhat surprising is that the magnitude of the estimated coefficients are such that diseconomies of scope set in at very small output levels. The estimated coefficient of the quadratic electricity term was also positive, which, in the single intercept case, ensures that there are globally decreasing electricity specific returns to scale. Mayo's estimates also yielded the result that overall returns to scale were exhausted at rather small electricity output levels for all of the gas electricity product mixes considered. This was true despite the negative estimated coefficient of the quadratic gas term (and the resulting global gas specific economies of scale).

Using 1981 data for the same sample, but leaving out utilities that generate more than 10 percent of their electricity using nuclear plants, Chappell and Wilder (1986) obtained estimates of the parameters of the single intercept quadratic cost function that yielded significantly different measures of scale and scope economies. That is to say, their estimated coefficients yielded measures of multiproduct, electricity specific and gas specific economies of scale that were not exhausted by even the largest firms in the sample, and economies of scope prevailed over most of the sample. They attributed the difference between their results and Mayo's to the fact that the relatively larger nuclear utilities had considerably higher ex post cost levels. In response, Mayo (1986) argued that nuclear generation of electric power is a different technique but not a different

technology, and therefore firms that employ it should not be excluded from the sample on a priori grounds. I agree with his position in principle, but the fact that the presence of a relatively few large, high cost firms can dramatically alter the range over which economies of scale and scope pertain is clearly a weakness of the simple quadratic cost specification.

From a methodological point of view, these studies dismiss too readily the use of multiple dummy variables in attempting to measure the fixed costs of the firms in the sample. The use of a single intercept term in a quadratic cost specification is overly restrictive because it assumes away product specific fixed costs and reduces the determination of scope economies to the interplay between the level of fixed costs and the magnitude of the coefficient of the quadratic interaction term. In the two-output case, it is possible to use dummy variables to attempt to measure the fixed costs associated with all possible product sets. That is, it is a simple matter to estimate the function $F\{S\}$ $\forall S \subseteq N$. What is more, it is possible to do this using only two dummy variables. Mayo did this in his Flexible Fixed Cost Quadratic (FFCQ) model. Unfortunately, he decided to favor the single intercept quadratic model on questionable statistical grounds.

Recall what it is one hopes to accomplish via the use of dummy variables. Under the maintained hypothesis that variable costs are a quadratic function of electricity and gas output levels, the object of using dummy variables is to distinguish the intercept (fixed cost) terms of the cost functions of three types of firms: those producing gas only, electricity only, and both electricity and gas. Let FG, FE and FB, respectively, denote the true intercept terms of these three cost functions. Then an appropriate estimating equation would be given by

$$C = \beta_0 + \beta_1 E + \beta_2 B + c.$$

Here, $E(B)$ are dummy variables that take on a value of 1 if electricity only (both electricity and gas) are produced by the firm and zero otherwise and c is a quadratic function of gas and electricity output levels. If this equation is estimated appropriately, the stand-alone fixed cost of gas production, FG, is estimated by the estimated value of the parameter β_0, FE by β_1, and FB by β_2.

Now consider the cost function estimating equation Mayo used:

$$C = \alpha_0 + \alpha_1 E + \alpha_2 G + c.$$

In this equation, $E(G)$ are dummy variables that take on a value of 1 whenever electricity (gas) is produced by the firm and zero otherwise, and c is a quadratic function of gas and electricity output levels. This specification directly estimates the incremental fixed costs of electricity and gas production as α_1 and α_2,

respectively. Thus, *FG* is measured by $\alpha_0 + \alpha_1$, *FE* by $\alpha_0 + \alpha_2$, and *FB* by $\alpha_0 + \alpha_1 + \alpha_2$.

Mayo performed an *F*-test that indicated that one could not reject the hypothesis that $\alpha_1 = \alpha_2 = 0$. However, the fact that one cannot be 90 percent or 95 percent certain that a coefficient is *not* zero does not make it appropriate to treat it as zero in one's calculations. (The estimate of α_0 in Mayo's Quadratic model is not significantly greater than zero, yet he uses its positive value in computing the degrees of scale and scope economies.)

4.4. Empirical studies of telecommunications

The U.S. Department of Justice filed an antitrust suit against the Bell System in 1974, seeking divestiture of the Bell Operating Companies, AT & T Long Lines, Western Electric and Bell Telephone Laboratories. Although it was not the main focus of the legal case raised by the DOJ, a part of AT & T's initial defensive position was that the break-up of the Bell System's "natural monopoly" would result in a loss of economic efficiency. This issue was also at least implicit in the policy debate concerning the entry of MCI and others into AT & T's long-distance monopoly markets. The Bell System's position was that these long-distance markets were natural monopolies and that wasteful duplication of facilities would result if competitive entry were permitted and protected.

Thus, empirically determining the extent of economies of scale in the telecommunications industry became more than a mere academic exercise. As in the case of railroads and trucking, the presence or absence of empirically estimated economies of scale was thought to be vitally important for public policy purposes. However, in the case of telecommunications, it was recognized rather early on that the multiproduct nature of the technology should play an important role in the empirical work. Unfortunately, none of the studies conducted during the late 1970s and early 1980s using U.S. data attempted to shed light upon the multiproduct cost concepts developed earlier in this chapter. However, multiproduct cost studies were made employing data from Bell Canada. All of this work is discussed in detail in an important survey article by Fuss (1983). The discussion that follows draws heavily from that source.

The studies covered by Fuss's survey include Fuss and Waverman (1981), Denny et al. (1981), Nadiri and Shankerman (1981), Breslaw and Smith (1980), Denny and Fuss (1980), and Christensen et al. (1983). All of these are translog-based studies. The survey restricts its attention to these papers because the maintained hypotheses (homogeneity, constant elasticities of substitution, Hicks-neutral technical change, etc.) in earlier studies were statistically tested and rejected using the more flexible translog specification. Of course, it is not immediately clear that flexibility of the cost function with respect to *input* prices

is the most important criterion to use when attempting to evaluate empirical estimates of its multiproduct *cost* properties.

The first issue that must be faced when using the translog flexible functional form to characterize a multiproduct cost function is how to handle observations which contain zero values for one or more output variables. The translog cost function is undefined at such points, since it is a quadratic function of the logarithms of the independent variables. This issue does not arise in the estimation of the traditional single product translog cost function, since the independent variables (output and factor prices) could take on only strictly positive values. Even if all output levels are strictly positive throughout the sample, calculation of important multiproduct cost measures such as the degrees of economies of scope or product specific economies of scale require evaluating the cost function at points at which one or more output levels are zero.

Fuss and Waverman (1981) circumvent that problem by employing a Box–Cox transformation of the output variables: $\hat{y}_i = (y_i^{\theta} - 1)/\theta$. Here, θ is a parameter to be estimated and $\hat{y}_i$ replaces $[\log y_i]$ in the translog estimation equation. Since $\hat{y}_i$ approaches $[\ln y_i]$ as θ approaches 0, the translog specification can be approximated arbitrarily closely by a sufficiently small value of θ. Similarly, when $\theta = 1$, a linear specification results. Thus, for θ in the unit interval, this *hybrid translog* cost function can be thought of as a mixture of linear and log-linear functional forms.

Even when the problems posed by zero output levels are solved, there remains the difficulty of calculating and interpreting the multiproduct cost constructs obtained from the estimated equations. For flexible functional forms, measures of economies of scale and other multiproduct cost constructs depend upon output levels as well as the values of estimated parameters. This is a desirable property, since one intuitively expects that the degree of economies scale or scope decline with size. However, this poses a difficult problem for the researcher attempting to summarize his results concisely, since the degrees of economies of scope and (various types of) economies of scale will be different at each point in the sample. The standard technique employed in practice is to normalize the data by dividing the values of each observed variable by its sample mean. Since both $[\log y_i]$ and $\hat{y}_i$ are zero when $y_i = 1$, most of the coefficients of second-order terms in the formulae for the degree of economies of scale and scope are eliminated when these measures are evaluated at the sample means. Thus, it has become standard to summarize regression results by reporting the magnitudes of interest at the sample mean of the variables in the regression equation. However, it is important to remember that this normalization in no way eliminates the substantial variation in, say, the degree of scale economies that may in fact be present over the range of output levels in the sample.

Using this approach, it is straightforward to show that, for both the translog and hybrid translog cost functions, the degree of multiproduct economies of scale

evaluated at the sample mean is given by $S = 1/[\sum \beta_i]$. In order to measure the extent of economies of scope or product specific economies of scale, it is necessary that the cost function specified be defined when one or more output levels are zero. Thus, these magnitudes cannot be measured from an estimated translog cost function. However, for the hybrid translog cost function, the degree of scale economies specific to output i evaluated at the sample mean, for example, is given by

$$S_i = \left\{\exp[\alpha_0] - \exp\left[\alpha_0 - \beta_i/\theta + \delta_{ii}/2\theta^2\right]\right\}/\{\alpha_i \cdot \exp[\alpha_0]\}.$$

Summarizing the results of the translog studies referred to above, the estimates of overall economies of scale, evaluated at sample means, range from 0.94 for Fuss and Waverman's (1981) hybrid translog to Nadiri and Shankerman's (1981) estimate of 2.12, with all studies except the former yielding estimates significantly greater than 1.[40] Since the calculation of the degrees of economies of scope or product specific economies of scale requires that the cost function in question be defined for zero output levels, only the Fuss and Waverman (1981) hybrid translog study could provide empirical tests for the presence or absence of these economies. They find no evidence of economies of scope in the operations of Bell Canada. They did find evidence of product specific increasing returns to scale for private line services at the sample mean. However, they determined that these increasing returns would be exhausted if Bell Canada were to serve the entire Canadian private line market in 1980. Thus, the strictures of Proposition 12 do not hold, and one *cannot* conclude that industry cost minimization requires that all private line services be provided by Bell Canada.

The two studies that yielded the upper and lower extremes of the estimates of the degree of overall scale economies also raise two important issues in empirical cost function studies: the treatment of technological change and the specification of the functional form to be estimated. The Nadiri and Shankerman (1981) study was focused on assessing technological change and the growth of total factor productivity in the Bell System. They obtained the highest estimate of the degree of overall scale economies. As mentioned above, the Fuss and Waverman study was the only one of the group employing the hybrid translog cost function. They obtained the lowest estimate of overall economies of scale.

Nadiri and Shankerman's study differed from the others under discussion in two primary respects: they used R&D expenditures to characterize technological change and they used U.S. rather than Canadian data. [Interestingly, the study of Christensen et al. (1983) shared these two characteristics and produced estimates of overall economies of scale nearly as large: 1.73.] While it may be the case that economies of scale are simply greater in the United States than in Canada, the

[40]See Fuss (1983, table 4, p. 19).

larger size of the U.S. telecommunications system would suggest exactly the opposite, i.e. that there should be more unexploited scale economies in Canada. Thus, it would be wise to give some weight to the possibility that the differing treatment of technological change played a determining role in the results.

Why is the specification and measurement of technological change such an important issue when attempting to estimate the degree of economies of scale experienced by a firm? In an industry (like telecommunications) that is experiencing *both* rapid technological change *and* demand growth, there are two effects that are lowering observed unit costs over time: technological advances that shift down the cost curve and output growth that moves the firm down along a falling average cost curve. When a cost function is to be estimated using time series data, it is important to separate the two effects in order to get an accurate estimate of the magnitude of either. For example, when output is growing, data that indicate a downward trend in unit costs could result from a downward shift over time of a constant average cost curve. Alternatively, such observations could be the result of output expansion over time down a stable falling average cost curve. Without careful measurement and specification of a measure of the state of technological progress, it is impossible to separate the two effects.

Even the beginning student of econometrics is continually reminded of the fact that any hypothesis tests or parameter estimates resulting from his analysis are conditional on the validity of the functional form specified for the estimating equation. That is why there has been considerable emphasis of late on the estimation of so-called flexible functional forms such as the translog which reduce to more restrictive functional forms (e.g. Cobb–Douglas) for certain parameter values. With respect to the empirical studies of the telecommunications industry under discussion, it should be remembered that, here, the translog is the restrictive functional form, being one limiting value of the hybrid translog. Fuss and Waverman tested the implied restriction and were able to reject it. It is interesting to note that their unrestricted estimate of the degree of overall scale economies of 0.92 is substantially below the next lowest estimate [Breslow and Smith's (1980) 1.29] and the modal estimate of about 1.45. However, when estimated over their sample, the translog specifications yields an estimate of the degree of overall economies of scale of 1.43, in line with the results of the other studies.

This evidence suggests that imposing the translog specification on the data may lead to an upward bias of the estimate of the degree of scale economies. That is the same conclusion reached by Guilkey and Lovell (1980) in their Monte Carlo study.

Throughout this discussion, the thoughtful reader will have noted that the policy issue motivating the discussion – i.e. whether or not telecommunications is a natural monopoly – has been addressed only tangentially. For the analysis

presented in Subsection 2.5 points out that overall economies of scale are neither necessary nor sufficient for a multiproduct industry to be a natural monopoly. Even when combined with evidence on economies of scope and product specific economies of scale, as in Fuss and Waverman, to reach definitive conclusions on the presence or absence of natural monopoly requires more than point estimates of such economies at the sample mean.

Evans and Heckman (1984) were the first to attack the problem directly. They test the subadditivity of a two output (toll and local) translog cost function estimated using time series data from the Bell System. The approach that they take is straightforward. A cost function is subadditive for some output level if it is cheaper to produce the quantity in question with a single firm than with any multifirm alternative. Thus, as noted in Subsection 2.5, in order to establish that a cost function is subadditive over some output region it is necessary to perform this calculation for each output level in the region and for each multifirm alternative industry structure. However, as Evans and Heckman recognize, to show that a cost function is *not* subadditive requires only that this test fails for *some* output level in the region for *one* multifirm alternative.

Evans and Heckman's strategy was to perform this test for all possible two firm alternatives to observed Bell System output levels that lie in their "admissible region". This region is defined by the intuitive notion that any output vector assigned to a firm in a hypothetical multifirm alternative must be at least as large (in both dimensions) as that of the smallest sample point and comprise a ratio of local to toll no smaller (larger) than the smallest (largest) actually observed in the sample used to estimate the cost function. This enabled them to avoid the problems caused by extrapolation outside the sample used to estimate the underlying cost function.

Their results can be summarized as follows. There were 20 data points, 1958–77, for which it was possible to find two firm alternatives in the admissible region. For *all* of these, Evans and Heckman found that there existed a two firm alternative industry configuration that would have resulted in a lowering of industry costs; often a statistically significant lowering.[41] These results enabled them to conclude, directly, that the *necessary* conditions for the subadditivity of the estimated Bell System cost function could not be satisfied, and that, therefore, the Bell System was not a natural monopoly. Note that Evans and Heckman were able to obtain their results even though the translog cost function they employed could not have been used to test for many of the *necessary* (economies of scope) or *sufficient* (economies of scope plus declining average incremental costs) conditions for cost subadditivity derived in Subsection 2.4. That is an important advantage of their direct approach when attempting to *disprove* subadditivity.

[41] See Evans and Heckman (1984, table 1).

5. Concluding remarks

The last decade has seen considerable advances in both theoretical and empirical work on the technological determinants of firm and industry structure. In particular, there has been a dramatic increase in the number of empirical studies that take explicit account of the intrinsic multiproduct nature of most real world industries. In addition to the papers discussed above, the Bibliography offers a (nonexhaustive) selection of empirical cost studies of hospitals, insurance firms, banks, airlines, railroads, motor carriers, automobile producers, to cite just a few examples.

Bibliography

Allen, B.T. (1983) 'Concentration, scale economies and the size distribution of plants', *Quarterly Review of Economics and Business*, 23(4):6–27.

Atkinson, S.E. and Halvorsen, R. (1984) 'Parametric efficiency test, economies of scale, and input demand in U.S. electric power generation', *International Economic Review*, 25(3):647–662.

Bailey, E.E. (1985) 'Airline deregulation in the United States: The benefits provided and the lessons learned', *International Journal of Transportation Economics*, 12(2):113–144.

Bailey, E.E. and Friedlaender, A.F. (1982) 'Market structure and multiproduct industries', *Journal of Economic Literature*, 20:1024–1041.

Bain, J. (1954) 'Economies of scale, concentration and entry', *American Economic Review*, 44:15–39.

Baumol, W.J. (1977) 'On the proper cost tests for natural monopoly in a multiproduct industry', *American Economic Review*, 67:43–57.

Baumol, W.J. and Fischer, D. (1978) 'Cost-minimizing number of firms and the determination of industry structure', *Quarterly Journal of Economics*, 92:439–467.

Baumol, W.J., Panzar, J.C. and Willig, R.D. (1982) *Contestable markets and the theory of industry structure*. New York: Harcourt Brace Jovanovich.

Beckenstein, A.R. (1975) 'Scale economies in the multiplant firm: Theory and empirical evidence', *Bell Journal of Economics*, 6:644–664.

Berechman, J. (1983) 'Costs, economies of scale, and factor demands in bus transport', *The Journal of Transport Economics and Policy*, 17(1):7–24.

Blackorby, C., Primont, D. and Russell, R. (1977) 'On testing separability restrictions with flexible functional forms', *Journal of Econometrics*, 5:195–209.

Braeutigam, R. and Daughety, A.F. (1983) 'On the estimation of returns to scale using variable cost functions', *Economic Letters*, 11:25–31.

Braeutigam, R.R. and Pauly, M.V. (1986) 'Cost function estimation and quality bias: The regulated automobile insurance industry', *Rand Journal of Economics*, 17:606–617.

Braeutigam, R.R., Daughety, A.F. and Turnquist, M.A. (1982) 'The estimation of a hybrid cost function for a railroad firm', *Review of Economics and Statistics*, 64:394–404.

Braeutigam, R.R., Daughety, A.F. and Turnquist, M.A. (1984) 'A firm specific analysis of economies of density in the U.S. railroad industry', *Journal of Industrial Economics*, 33:3–20.

Breslaw, J. and Smith, J. (1980) 'Efficiency, equity and regulation: An econometric model of Bell Canada', final report to the Canadian Department of Communications.

Bruning, R.E. and Olson, R.E. (1982) 'The use of efficiency indexes in testing for scale economies in the motor carrier industry', *Journal of Transport Economics and Policy*, 16(3):227–293.

Caves, D., Christensen, L. and Swanson, J. (1980) Productivity in U.S. railroads 1951–1974', *Bell Journal of Economics*, 11(1):166–181.

Caves, D.W., Christensen, L.R. and Swanson, J.A. (1981a) 'Economic performance in regulated

environments: A comparison of U.S. and Canadian railroads', *Quarterly Journal of Economics*, 96:559–581.
Caves, D.W., Christensen, L.R. and Swanson, J.A. (1981b) 'Productivity growth, scale economies and capacity utilization in U.S. railroads, 1955–74', *American Economic Review*, 71:994–1002.
Caves, D.W., Christensen, L.R. and Tretheway, M.W. (1980) 'Flexible cost functions for multiproduct firms', *Review of Economics and Statistics*, 62(3):477–481.
Caves, D.W., Christensen, L.R. and Tretheway, M.W. (1984) 'Economies of density versus economies of scale: Why trunk and local service airline costs differ', *Rand Journal of Economics*, 15:471–489.
Chappell, Jr., H.W. and Wilder, R.P. (1986) 'Multiproduct monopoly, regulation, and firm costs: Comment', *Southern Economic Journal*, 52(4):1168–1174.
Christensen, L. and Greene, W. (1976) 'Economies of scale in U.S. electric power generation', *Journal of Political Economy*, (4):655–676.
Christensen, L., Cummings, D. and Schoech, P. (1983) 'Econometric estimation of scale economies in telecommunications', in: L. Courville, A. de Fontenay and R. Dobell, eds., *Economic analysis of telecommunications: Theory and applications*. Amsterdam: North-Holland.
Clark, J.M. (1923) *Studies in the economics of overhead costs*. Chicago: University of Chicago Press.
Clemens, E. (1950–1951) 'Price discrimination and the multiproduct firm', *Review of Economic Studies*, 19(4):1–11; reprinted in: R. Heflebower and G. Stocking, eds., *Readings in industrial organization and public policy*. Homewood, Ill.: Irwin, 1958, 262–276.
Cowing, T.G. and Holtmann, A.G. (1983) 'Multiproduct short run hospital cost functions: Evidence and policy implications from cross-section data', *Southern Economic Journal*, 50:637–653.
Daly, M.J. and Rao, P.S. (1985) 'Productivity, scale economics, and technical change in Ontario Hydro', *Southern Economic Journal*, 52(1):167–180.
Daly, M.J., Rao, P.S. and Geehan, R. (1985) 'Productivity, scale economics and technical progress in the Canadian life insurance industry', *International Journal of Industrial Organization*, 3(3): 345–361.
De Borgen, B.L. (1984) 'Cost and productivity in regional bus transportation: The Belgian case', *Journal of Industrial Economics*, 33:37–54.
Demsetz, H. (1973) 'Joint supply and price discrimination', *Journal of Law and Economics*, 16(2):389–405.
Denny, M. and Fuss, M. (1977) 'The use of approximation analysis to test for separability and the existence of consistent aggregates', *American Economic Review*, 67:404–418.
Denny, M. and Fuss, M. (1980) 'The effects of factor prices and technological change on the occupational demand for labour: Evidence from Canadian telecommunications', Institute for Policy Analysis working paper no. 8014, University of Toronto.
Denny, M., Everson, C., Fuss, M. and Waverman, L. (1981) 'Estimating the effects of diffusion of technological innovations in telecommunications: The production structure of Bell Canada', *Canadian Journal of Economics*, 14:24–43.
Doherty, N.A. (1981) 'The measurement of output and economies of scale in property-liability insurance', *Journal of Risk and Insurance*, 48:390–402.
Evans, D.S., ed. (1983) *Breaking up Bell*. New York: North-Holland.
Evans, D.S. and Heckman, J.J. (1984) 'A test for subadditivity of the cost function with application to the Bell system', *American Economic Review*, 74:615–623.
Evans, D.S. and Heckman, J.J. (1986) 'A test for subadditivity of the cost function with application to the Bell system: Erratum', *American Economic Review*, 76:556–558.
Fare, R., Jansson, L. and Knox Lovell, C.A. (1985) 'Modelling scale economies with ray-homothetic production functions', *Review of Economics and Statistics*, 67:624–636.
Ferguson, C.E. (1969) *The neoclassical theory of production and distribution*. Cambridge: Cambridge University Press.
Friedlaender, A.F., Winston, C. and Wang, K. (1983) 'Costs, technology and productivity in the U.S. automobile industry', *Bell Journal of Economics*, 14:1–20.
Friedlaender, A.F. and Spady, R. (1981) *Freight transport regulation: Equity, efficiency and competition in the rail and trucking industries*. Cambridge, Mass.: MIT Press.
Fuss, M.A. (1983) 'A survey of recent results in the analysis of production conditions in telecommunications', in: L. Courville, A. de Fontenay and R. Dobell, eds., *Economic analysis of telecommunications: Theory and applications*. Amsterdam: North-Holland.

Fuss, M.A. and Waverman, L. (1981) 'The regulation of telecommunications in Canada', final report to the Economic Council of Canada.
Gillen, D.W. and Oum, T.H. (1984) 'A study of the cost structure of the Canadian intercity motor coach industry', *Canadian Journal of Economics*, 17(2):369–385.
Gilligan, T. and Smirlock, M. (1984) 'An empirical study of joint production and scale economies in commercial banking', *Journal of Banking and Finance*, 8:67–77.
Gilligan, T., Smirlock, M. and Marshall, W. (1984) 'Scale and scope economies in the multiproduct banking firm', *Journal of Monetary Economics*, 13:393–405.
Ginsberg, W. (1974) 'The multiplant firm with increasing returns to scale', *Journal of Economic Theory*, 9:283–292.
Gold, B. (1981) 'Changing perspectives on size, scale and returns: An interpretive survey', *Journal of Economic Literature*, 19:5–33.
Gort, M. and Wall, R.A. (1984) 'The effect of technical change on market structure', *Economic Inquiry*, 22(4): 668–675.
Guilkey, D. and Lovell, C.A.K. (1980) 'On the flexibility of the translog approximation', *International Economic Review*, 21:137–148.
Harmatuck, D.J. (1981) 'A motor carrier joint cost function', *The Journal of Transport Economics and Policy*, 15(2):135–153.
Harmatuck, D.J. (1985) 'Short run motor carrier cost functions for five large common carriers', *The Logistics and Transportation Review*, 21(3):217–237.
Harris, R.G. and Winston, C. (1983) 'Potential benefits of rail merger: An econometric analysis of network effects on service quality,' *Review of Economics and Statistics*, 65:32–40.
Hicks, J.R. (1935) 'Annual survey of economic theory: Monopoly', *Econometrica*, 3:1–20; reprinted in: G. Stigler and K. Boulding, eds., *Readings in price theory*. Chicago: Irwin, 1952, 361–383.
Hirshleifer, J. (1984) *Price Theory and Applications*, 3rd ed. Englewood Cliffs, N.J.: Prentice-Hall.
Kellner, S. and Matthewson, G.F. (1983) 'Entry, size distribution, scale, and scope economies in the life insurance industry', *Journal of Business*, 56(1):25–44.
Kim, H.Y. and Clark, R.M. (1983) 'Estimating multiproduct scale economies: An application to water supplies', U.S. Environmental Protection Agency, Municipal Environment Research Laboratory, Cincinnati, Ohio.
Kott, P.S. (1983) 'Return to scale in the U.S. life insurance industry: Comment', *Journal of Risk and Insurance*, 50(2):506–507.
Mayo, J.W. (1984) 'Multiproduct monopoly, regulation, and firm costs', *Southern Economic Journal*, 51(1):208–218.
Mayo, J.W. (1984) 'The technological determinants of the U.S. energy industry structure', *Review of Economics and Statistics*, 66:51–58.
McFadden, D. (1978) 'Cost, revenue and profit functions', in: M.A. Fuss and D. McFadden, eds., *Production economics: A dual approach to theory and applications*. Amsterdam: North-Holland.
Murray, J.D. and White, R.W. (1983) 'Economies of scale and economies of scope in multiproduct financial institutions: A study of British Columbia credit unions', *Journal of Finance*, 38:887–901.
Nadiri, M.I. and Shankerman, M. (1981) 'The structure of production, technological change and the rate of growth of total factor productivity in the Bell system,' in: T. Cowing and R. Stevenson, eds., *Productivity measurement in regulated industries*. New York: Academic Press.
Nerlove, M. (1963) 'Returns to scale in electricity supply', in: C. Christ, ed., *Measurement in economics: Studies in mathematical economics and econometrics in memory of Yehuda Grunfeld*. Stanford: Stanford University Press.
Pagano, A.M. and McKnight, C.E. (1983) 'Economies of scale in the taxicab industry: Some empirical evidence from the United States', *The Journal of Transport Economics and Policy*, 17(3):299–313.
Panzar, J.C. (1976) 'A neoclassical approach to peak load pricing', *Bell Journal of Economics*, 7:521–530.
Panzar, J.C. and Willig, R.D. (1975) 'Economies of scale and economies of scope in multi-output productions', Bell Laboratories economic discussion paper no. 33.
Panzar, J.C. and Willig, R.D. (1981) 'Economies of scope', *American Economic Review*, 71(2):268–272.
Praetz, P. (1980) 'Returns to scale in the United States life insurance industry', *The Journal of Risk and Insurance*, 47:525–533.

Praetz, P. (1983) 'Returns to scale in the United States life insurance industry: Reply', *The Journal of Risk and Insurance*, 50(2):508–509.

Praetz, P. and Beattie, M. (1982) 'Economies of scale in the Australian general insurance industry', *Australian Journal of Management*, 7(2):117–124.

Sakai, Y. (1974) 'Substitution and expansion effects in production theory: The case of joint production', *Journal of Economic Theory*, 9(3):255–274.

Scherer, F.M. (1980) *Industrial market structure and economic performance*, 2nd ed. Chicago: Rand McNally.

Scherer, F.M., Beckenstein, A.R., Kaufer, E. and Murphy, R.D. (1975) *The economics of multiplant operation: An international comparison study*. Cambridge, Mass.: Harvard University Press.

Sharkey, W.W. (1982) *The theory of natural monopoly*. Cambridge U.K.: Cambridge University Press.

Sickles, R.C. (1985) 'A nonlinear multivariate error components analysis of technology and specific factor productivity growth with an application to the U.S. airlines', *Journal of Econometrics*, 27:61–78.

Skogh, G. (1982) 'Returns to scale in the Swedish property-liability insurance industry', *The Journal of Risk and Insurance*, 49(2):218–228.

Spady, R. (1985) 'Using indexed quadratic cost functions to model network technologies', in: A.F. Daughety, ed., *Analytical studies in transport economics*. Cambridge, U.K.: Cambridge University Press.

Spady, R. and Friedlaender, A.F. (1978) 'Hedonic cost functions for the regulated trucking industry', *Bell Journal of Economics*, 9(1):159–179.

Sueyoshi, T. and Anselmo, P.C. (1986) 'The Evans and Heckman subadditivity test: Comment', *American Economic Review*, 76:854–855.

Talley, W.K., Agarwal, V.B. and Breakfield, J.W. (1986) 'Economies of density of ocean tanker ships', *The Journal of Transport Economics and Policy*, 20(1):91–99.

Tauchen, H., Fravel, F.D. and Gilbert, G. (1983) 'Cost structure and the intercity bus industry', *The Journal of Transport Economics and Policy*, 17(1):25–47.

Teece, D.J. (1980) 'Economics of scope and the scope of the enterprise', *Journal of Economic Behavior and Organization*, 1:223–247.

Varian, H. (1984) *Microeconomic analysis*. New York: Norton.

Victor, P.A. (1981) 'A translog cost function for urban bus transit', *Journal of Industrial Economics*, 23:287–304.

Wang Chiang, S.J. and Friedlaender, A.F. (1984) 'Output aggregation, network effects, and the measurement of trucking technology', *Review of Economics and Statistics*, 66:267–276.

Wang Chiang, J.S. and Friedlaender, A.F. (1985) 'Truck technology and efficient market structure', 67:250–258.

White, L.J., (1979) 'Economics of scale and the question of "natural monopoly" in the airline industry', *Journal of Air Law and Commerce*, 44:545–573.

Winston, C., et al. (1987) *Blind intersection? Policy and automobile industry*. The Brookings Institution, Washington D.C.