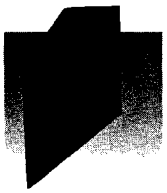
Productive efficiency in banking

Douglas D. Evanoff and
Philip R. Israilevich



Then a new CEO came in who asked, ... "What do we have to produce by way of results?" Every one of his store managers knew the answer, "We have to increase the amount each shopper spends per visit." Then he asked, "Do any of our stores actually do this?" Three or four—out of 30 or so—did it. "Will you then tell us," the new CEO asked, "what your people do that gives you the desired results?"[1]

Introduction

In the above epigraph the managers are attempting to identify, in a particular context, the firms which are doing the best job of accomplishing the company objectives. Such firms are known as the best practice firms. Economists typically make similar inquiries concerning the production process. They address the issue by theoretically defining the best practice firm, empirically identifying it, determining its resource utilization, and then evaluating how others compare to it. More generally, economists, like the new CEO, are concerned with productive efficiency.

Because of changes taking place in the banking industry, the importance of efficiency has increased substantially. As geographic and product deregulation occurs, the resulting increase in competition should place banks in a situation where their success will depend on their ability to adapt and operate efficiently in the new environment. Banks unable to do so will have difficulty surviving.

Most studies of bank efficiency have concentrated on cost advantages resulting from the scale of production. In fact, this is probably one of the most researched topics in banking.[2] There are, however, other aspects of efficiency which students of the industry have just begun to evaluate. For example, do the producers of banking services effectively combine their productive inputs? Once employed, do they use the inputs effectively? If not, how inefficient are they? What allows them to continue to do this and stay in business? Given its importance in the deregulated environment, it is imperative that the various aspects of bank efficiency be understood and empirically analyzed.

In this article we discuss the concept of efficiency in production, define its various aspects and the means to measure it, and review the relevant literature concerning inefficiency in the banking industry. Our major conclusion is that there appears to be significant inefficiency in banking. Inefficiency resulting from operating at an inappropriate scale of operation is probably in the range of 10-20 percent of costs. However, by emphasizing the role of scale, researchers have essentially overlooked a major portion of bank inefficiency. The evidence suggests that inefficiencies resulting

The authors are economists at the Federal Reserve Bank of Chicago. Helpful comments on earlier drafts by Herb Baer, Paul Bauer, Allen Berger, Dave Humphrey, Curt Hunter, Carl Pasurka, and Sherrill Shaffer are gratefully acknowledged. The views expressed, however, are those of the authors and are not necessarily shared by others.

from the suboptimal utilization of inputs is larger than that resulting from other factors. According to a majority of studies, banks operate relatively efficiently with respect to the optimal combination of inputs, yet many are very inefficient in converting these inputs into outputs. This inefficient utilization of inputs accounts for an additional 20-30 percent of costs. This is particularly interesting because it implies that, to a great extent, the future viability of an individual bank is under its own control. To the extent that bank inefficiency can be accurately measured, it appears that the largest inefficiencies are not the result of regulation or technology, but result directly from an underutilization of factor inputs by bank management. This inefficiency will most likely decline in the future as bankers respond to increased competitive pressures and strive to become more efficient. Failing this, the inefficient firms will become prime merger candidates to be acquired and restructured.

The article proceeds as follows. In the next section we define, discuss, and illustrate the components of production efficiency. We then evaluate the alternative means to generate measures of efficiency. A review of the literature on bank efficiency is then presented. The final section summarizes and evaluates policy concerns. We have also included an extensive reference list for readers interested in more detailed analysis of productive efficiency.

Production efficiency

The economic theory of the firm assumes that production takes place in an environment in which managers attempt to maximize profits by operating in the most efficient manner possible. The competitive model suggests that firms which fail to do so will be driven from the market by more efficient ones. However, when natural entry barriers or regulation weaken competitive forces, inefficient firms may continue to prosper. That is, true firm behavior may vary from that implied by the competitive model as managers attempt to maximize their own well-being instead of profits, or find that they are not required to operate very efficiently to remain in business.

Variations from productive efficiency can be broken down into input and output induced inefficiencies. By input inefficiency we mean that, for a given level of output, the firm is not optimally using the factors of production . *Overall input inefficiency* resulting from the suboptimal use of inputs can be decomposed into *allocative* and *pure technical* inefficiency. Allocative inefficiency occurs when inputs are combined in sub-optimal proportions. Regulation is typically given as a major reason for this occurrence. Pure technical inefficiency occurs when more of each input is used than should be required to produce a given level of output. This occurrence is more difficult to explain, but is typically attributed to weak competitive forces which allow management to "get away" with slackened productivity. Combining these two notions of inefficiency we get the overall inefficiency resulting from the improper use of inputs.[3] The distinction between the two types of inefficiency is important because they may be caused by totally different forces.

Productive efficiency requires optimizing behavior with respect to outputs as well as inputs. With respect to outputs, optimal behavior necessitates production of the level and combination of outputs corresponding to the lowest per unit cost production process. An optimal output level is possible if economies and diseconomies of scale exist at different output levels. Economies of scale exist if, over a given range of output, per unit costs decline as output increases. Increases in per unit cost correspond to decreasing returns to scale. A *scale efficient* firm will produce where there are constant returns to scale; that is, changes in output result in proportional changes in costs. Because it involves the choice of an inefficient level, scale inefficiency is considered a form of technical inefficiency. Thus total technical inefficiency includes both pure technical and scale inefficiency; that is, inefficient levels of both inputs and outputs.

Additional cost advantages may result from producing more than one product. For example, a firm may be able to jointly produce two or more outputs more cheaply than producing them separately. If the cost of joint production is less than the cost resulting from independent production processes, *economies of scope* are said to exist. Diseconomies of scope exist if the joint production costs are actually higher than specialized or stand-alone production of the individual products.

A final point should be mentioned concerning the various categories of inefficiency. Pure technical inefficiency is entirely under the control of, and results directly because of, the

behavior of the producer. Output inefficiency and allocative inefficiency may, from the perspective of the firm, be unavoidable. For example, a firm optimally using factor inputs may find that per unit cost declines over the entire range of market demand. While increasing production would generate cost savings or efficiencies, the characteristics of market demand may not justify it. Failure to exploit scope advantages may also result from factors outside of the control of the firm. In banking, the array of allowable activities is obviously constrained by regulation. This may preclude potential gains from the joint production of various financial services. Finally, as mentioned earlier, allocative inefficiency may occur as a direct result of regulation. For example, during the 1970s, banks were restricted with respect to the explicit rates they could pay depositors. As market rates rose above allowable levels, banks frequently substituted implicit interest payments in the form of improved service levels; for example, more offices per capita or per area, see Evanoff (1988). This resulted in an over-utilization of physical capital relative to other factor inputs. In this case, regulation was the driving force behind the resulting allocative inefficiency. The point is that much inefficiency *may* be beyond the control of the individual firm.

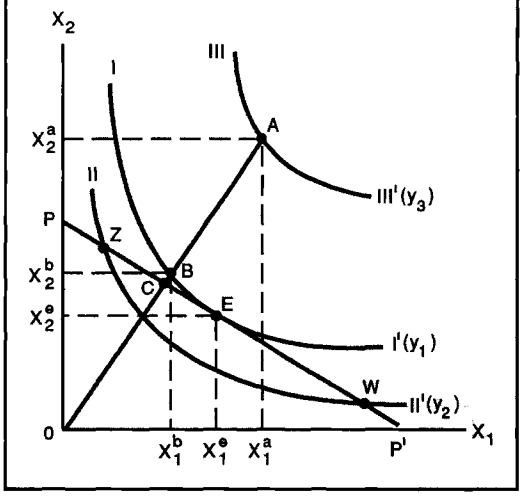
In the following sections we illustrate the inefficiencies described above and discuss alternative methods used to empirically capture them. The reader who is most interested in an analysis of efficiency in banking may skip directly to the section entitled "The role of production inefficiency in banking: A survey of the literature."

Illustrating input efficiency

The notions of input inefficiencies can be illustrated as shown in Figure 1. Assume that x_1 and x_2 are two factor inputs required to produce a single output, y. Isoquant *I-I'* depicts various efficient combinations of the two inputs which can be used to produce a specific level of output, y_1. Isoquants further to the right correspond to higher levels of output, those to the left to lower levels of output. For example, the output level associated with isoquant *II-II'* is less than y_1. For a given set of input prices, the isocost line, *P-P'*, represents the various combinations of inputs which generate the same level of expenditures. Isocost lines further to the right

FIGURE 1

Input efficiency



correspond to higher level of expenditures on inputs. The slope of the isocost line is, obviously, determined by input prices.

If the objective of the producer is to produce a particular level of output at minimum cost then the optimal input combination in Figure 1 is at point *E*. That is, given factor prices, output y_1 can be optimally produced by employing x_1^e units of input x_1 and x_2^e units of input x_2. Any other combination of the inputs along the *P-P'* isocost line would generate less output for the same cost. For example, the input combinations corresponding to points *W* or *Z* would result in similar expenditures on inputs, but generate the lower level of output associated with isoquant *II-II'*. Alternatively, the production of y_1 using any combination of inputs other than that corresponding to point *E* would cost more. Therefore, at point *E*, input efficiency exists.[4]

To illustrate input inefficiency, suppose that the *observed* combination of inputs used by a particular firm to produce y_1 is at point *A* in Figure 1. We know that inefficiency exists because *E* was shown above to correspond to the most efficient combination of inputs to produce y_1. Comparing the input utilization at point *A* to that at *E* we can derive the level of inefficiency resulting from the suboptimal use of inputs. In order to illustrate allocative and pure technical inefficiency, we have drawn a line from the origin to point *A*. Along this line different levels of factor inputs are employed but the ratio between the two inputs is fixed at

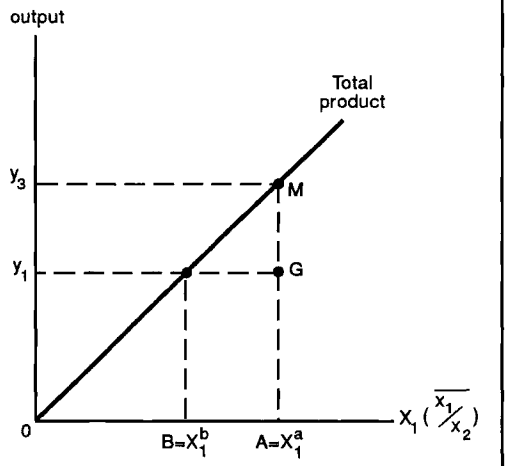
the actual ratio (that is, the ratio at point A). Reference points along this line and on isoquant I-I' and isocost line P-P' are highlighted. Consider allocative inefficiency first. Point C represents a level of costs equal to that of the efficient production process at point E because it is on line P-P'. Point B corresponds to an output level equal to y_1 because it is on isoquant I-I'. Therefore, the distance CB corresponds to additional production expenses resulting from the suboptimal allocation of inputs. That is, allocative inefficiency exists because we are not on the isocost line, P-P'. Formally, OC/OB is a measure of allocative efficiency. Values less than 1.0 reflect inefficiency.[5]

For this same example, we can also depict pure technical inefficiency resulting from producing at point A. We have seen that producing y_1 using x_1^a and x_2^a involves allocative inefficiency because point A is to the right of line P-P' and ray $0A$ does not go through point E. However, there is additional inefficiency because point A is above isoquant I-I'. That is, the combination of inputs associated with point A should enable the firm to produce a level of output greater than y_1. (It should be able to produce output y_3 corresponding to isoquant III-III'.) Given that the isocost line depicts total expenditures used in production, distance CA constitutes a less than optimal usage of all inputs and corresponds to additional production expenses. Therefore, *overall input inefficiency* is measured as OC/OA. Because OC/OB is attributed to allocative inefficiency, the remaining portion, OB/OA, can be attributed to pure technical inefficiency. Since these are radial measures, overall input inefficiency is the product of the two subcomponents, that is, $OC/OA = (OC/OB) \cdot (OB/OA)$.

The pure technical inefficiency shown in Figure 1 can also be illustrated in terms of output, instead of input, using a total output or total product relationship as depicted in Figure 2. The ratio of input usage, x_1/x_2, is held fixed by assumption in Figure 2 to represent input combinations along the ray OA in Figure 1. Since the fixed input ratio precludes the analysis of allocative efficiency, we are analyzing only pure technical efficiency. Because changes in inputs result in proportional changes in output (the total product curve is linear) we have constant returns to scale as was assumed in Figure 1. Employing x_1^b units of input x_1 we

FIGURE 2

Pure technical efficiency measured in terms of outputs



could produce an output level y_1 *if* the inputs were fully utilized. This corresponds to point B in Figure 1. Similarly, using x_1^a units of input x_1 we should be able to produce y_3. Again, this corresponds to point A in Figure 1. However, if inputs are not used effectively, that is if technical inefficiency exists, the resulting production point will be below the total product curve. That is, pure technical inefficiency occurs when we operate beneath the total product relationship. For example, the pure technical inefficiency depicted in Figure 1 corresponds to that found at point G in Figure 2, where inputs are under-utilized and x_1^a only generates an output level of y_1. If we are producing y_1 at point A in Figure 1 or, equivalently, at point G in Figure 2, pure technical inefficiency is measured with respect to inputs as OB/OA and with respect to outputs as AG/AM. The inefficiency measures are equivalent. This illustration is important because it indicates that technical inefficiency can be measured in terms of either inputs or outputs. Below we drop the constant returns to scale assumption and expand on this output inefficiency measure.

Illustrating output efficiency

Point E in Figure 1 corresponds to the least cost, most efficient means to produce y_1. However, because of particular characteristics of the production technology, this level of output may not be the optimal one to produce. For example, it may be that over a certain range of outputs, economies of scale exist. Production

efficiency, therefore, requires optimal decisions concerning both input and output levels. In Figure 3 we have dropped the assumption of constant returns to scale. The production process is now characterized by increasing returns up to point R, constant returns at R, and decreasing returns at output levels above R. Now the firm corresponding to point G in Figure 3 is technically inefficient for two reasons. First, there is pure technical inefficiency resulting from the under-utilization of inputs; that is, we are beneath the total product curve. If inputs are fully utilized, input x_1^a should produce the higher output level corresponding to point M, that is, y_3. Second, we have decreasing returns to scale at the current level of output since the production process is not represented as the linear relationship OH. The output not produced because of scale inefficiency can be measured as HM. This output is what could have been produced if inputs were used efficiently *and* constant returns to scale existed at this output level. Therefore, for the input usage depicted at point A, the input efficient firm could produce at point M, and the input *and* scale efficient firm could produce at point H. As explained above, scale inefficiency is generally considered a form of technical inefficiency because it involves the choice of an inefficient level. Thus, *total technical inefficiency* includes pure technical and scale inefficiency; that is, inefficiency in the use of both inputs and outputs.

FIGURE 3

Technical efficiency with varying returns to scale

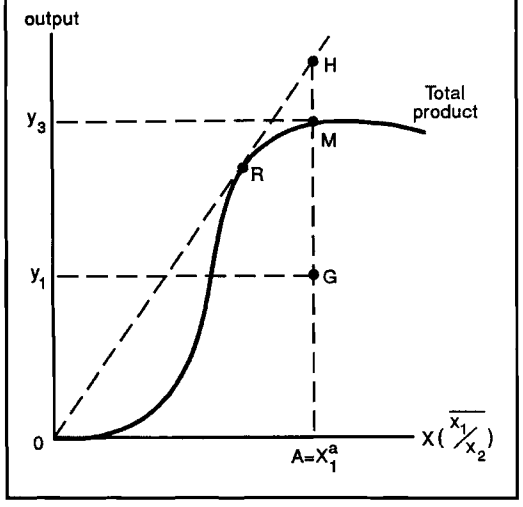
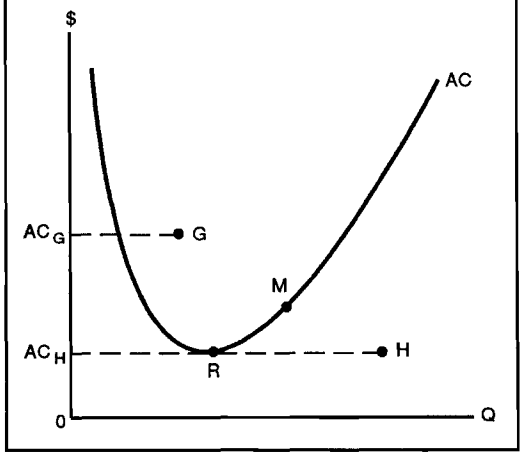


Figure 4 depicts the reference points just discussed in Figure 3 in terms of production cost. The total product relationship in Figure 3 corresponds to the average cost relationship depicted in Figure 4. Points H and R each correspond to constant returns to scale and, therefore, correspond to minimum points on average cost relationships. Total technical

FIGURE 4

Pure technical efficiency in terms of costs



inefficiency can be depicted here as the ratio of the average costs. For the example just discussed, total technical inefficiency is equal to AC_H/AC_G. The alternative measures of inefficiency illustrated in Figures 1 through 4 are equivalent and correspond to the alternative means of calculating inefficiency estimates commonly cited in the literature.

In the above discussion we assumed the production of a single output for illustrative purposes. Additional cost advantages may result from multiproduct production. For example, economies may exist for the joint production of two or more outputs, relative to the stand-alone production of the individual products. That is, scope advantages may exist. More formally, economies of scope exist in the joint production of Q_1 and Q_2 if

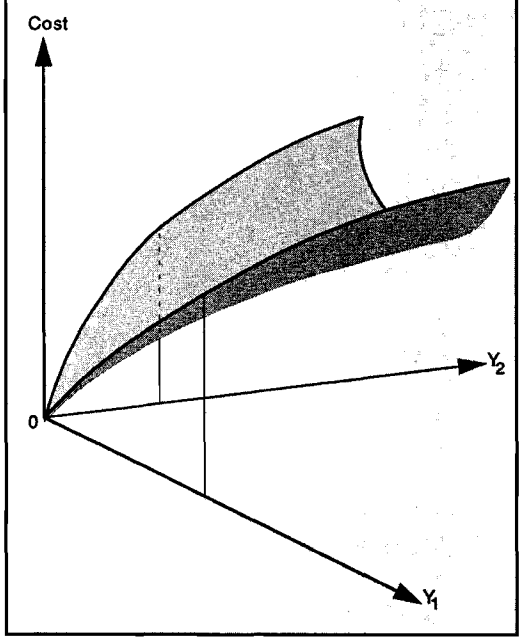
$$(1)\quad [C_1 + C_2] > C_{12}$$

where C_1 and C_2 are the cost of producing Q_1 and Q_2 independently (that is, as stand-alone processes), and C_{12} is the cost of joint production. With multiproduct production, some

fixed cost of production can be spread across the outputs and there may be synergies when the two products are produced jointly. A multiproduct cost relationship which exhibits production synergies between the two outputs, y_1 and y_2, is illustrated in Figure 5. Joint production moves the cost off the "lip" of the relationship onto the inner surface. Potential cost gains obviously exist.

FIGURE 5

A multiproduct cost relationship



Measuring production inefficiencies

The relationships depicted in the above figures, as well as all standard textbook presentations of the production process, present extreme values; that is, the *maximum* output that can be produced from a given set of inputs, or the minimum cost required to produce a given level of output. However, when attempts are made to generate estimates of the production process we typically abstract from the extreme values. The traditional approach to evaluating the production process is to assume the standard competitive model is appropriate and to estimate an average production, cost, or profit function.[6] Realizing that this restrictive model may not adequately describe the production process (and definitely avoids efficiency issues), methods have been developed which allow for variations in this approach. We discuss these variations in this section. The methodologies differ from each other in a number of ways, not the least of which is a result of differences in assumptions imposed in the analysis. The restrictiveness of these assumptions is determined by the individual data sets. Each of the methods discussed here is superior to the basic competitive model *as long as the assumptions employed are correct.* More will be said about this later.

While the concept of firm efficiency is rather straightforward, various difficulties are encountered when attempting to measure it. Essentially, one needs to derive the *best practice firm*, or the *production frontier* which depicts the maximum performance possible by firms, and contrast existing firms to this standard. Ideally, we would compare firm performance to the *true* frontier, however, the best that can be achieved is an empirical frontier or best practice firm generated from the observed data. Once the best practice firm is established, input related pure technical and allocative efficiency, and output related scale and scope efficiency, can be analyzed. For example, assuming constant returns to scale in Figure 1, all firms can be compared to one producing at point *E*.

Differences in estimates of firm efficiency typically result from different means of generating the best practice firm. There are two general approaches used to model this relationship. First, the parametric or econometric approach employs statistical methods to estimate an efficient cost frontier. Second, the nonparametric or deterministic approach is based on the linear programming approach for optimal allocation of resources called data envelopment analysis (DEA). This technique is used to directly generate individual frontiers for each firm. Below we discuss alternative methodologies within these two broad categories.

It should be emphasized that empirical measures of inefficiency are no different from estimated parameters in any economic model. The model may mistakenly reflect measurement errors or specification errors for productive inefficiency. As the literature on banking develops, more comprehensive models should be analyzed.

Parametric approach: Shadow price models

To generate estimates of allocative efficiency, one can use the parametric approach developed by Lau and Yotopoulos (1971) and refined by Atkinson and Halvorsen (1980, 1984).[7] This method assumes that firms are combining the factor inputs correctly, but that the combination is not necessarily based on observed prices. Rather, there are factors in addition to explicit market prices which enter the firm's employment decision process. These additional factors are combined with the explicit prices to generate *shadow prices* which are more comprehensive and which determine factor utilization. These additional factors typically include distortions induced by unionism, regulation, or managerial goals other than profit maximization. These alternative goals may include profit satisficing or expense preference behavior.[8]

More formally, a basic contention of economic theory is that, in competitive markets, the optimal level of employment for each factor of production can be determined by employing additional units until the last dollar spent on each factor yields the same amount of productivity. That is,

$$(2)\quad \frac{f_i}{P_i}=\frac{f_j}{P_j},\ \text{for } i \neq j = 1, \ldots, m,$$

where $f_i \equiv \partial f / \partial X_i$ is the marginal product of input i, and P_i is the price of input i, or

$$(3)\quad \frac{f_i}{f_j}=\frac{P_i}{P_j},\ \text{for } i \neq j = 1, \ldots, m,$$

where f_i / f_j is the marginal rate of technical substitution between the inputs. This relationship corresponds to the tangency of the isoquant and the isocost curve (point E) in Figure 1.

Given input prices and the predetermined level of output as the only constraint, the optimal combination of inputs, as in Equation (3), can be derived to minimize cost. However, if additional constraints exist (for example, regulatory constraints), they need to be accounted for and incorporated into the optimization process. Concerning the employment decision, Equation (3) becomes

$$(4)\quad \frac{f_i}{f_j}=\frac{P_i^*}{P_j^*},\ \text{for } i \neq j = 1, \ldots, m,$$

where P_i^* is the effective or shadow price of input i, and the marginal rate of technical substitution between the inputs is set equal to the ratio of the shadow prices of the inputs. Given competitive markets and the absence of additional binding constraints, shadow and actual prices are equal and the employment decision is not affected.

Because the shadow prices of the inputs are not directly observable, Lau and Yotopolous developed a means to estimate them along with other parameters of the cost relationship. Assuming shadow prices are proportional to market prices, shadow prices can be approximated by

$$(5)\quad P_i^* = k_i P_i,\ \text{for } i = 1, \ldots, m,$$

where k_i is input-specific.[9] Again, if the additional constraints are not binding, all shadow prices equal the respective market prices, that is, $k_i = 1$ for all i.

Standard econometric techniques can then be used to generate cost estimates employing the additional information. That is, the standard cost structure

$$(6)\quad C = C(P, Q, Z),$$

where C depicts costs, Q outputs, P explicit factor prices, and Z additional pertinent exogenous variables, is replaced with

$$(7)\quad C^s = C^s(kP, Q, Z),$$

where kP denotes shadow factor prices, and k is estimated along with the other parameters in the cost function.[10]

The shadow price model also allows one to calculate the optimal (unobserved) input combination given observed prices, P. This combination is relevant for measuring the cost differences resulting from production under competitive conditions and those when additional binding constraints exist. In the banking industry, these additional constraints are typically thought to be regulatory induced. The cost differences can be determined by contrasting costs when market prices equal shadow

prices ($k = 1$) to that found using the estimated shadow prices ($k = \hat{k}$ where $\hat{k}$ denotes the estimated value for k). The difference between the two cost values will be the result of combining inputs in a suboptimal manner.

Estimation of the cost function will yield k_i values which can be considered to reflect the effect of binding constraints on average. Ideally, the k_i measure would be firm specific. However, statistical problems typically make this prohibitive in terms of the degrees of freedom required for the estimation procedure. All the parametric approaches cited below have this same shortcoming. Some progress toward resolving this shortcoming has recently been made; see Evanoff and Israilevich (1991, 1990a).

One of the advantages of the shadow price model approach is that it allows for the estimation of returns to scale and scope along with allocative efficiency. However, pure technical efficiencies can not be measured by this approach although, as discussed later, this shortcoming can also be partially resolved.

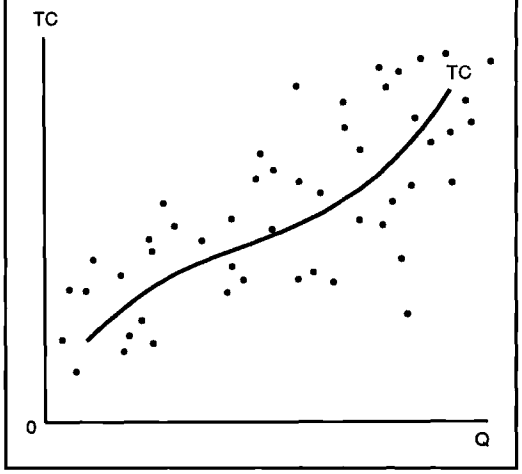
Parametric approach: Stochastic cost frontiers

Another more comprehensive parametric approach for measuring efficiency is to use stochastic frontier models. With this approach, the cost frontier is empirically estimated and firm specific deviations from the frontier are attributed to productive inefficiencies. A number of alternative parametric techniques can be used to generate the frontier. The major difference between these techniques is in the maintained assumptions which, obviously, can produce significantly different results. The restrictiveness of these assumptions is determined by the individual data sets. Here we summarize alternative parametric methods used to develop the frontier.

Using a parametric approach, the standard cost structure is typically generated by imposing a specific functional form on the data and obtaining the "best fit" by minimizing deviations from the estimated structure. For example, the estimated total cost relationship may be fitted to the data to produce a relationship such as *TC* in Figure 6. However, when evaluating efficiency, we are interested in the *best practice firm* or the cost frontier. We are not interested in the average relationship, rather we are looking for a minima in the data. Therefore,

FIGURE 6

Total cost relationship



adjustments to the standard estimation procedures are required. Typically the standard parametric procedure is adjusted by employing a more complex error structure. A "composed" error can be used which consists of two components: one is the standard statistical noise which is randomly distributed about the relationship, and the other consists entirely of positive deviations from the cost structure (that is, a one-sided disturbance term) and represents inefficiency.[11] Stated crudely, the resulting frontier is simply a transformation of *TC* in Figure 6 (shifted downward) to generate the best cost relationship instead of the average relationship.

For example, and more formally, assume a stochastic frontier model which consists of the following cost and share equations:

$$(8)\quad \ln C_h^A = \ln C_h^F + \ln T_h + \ln A + u_h;$$

$$(9)\quad M_{ih}^A = M_{ih}^F + \underbrace{b_i + u_{ih}}_{B_{ih}},\ \text{for } i = 1, \ldots, m;$$

where *ln* denotes the natural log, and C_h^A and M_{ih}^A are *observed* cost and factor shares for firm *h*. C_h^F is the lowest production cost relationship or the cost frontier, lnT_h reflects additions to cost resulting from pure technical inefficiency, *lnA* reflects additions to cost resulting from allocative inefficiency, and u_h is a random error. M_{ih}^F is the efficient share equation, b_i depicts share distortions resulting from allocative inefficiency, u_{ih} captures random distortions from effi-

cient shares, and B_{ih} is the composed error term.

Measures of technical inefficiency are calculated as firm specific deviations from the frontier and are derived from the additional error term discussed above. Since technical inefficiency can result only in increases to total cost, this error structure must consist entirely of non-negative values. That is, this component of the error structure is one-sided relative to the frontier. Choice of a specific one-sided distribution could obviously influence the empirical results.[12]

As with the shadow price model, allocative inefficiency is computed as an average for the sample and is not firm specific. *lnA* is non-negative as deviations from use of the optimal combination of inputs can lead only to additions to cost. However, b_i can be positive or negative suggesting over- or under-utilization of a particular input.

Obviously, *lnA* and b_i are related because suboptimal combinations of factor inputs ($b_i \neq 0$) result in additions to cost. However, empirically modeling this relationship is problematic. One standard means to do it is to impose restrictions on the relationship reflecting prior knowledge. For example, assuming increased costs occur only when mistakes are made ($A = 0$ only when $b_i = 0$), and that large mistakes cost more than small ones, one can impose a relationship between allocative mistakes and cost increases:[13]

$$(10)\quad \ln A = b' F b;$$

where F is a diagonal matrix with nonnegative elements. Positive elements of F represent weights for each b_i. For example, f_{ii} represents the relative effect of allocative distortions from factor i on the increased production costs. To summarize, the additional cost of allocative inefficiency is a weighted sum of squared mistakes from the misallocation of each input. The (nonnegative) weights are additional parameters to be estimated.

An alternative approach to generate a cost frontier is to utilize a cost structure consisting of cost and share equations, but to sever the link between the error terms of the cost and share equations. That is, the share equations are used only for efficiency gains in parameter estimation; not to link suboptimal combinations of inputs to increases in cost. Under this approach, both allocative and technical inefficiencies are depicted as one-sided errors from the cost frontier. Therefore, the estimated system of Equations 8 and 9 becomes

$$(11)\quad \ln C_h^A = \ln C_h^F + v_h + u_h,$$

$$(12)\quad M_{ih}^A = M_{ih}^F + u_{ih},\ \text{for } i = 1, \ldots, m,$$

where the error term depicting inefficiency, v_h, can be decomposed into its two components (that is, $lnT + lnA$) using techniques developed by Kopp and Diewert (1982) and refined by Zieschang (1983). This approach essentially ignores information concerning the relationship between disturbances in the cost and share equations, but is easier to work with than the above approach and does not necessarily generate results inferior to more complicated linkage approaches. This is particularly true if the more complicated approach, which is typically based on a set of untested assumptions, incorrectly models the linkage.

This attempt to simplify the methodology brings us to the most recent approach introduced by Berger and Humphrey (1990). These authors take the view that the preceding methodologies impose rather restrictive *ad hoc* assumptions concerning the data, the validity of which are questionable. For example, the assumed linkage between the error structure in the share cost equations, discussed above, could be inaccurate as could the assumptions concerning the one-sided error distribution. To partially remedy these problems the authors developed a "thick frontier" approach. Instead of imposing restrictive characteristics on the cost relationship to generate a true frontier or frontier *edge*, a thick frontier is estimated from a subsample of the data which, based on *a priori* information, is considered to be an efficient subgroup. This group is then compared to another group which, based on *a priori* information, is considered an inefficient subgroup. Therefore the authors are able to relax the restrictive assumptions employed in the methodologies discussed above, but at the cost of using a somewhat *ad hoc* means to categorize the data into efficient and inefficient groups.

This approach was implemented using banking sector data by assuming subgroups could be delineated based on their average cost

per dollar of output. The data were then stratified by size and divided into quartiles and the lowest and highest cost quartiles were contrasted. After accounting for differences resulting from market characteristics, the remaining differences between the two groups were assumed to constitute inefficiency. This can be distributed into its allocative and technical components using procedures similar to those of Kopp and Diewert discussed above. Obviously, this approach lacks precision and also imposes some rather *ad hoc* assumptions to develop the subgroups and produce the frontier. However, the assumptions may be less restrictive than those made in the more elaborate models discussed above. In fact, some of the maintained assumptions in these models were statistically tested and rejected by Berger and Humphrey. As a result, the relatively easy-to-implement approach may perform quite well in generating a rough measure of the extent of production inefficiency in an industry.[14]

Nonparametric approach

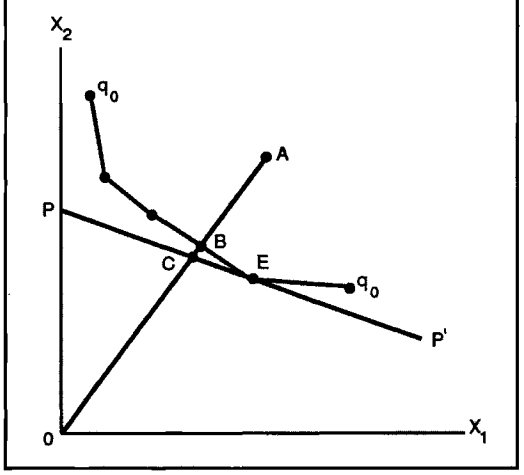
While intuitively appealing, and somewhat similar to the procedures commonly used to estimate standard cost relationships, the parametric approaches have been criticized for requiring more information than is typically available for estimation of the cost frontier. In an attempt to decrease the required information, some have chosen to use a nonparametric, linear programming approach known as data envelopment analysis (DEA).

Although there are various permutations to the DEA approach, the basic objective is to "envelop" the data by producing a piecewise linear fit via linear programming techniques. That is, instead of using regression techniques to fit a smooth relationship, a piecewise linear surface is produced which borders the observations, for example, the broken line q_o-q_o in Figure 7. The technique identifies observations for which the firm is producing a given level of output with the fewest inputs. These will be observations on the frontier. All other observations will be given an efficiency measure based on the distance from the frontier and indicating the extent to which inputs are being effectively utilized. This is comparable to the measure of pure technical inefficiency, *OB/OA*, for observation *A* in Figure 1.

The technique allows for the derivation of a frontier for each firm in the sample based on

FIGURE 7

DEA measured efficiency



the output and input utilization of all firms in the sample. As a simple example for the two input, one output case, the linear programming problem for technical inefficiency could be written as

$$(13)\quad \text{Min } \Theta^A,$$

$$\text{subject to } q^A \leq \mu^1 q^1 + \mu^2 q^2 + ... + \mu^n q^n$$

$$\Theta^A x_1^A \geq \mu^1 x_1^1 + \mu^2 x_2^1 + ... + \mu^n x_1^n$$

$$\Theta^A x_2^A \geq \mu^1 x_2^1 + \mu^2 x_2^2 + ... + \mu^n x_2^n$$

$$\mu^i \geq 0,$$

where Θ^A is the fraction of the actual inputs which could be used to optimally produce the given level of output, q^A, for observation *A*; x_1 and x_2 are quantities of the two inputs; μ^is are the weights generated for each observation via the linear programming optimization process to obtain the optimal value for Θ; *A* is the observation we are evaluating, and superscripts denote individual firms. Again, $\Theta^A = OB/OA$ for firm *A* in Figure 1 or Figure 7. Therefore, we are finding the lowest fraction of the inputs used which would produce an output level at least as great as that actually produced by firm *A*. Additional linear programs can be solved to derive allocative inefficiency. A more complete description and an example of DEA analysis which has been applied to the banking industry is presented in the Box.

Example of a data envelopment analysis (DEA) program applied to banking

Technical inefficiency is measured as the difference between the observed behavior of bank A to that which would occur if bank A were on the production frontier. Therefore, the unobserved frontier must be projected. This is done via DEA analysis by developing a program which determines the minimum required amount of inputs necessary for bank A to produce as much, or more, of each of the outputs currently being produced. The input vector is chosen based on the observed behavior of the sample firms. Again, this reduces to a linear programming problem. For example, for bank A, the technically efficient combination of inputs is determined as

$$(1)\ \text{Min}\ \Theta^A,$$

$$\text{s.t.}\ \ q_i^A \le \sum_{h=1}^{H} \mu^h \cdot q_i^h, \qquad i = 1, \ldots, m,$$

$$\Theta^A \cdot x_j^A \ge \sum_{h=1}^{H} \mu^h \cdot x_j^h, \qquad j = 1, \ldots, n,$$

$$z_s^A \le \sum_{h=1}^{H} \mu^h \cdot z_s^h, \qquad s = 1, \ldots, s_r,$$

$$z_s^A \ge \sum_{h=1}^{H} \mu^h \cdot z_s^h, \qquad s = s_{r+1}, \ldots, S,$$

$$\mu^h \ge 0,\ h = 1, \ldots, H, \sum_{h=1}^{H} \mu^h = 1,$$

where Θ^A is our radial measure of technical efficiency for firm A, q_i is the output vector, μ^h is a vector of weights assigned to each observation (an intensity vector) which determines the combination of technologies of each firm to form the production frontier, x_j^h is the observed amount of input j used by firm h, and z is a vector of additional exogenous variables.[1] There are two types of these exogenous variables; those that need to be maximized, z_s^h for $s = 1, \ldots, s_r$, and those that should be minimized, z_s^h for $s = s_r + 1, \ldots, S$. An example of these exogenous variables in banking would be the number of branch offices. Banks would, *ceteris paribus*, want to minimize the number of branch offices required to provide a given level of output. The output of each firm in the sample is weighted in such a way that the combination of observed outputs, i, is not less than the output actually produced by firm A. Thus the frontier for firm A is constructed as a weighted technology from the sample. If $\Theta^A = 1$, then firm A is as efficient as any firm in the sample, that is, firm A is on the frontier. If $\Theta^A < 1$ then firm A is inefficient.

Allocative inefficiency for firm A can be derived by determining overall inefficiency and technical inefficiency, and then taking the difference between the two. To determine overall inefficiency, take the observed input prices w_j^A faced by the bank A and assume cost minimizing behavior:

$$(2)\ \text{Min}_{x_j^A} \sum_{j=1}^{n} w_j^A \cdot x_j^A,$$

$$\text{s.t.}\ \ q_i^A \le \sum_{h=1}^{H} \mu^h \cdot q_i^h, \qquad i = 1, \ldots, m,$$

$$x_j^A \ge \sum_{h=1}^{H} \mu^h \cdot x_j^h, \qquad j = 1, \ldots, n,$$

$$z_s^A \le \sum_{h=1}^{H} \mu^h \cdot z_s^h, \qquad s = 1, \ldots, s_r,$$

$$z_s^A \ge \sum_{h=1}^{H} \mu^h \cdot z_s^h, \qquad s = s_{r+1}, \ldots, S,$$

$$\mu^h \ge 0,\ h = 1, \ldots, H, \sum_{h=1}^{H} \mu^h = 1.$$

The optimization process determines the minimum input vector, x^{A*} for the observed price vector w^A. Scalar $w^A \cdot x^{A*}$ is the minimum production cost for the vector of outputs q^A. Overall inefficiency for any firm, h, is therefore the ratio of cost of the observed and the best practice bank:[2]

$$(3)\ O^h = (w^h \cdot x^h) / (w^h \cdot x^{h*}) - 1.$$

The difference between the costs of technically efficient production and overall efficient production determines the cost resulting from allocative inefficiency. That is, $A^h = [(w^h \cdot \Theta^{h*} \cdot x^h) / (w^h \cdot x^{h*})] - 1$ is an index of allocative inefficiency for firm h, and Θ^{h*} is the optimal value of Θ^h determined in Equation (1).[3]

FOOTNOTES

[1]The sum of the weights μ^h used in the optimization process is restricted to unity to allow for varying returns to scale. See Afriat (1972). The appropriate number of constraints for exogenous variables is difficult to determine and the estimated inefficiency for a given model typically varies inversely with the number chosen.

[2]The inequality in the linear program implies free disposability of both inputs and outputs.

[3]Technical inefficiency, determined in Equation 1, obviously is the difference between overall and allocative inefficiency: $T^h = O^h - A^h$.

Comparison of the parametric and nonparametric approaches

Using either the parametric or DEA approach, the goal is to generate an accurate frontier. However the two methods use significantly different approaches to achieve this objective. Because the parametric approach generates a stochastic cost frontier and the DEA approach generates a production frontier, and because the methodologies are fundamentally different, one should expect differences in the efficiency projections. Which methodology is preferable?

There are advantages and disadvantages with each of the procedures. The parametric approach for generating cost relationships requires (accurate) information on factor prices and other exogenous variables, knowledge of the proper functional form of the frontier and the one-sided error structure (if used), and an adequate sample size to generate reliable statistical inferences. The DEA approach uses none of this information, therefore, less data is required, fewer assumptions have to be made, and a smaller sample can be utilized.[15] However, statistical inferences cannot be made using the nonparametric approach.

Another major difference is that the parametric approach includes a random error term around the frontier, while the DEA approach does not. Consequently, the DEA approach will account for the influence of factors such as regional factor price differences, regulatory differences, luck, bad data, extreme observations, etc., as inefficiency.[16] Therefore, one would expect the nonparametric approach to produce greater measured inefficiency.[17] The importance of this difference should not be understated because single outliers can significantly influence the calculated efficiency measure for each firm using the DEA approach.

Obviously, one would like to be able to take comfort in the fact that either approach generates similar results. This is more likely to occur if the sample analyzed has homogeneous units which utilize similar production processes. However, similar results have not been found in the literature. In fact, *it is common for studies contrasting results produced from the two methodologies to find no correlation between the efficiency estimates.* This has also occurred in studies of efficiency for the banking sector. We next review some of that literature.

The role of production inefficiency in banking: A survey of the literature

In this section we review the literature on productive efficiency for financial institutions. Most of the studies reviewed, particularly those analyzing input efficiency, were conducted recently and involve flexible functional forms and state of the art research techniques. For a more comprehensive review of much of the earlier literature on output efficiency, which typically utilized somewhat restrictive functional forms and single output measures, the reader is referred to Gilbert (1984).

Output efficiency

The production process has been one of the most extensively investigated topics in banking. A major purpose of most of these studies has been to obtain estimates of scale elasticities, that is, to evaluate how bank costs change with changes in the level of output.[18] More recently, efforts have also been made to estimate economies of scope; that is, advantages from the joint production of multiple outputs.

Concerning scale economies, if changes in bank costs are proportional to changes in output then the scale elasticity measure equals 1.0 and all cost advantages resulting from the scale of production are being fully exploited. If the changes are not proportional, that is, varying returns to scale exist, then efficiency gains could be obtained by leaving the production process unchanged, but altering the quantity of output produced. Scale elasticities less than one imply that increases in output would produce less than proportional increases in costs. Efficiency gains, therefore, could be obtained by increasing the scale of production. This is typically a justification given for bank merger activity. Efficiency gains could be obtained by reducing production levels if decreasing returns to scale exist; that is, the scale elasticity is greater than 1.0.

Although much effort has been spent evaluating scale economies, it is one of the most disagreed upon topics in banking. For example, a number of studies find cost advantages from size to be fully exhausted at relatively low levels of output. Even when potential economies exist they appear to be relatively small. Some of these studies are summarized in Table 1 which presents the estimated scale elasticity for the average bank in the sample, the range of the estimates for all banks, and the

TABLE 1

Economies of scale estimates for small banks

Author		Scale elasticity at sample mean[a]	Range of scale elasticity measure	Relevant range for significant scale (dis)economies[b]
Benston, Hanweck and Humphrey (1982)	U	1.09	0.89 - 1.24	Diseconomies above $25 million
	B	1.10	0.97 - 1.16	Diseconomies above $25 million
Berger, Hanweck and Humphrey (1987)	U	1.04	0.87 - 1.21	Diseconomies above $100 million
	B	1.03	1.00 - 1.03	No significant (dis)economies
Cebenoyan (1988)	U	1.08*	0.88 - 1.39	Diseconomies above $50 million
	B	0.97	0.92 - 1.03	Economies above $100 million
Gilligan and Smirlock (1984)[c]	U	0.99	0.98 - 1.10	Economies above $10 million and diseconomies above $50 million
Gilligan, Smirlock and Marshall (1984)	U	1.03*	0.93 - 1.27	Economies below $25 million and diseconomies above $100 million
	B	1.02*	0.94 - 1.17	Economies below $25 million and diseconomies above $100 million
Kolari and Zardkoohi (1987)	B	(n.a.)	0.99 - 1.02	No significant (dis)economies
	U	(n.a.)	0.88 - 0 .93	Economies below $100 million
Lawrence and Shay (1986)		0.99	0.91 - 0.99	Economies below $100 million

[a]Calculated as $(d\ lnC/d\ lnQ)$ for single output measures or $\Sigma\ (d\ lnC/d\ lnY_i)$ for all i=outputs. Benston, Hanweck and Humphrey (1982) calculated a scale elasticity augmented for output expansion via office expansion.

[b]In these studies the banks are grouped by deposit size for calculation of the scale elasticity measure. The figures presented are for the minimum bound on the group where statistically significant (dis)economies were realized.

[c]Gilligan and Smirlock did not use the FCA data, as did the other studies, but did evaluate institutions similar in size to those in the FCA sample.

*Denotes statistically significant difference from 1.0.

Note: U and B represent unit and branch bank subsamples, respectively. Many of the studies provided results for a number of years and/or are based on alternative output measures. When multiple sets of findings were provided, the results reported here are for the most recent year, based on earning assets as the output measure, and use the intermediation approach (i.e., dollar value of funds transformed to assets).

level of output at which significant advantages or disadvantages from the scale of production occurs. Basically, the results imply that scale advantages are fully exhausted once an institution achieves a size of approximately $100-200 million, a relatively small bank in the United States.[19] Higher output levels result in either constant or decreasing returns to scale.

The implications from these results are that *very* small banks are inefficient because they operate under increasing returns to scale, and inefficiencies may exist for banks above approximately $100-200 million in deposits. The extent of the inefficiency, however, would not appear to be very large: scale elasticities typically range from .95 to 1.05. These findings would appear to run counter to the arguments typically found in the popular banking press which imply that merger activity, desires to expand geographically, and product expansion are all driven by the desire to reap cost advantages; for example, see Moynihan (1991).

However, this may partially result from the fact that, until very recently, most of the bank cost studies excluded large institutions; the very ones which are most interested in expanding. Most of the studies presented in Table 1 utilized the Federal Reserve's Functional Cost Analysis (FCA) survey data which typically includes only institutions with less than one billion dollars in assets. Although banks in this size group constitute over 95 percent of all

banks in the United States, they constitute only about 30 percent of the nation's banking assets. It excludes the larger banks which are most active in merger activity (Rhoades 1985) and most vocal about expanded product and geographic expansion powers.

Table 2 provides a summary of results from recent studies which have analyzed larger financial institutions; typically in excess of one billion dollars. The evidence suggests that scale advantages exist well beyond the $100-200 million range. While typically significant in a statistical sense, the scale elasticity measure is close to 1.0. Again, the measures tend to range from .95 to 1.05. Therefore, the studies employing data for larger banks tend to argue against the finding that inefficiencies resulting from diseconomies of scale set in at relatively low levels of output. However, the most typical conclusion the authors draw from these bank cost studies is that potential gains from altering scale via internal growth or merger activity are relatively minor.[20]

It should be emphasized, however, that the scale elasticity measure is *not* a measure of inefficiency. This may partially explain some of the disagreement between past research studies claiming potential savings from growth are not very great because scale elasticity measures are not very different from 1.0, and the popular banking press which typically claims that significant cost savings could be gained by expanding the bank scale of operation. Relatively minor scale elasticity deviations from 1.0 can actually result in nontrivial inefficiency.[21] To determine potential gains from scale advantages, the relative comparison is the production costs of existing banks to that of the most efficient sized bank. For example, assuming scale advantages are exhausted at a $5 billion bank, how does the production cost for ten existing $500 million banks compare to that resulting from the one large bank? The scale elasticity measure is required to estimate the cost difference, however it by itself is not a measure of inefficiency.[22]

TABLE 2

Results from large bank cost studies

Author	Range of calculated scale elasticities[a]	Size at which economies of scale are exhausted[b]
Berger and Humphrey (1990)	0.98 - 1.03[e] 0.92 - 1.06[f]	0.3 billion[e] 0.08 billion[f]
Clark (1984)	0.95 - 0.96	Non-exhausted through $500million[h]
Evanoff and Israilevich (1990)[i]	0.98	5.5 billion
Hunter and Timme (1986)[i]	1.05 0.97	$4.2 billion[c] $12.5 billion[d]
Hunter, Timme and Yang (1990)	0.86 - 1.14	$25.0 billion
Noulas, et al. (1990)	0.97 - 1.09	$6.0 billion
Shaffer (1988)	0.94[g]	Non-exhausted through $140 billion[h]
Shaffer (1984)	0.95	Non-exhausted through $50 billion[h]
Shaffer and David (1991)[i]	0.92	$37.0 billion

[a]The reported values are based on elasticity calculations for alternative asset size groups (when available). Statistical significance is not taken into account for figures reported in this column—that is, the calculated values may not be significantly different from 1.0 in a statistical sense.

[b]The values should be considered approximations. The authors frequently reported scale elasticity measures for a group of banks covering a relatively broad size range, for example, 10-25 billion. If the calculated value was insignificantly different from 1.0, then banks up to $25 billion were said to have constant returns to scale. Unlike the figures reported in the previous column, whether or not a calculated scale elasticity is significantly different from 1.0 in a statistical sense is all important for figures in this column.

[c]For one bank holding companies. The value is probably biased downward. This is the sample mean value at which the calculated scale elasticity was insignificantly different from 1.0.

[d]For multibank holding companies. See note c.

[e]Branch bank results for the low cost banks.

[f]Unit bank results for low cost banks.

[g]For a $10 billion bank.

[h]Non-exhausted for the entire sample.

[i]The values are calculated at the sample mean.

Note: Many of the studies provided results for a number of years and/or are based on alternative output measures. When multiple sets of findings were provided, the results reported here are for the most recent year, are based on earning assets as the output measure, and use the intermediation approach.

TABLE 3

Estimated scale inefficiencies in banking

Author	Calculated scale inefficiency (percent)
Aly, et al. (1990)	3.3[c]
Berger and Humphrey (1990)	4.2[b] 12.7[u]
Clark (1984)	18.3[a]
Elyasiani and Mehdian (1990a)	38.9[c]
Evanoff and Israilevich (1990)	16.6
Gilligan, Smirlock and Marshall (1984)	5.0[u] 4.3[b]
Hunter, Timme, and Yang (1990)	26.6
Lawrence and Shay (1986)	5.5
Noulas, et al. (1990)	2.7
Shaffer (1984)	12.0[d]
Shaffer (1988)	10.0[d]

[a]The scale elasticity for the "efficient" firm was .9637 since scale advantages were not exhausted in the data sample. The calculated inefficiency would be larger if we extrapolated outside the study data sample.

[b]Denotes branch banks.

[c]Taken directly from the cited study.

[d]The inefficiency measure is biased downward because data limitations necessitated using an "inefficient" size bank which was not the most inefficient in the sample.

[u]Denotes unit banks.

Note: The reported inefficiencies were derived assuming prices, exogenous variables, and product mix were constant across banks (for example, at the sample mean), and that the cost representation could be approximated by $lnC = a + b\,(lnQ) + .5c(lnQ)^2$ (where Q represents output). Evaluating *only* inefficiency resulting from production in the range of increasing returns to scale, the data were centered about the values of the inefficient bank. Hence, *for this bank*, the scale elasticity measure is simply the coefficient b. The cost of production for the scale efficient bank is $lnC = a + b\,ln(F \cdot Q) + .5c[ln(F \cdot Q)]^2$ where F is the size of the efficient firm relative to the inefficient one. The scale elasticity for the efficient bank is $d\,lnC/d\,ln(F \cdot Q) = b + c\,ln(F \cdot Q) = 1.0$. Scale inefficiency is the difference between cost values of the two banks relative to F, that is, $[\,F + C_E/C_I\,] - 1$, where C_I and C_E denote costs of the inefficient and efficient bank, respectively. The same methodology could be used to calculate inefficiency resulting from production in the range of decreasing returns to scale. In the studies considered, scale measures are typically reported for various size ranges. Unless noted, the calculated inefficiency is based on the smallest bank in the size group in which statistically significant economies of scale existed, relative to the largest bank in the size category in which minimum efficient scale existed (that is, the scale measure was not significantly different from 1.0 in a statistical sense). Details are available from the authors. By holding product mix constant we restrict the cost savings to scale effects only; precluding any savings resulting from altering the mix. This implicitly assumes either that the mix is actually invariant over the banks considered or that the scale efficient bank analyzed is equal in size to the scale efficient bank observed in the data. Given the assumptions employed and the relatively broad size categories reported in the studies considered, the reported inefficiencies should be considered rough approximations.

Using the actual scale estimates and sample data from a number of bank cost studies, measures of scale inefficiency were calculated and are presented in Table 3. The reported inefficiencies are for banks producing in the range of increasing returns to scale. They suggest that potential gains resulting from scale inefficiency are not trivial. While some of the studies suggest inefficiencies in the range of five percent, estimates in the 10-20 percent range are not uncommon, and they range up to nearly 40 percent. The major point is that although their importance is typically played down in the bank cost literature, scale inefficiencies appear to be significant enough to warrant efforts by banks to achieve an efficient scale.[23]

The evidence concerning efficiency gains from economies of scope is not conclusive. Studies to date typically focus on the outputs currently produced and find very slight or no potential for efficiency gains; for example, see Benston, et al. (1982), Cebenoyan (1990), Clark (1988), Hunter, Timme, and Yang (1990), Lawrence and Shay (1986), and Mester (1987).[24] However, the methodologies used to evaluate advantages from joint production have typically been criticized on the grounds that most functional forms utilized for bank cost analysis are ill suited for analyzing economies of scope. Additionally, the evaluation of potential efficiency gains is commonly precluded as a result of regulation. Since numerous products cannot be provided by banks, there is no available quantitative means to evaluate the joint cost relationship or potential efficiency gains.

Input efficiency

While much research has been conducted evaluating output

efficiency, only recently has input efficiency been considered. The evidence suggests that the assumption of input efficiency, common in most studies of bank production, is typically violated. Table 4 presents summary findings for recent studies evaluating input efficiency in banking.

While substantially different techniques were used in the studies reviewed, the results are surprisingly similar. Total input inefficiency is commonly in the range of 20-30 percent, and is as high as 50 percent in one of the studies. This implies that significant cost savings could be realized if bank management more efficiently utilized productive inputs.

Breaking down the study findings into more detail, allocative inefficiency is typically found to be relatively minor and, with one exception, dominated by technical inefficiency.[25] Evanoff and Israilevich (1990a, 1990b, 1991) found that the allocative inefficiency that does exist results from the overuse of physical capital relative to other inputs. As mentioned earlier, this is consistent with expectations since past bank regulation did not allow price competition in the market for deposits. As a result, it appears that banks simply responded by competing using alternative means such as service levels. The introduction of numerous branch offices resulted in brick-and-mortar competition instead of price competition. While the typically small allocative inefficiency estimate cannot be ignored as a potential source of future cost savings in banking, it does suggest that the frequent criticism of bank regulation based on the burden it imposes on the bank production process may be somewhat exaggerated. Apparently the optimal mix of factor inputs is only marginally affected by regulation.

The results presented in Table 4 suggest that the major source of input inefficiency is

TABLE 4

Input inefficiency in banking

Author	Approach	Overall input inefficiency	Allocative inefficiency	Pure technical inefficiency
		(----------percent----------)		
Berger and Humphrey (1990)[b]	Parametric	24.8	Minimal	
Berger and Humphrey (1990)[u]	Parametric	20.2	Minimal	
Elyasiani and Mehdian (1990b)[d]	Parametric			13.6
Evanoff and Israilevich (1990)[a]	Parametric	22.0	1.0	21.0
Evanoff, Israilevich, and Merris (1990)	Parametric		1.8	
Ferrier and Lovell (1990)	Parametric	26.0	17.1	8.9
Aly, et al. (1990)[e]	Nonparametric	50.7	14.9	35.8
Elyasiani and Mehdian (1990a)	Nonparametric			11.7
Ferrier and Lovell (1990)	Nonparametric	21.0	5.0	16.0
Gold and Sherman (1985)[c]	Nonparametric			27.9

[a]For the 1972-87 period. Subperiods produced different results.

[b]For branch banks.

[c]For the most inefficient decision making unit.

[d]For 1980. Scale inefficiency was also calculated to be 38.9%.

[e]Scale inefficiency was also calculated to be 3.1%.

[u]For unit banks.

Note: The figures presented are the level of inefficiency relative to the firm using its inputs efficiently. The studies frequently reported inefficiency relative to the observed firm or efficiency as a percentage of input utilization (see Figure 1 for an illustration of input inefficiency measures). These measures were converted to the measure presented here. Gaps in the results are due to the fact that not all studies considered all components of inefficiency.

pure technical inefficiency. Simply put, firms use too much input per unit of output. Combined with the finding of relatively small allocative inefficiency, this implies that bank managers do a relatively good job of choosing the proper input mix, but then simply underutilize all factor inputs.[26] This inefficiency obviously cannot be sustained over time if the banks are subject to competitive forces. Comparing the findings summarized in Tables 3 and 4, it is apparent that the inefficiencies resulting from the suboptimal use of inputs are somewhat larger than those resulting from producing suboptimal levels of output.[27]

Causes of inefficiency and implications for the future

There are a number of possible explanations for the inefficiency in banking. Basically, the expected causes should be the same as those found in any industry. As discussed earlier, economic theory suggests that allocative inefficiency is driven by market distortions from factors such as regulation. Pure technical inefficiency may be the result of weak market forces (induced by market structure or regulation) which allow bank management to become remiss and to continue their inefficient behavior. Scale and scope induced inefficiency may be the result of either market or regulatory forces which make the optimal level and mix of outputs unachievable. Some analysts would also argue that bank size should be a determinant of efficiency. According to this argument, larger banks may have more astute management and/or be more cost conscious because of greater pressure from owners concerning bottom-line profits. Additionally, these banks are typically located in the larger, more competitive markets which may induce a more efficient production process.

The evidence suggests that these forces are indeed operative in determining efficiency levels in banking. Analyzing data for large banks over the 1972-87 period, Evanoff and Israilevich (1990a, 1990b, 1991) found allocative inefficiency to be related to alternative measures of regulatory stringency. It was also found to be greater in regions characterized by more restrictive state level regulation, and significantly less after industry deregulation occurred in the early 1980s. Allocative efficiency has not been found to be related to bank size (for example, Aly, et al. 1990). This, however, should not be surprising since inefficiency may occur although the bank is operating efficiently in response to shadow market prices (that is, those including market distortions).

The evidence also suggests that pure technical inefficiency is induced by regulation, and some evidence exists suggesting that it results from elements of market structure. Berger and Humphrey (1990) found that the inefficiency was greater, on average, for banks located in the more restrictive unit banking states than those in states allowing branching. Additional analysis of data used in Evanoff and Israilevich (1990b) produced similar findings.[28] Pure technical inefficiency has also been found to be negatively associated with bank size [Berger and Humphrey (1990), Aly, et al. (1990), Elyasiani and Mehdian (1990a), Rangan, et al. (1988)]. To the extent that small institutions are located in the smaller, least competitive markets, the absence of market pressures could be producing the higher levels of inefficiency. Aly, et al. (1990) tested this contention more directly by relating pure technical inefficiency to bank location. Banks located in large metropolitan areas were found to be significantly more efficient than those in smaller markets, suggesting market structure may influence efficiency levels. The evidence here, however, is also not conclusive. It *may* be that cost savings realized by urban banks may exist because the increased population density makes possible less costly delivery systems. This cost savings may be interpreted as being driven by greater market competition in the urban markets while actually it is simply a function of demographics.

Scale and scope diseconomies are also expected to be partially determined by regulatory forces. Unit banking restrictions force banks to expand at one physical location instead of allowing them to expand by opening additional offices to serve customers. Diseconomies of scale may set in at the individual office causing higher cost for larger single office institutions. Expansion via new offices has been shown to be more cost effective. Review of the findings presented in Tables 1 and 2 indicates that diseconomies of scale are typically larger in unit banking markets. Analysis which combines data for both unit and branch banks usually find that the larger unit banks typically operate under conditions of

diseconomies of scale; see for example, Evanoff, Israilevich, and Merris (1990). LeCompte and Smith (1990) also found that inefficiency resulting from not producing the proper mix of outputs, and therefore failing to take advantage of economies from joint production, was greater under conditions of more stringent regulation.

What are the implications of these findings? Given the important role regulation apparently serves in determining efficiency levels, the recent trend toward industry deregulation should result in improved efficiency. Reductions in entry barriers resulting in less regulatory-created market protection, and fewer regulatory-induced market distortions should significantly increase competitive pressures. The beneficial aspects of increased competition will be accomplished by weeding out the less efficient firms. Obviously, in an environment of deregulation and increased competition, reducing pure technical inefficiency could be a major determinant of firm survival.

Merger activity in the financial services industry will probably increase in the future as banks strive to compete in the deregulated market. The deregulation will provide banks with both the desire and the ability to expand acquisition activity. Scale inefficient firms will be absorbed to exploit cost advantages. Firms whose management does an inadequate job of utilizing factor inputs may soon find it difficult to survive in the more competitive market. They will be required to eliminate the inefficiency or become prime targets for acquiring firms looking to "trim the fat" from new acquisitions.[29] Given that pure technical inefficiency is so significant in banking, and given that it is the one aspect of efficiency over which the firm has direct control, one would expect significant increases in bank productive efficiency in the coming years.

Summary and conclusions

The purpose of this study has been to discuss productive efficiency at a conceptual level and to review the relevant literature for the banking industry. We categorize efficiency into input and output related measures. Output inefficiency results from producing suboptimal output levels or a suboptimal combination of outputs. Input inefficiency results from production using a sub-optimal input mix (allocative inefficiency) and not effectively utilizing the inputs employed (pure technical inefficiency). A review of existing bank cost studies suggests that banks have substantial room to increase productive efficiency and, as a result, to significantly lower costs. Although the range of findings in the studies surveyed is relatively broad, it is not uncommon to find 10-20 percent bank scale inefficiency generated by producing at suboptimal output levels. Allocative inefficiency is typically found to be relatively minor; usually less than five percent. Pure technical inefficiency, however, is apparently quite significant; in the range of 20-30 percent. Combining these three effects results in substantial potential cost savings for banks.

What are the major causes of bank inefficiency? The evidence suggests that industry regulation is a dominant source. Allocative inefficiency, although relatively minor, is directly induced by regulation. Inefficiencies resulting from not producing the optimal combination of outputs have also been shown to be related to regulation. However, the major source of inefficiency, pure technical inefficiency, is managerially induced. That is, the absence of competitive forces, which is also influenced by industry regulation, has allowed banks to continue to operate in spite of the fact that management has not effectively utilized the resources available to them.

Given that the industry is undergoing a process of significant deregulation, the findings from these studies have both positive and negative implications for banking. As deregulation continues, the increased competitive pressures will force banks to operate more efficiently. Those unable to do so by adjusting to the new competitive environment will have difficulty surviving. However, one of the major sources of inefficiency, pure technical inefficiency, is directly under the control of the banks themselves. Therefore, they will have control of their own destiny. In light of the recent significant number of branch closings and cost saving campaigns aimed at reducing payrolls, it would appear that efforts to improve bank efficiency are already underway.

FOOTNOTES

[1]Drucker (1991).

[2]However, the research process continues because of differences in results from previous studies, methodologies, assumptions, output definitions, etc. For a review of some of these studies see Gilbert (1984), Clark (1988), or Humphrey (1990).

[3]These definitions of productive inefficiency were introduced by Farrell (1957). They are radial measures and coincide with much of the discussion that follows. For alternative measures of (in)efficiency see Färe and Lovell [1978].

[4]At this stage we are ignoring the potential for economies of scale. Farrell assumed a linearly homogeneous production process; that is, constant returns to scale. We assume, as discussed below, that any gains from scale advantages would result in a higher level isoquant for a given efficient combination of inputs.

[5]For example, if production was allocatively efficient the measure would, obviously, be $OE/OE = 1.0$; that is, points B, C, and E would coincide.

[6]Typically we assume profit maximization as the objective under competitive conditions, that is, frictionless markets and the absence of monopoly power and regulatory distortions. In this model the production, cost, and profit functions are essentially alternative means of evaluating the same optimization process, that is, the production process. The cost relationship is frequently evaluated instead of the production function because less information is required. When discussing frontier analysis, it is irrelevant whether the cost or production side is considered. However, empirically, the choice of a cost, production, or profit representation may generate different results because the researcher is required to use approximations for the true functional forms of these representations.

[7]As a byproduct, the methodology also allows for the estimation of scale and scope induced inefficiency. It does not, however, allow for estimates of pure technical inefficiency.

[8]However, this is an empirical approach, therefore the true cause of the distortion in factor prices could be generated by a number of things, including data measurement errors, etc.

[9]Using this methodology, the shadow price approximations can be interpreted as first-order Taylor's series expansions of arbitrary shadow-price functions. It should be emphasized that there is nothing special about the linear relationship. Alternative specifications can and should be considered; for example, see Evanoff and Israilevich (1991, 1990a).

[10]Typically, factor share equations are derived from the cost relationship via Shephard's Lemma and the system of cost and share equations are jointly estimated. The additional share equations provide increased efficiency of the estimates. The share equations are derived from the *shadow* cost relationship.

[11]For a lucid description of these models, see Bauer (1990). The foundation for this approach was developed by Aigner, Lovell and Schmidt (1977). See also Färe, Grosskopf, and Lovell (1985).

[12]One sided distributions which have been used in estimation include the half-normal, exponential, truncated normal and the Gamma distribution. Again, see Bauer (1990) and the sources cited.

[13]This is the linkage employed in Ferrier and Lovell (1990) in their study of bank efficiency.

[14]The approach can also be combined with others to incorporate additional information. For example, Evanoff and Israilevich (1990b) augmented the thick frontier approach by estimating shadow cost functions for high and low cost banks. In this way, estimates of allocative inefficiency could be obtained directly from the model instead of using an auxiliary, somewhat arbitrary, procedure to decompose the total inefficiency into its component parts.

[15]Technical inefficiency can be calculated ignoring this information. Measures of allocative efficiency require information on factor prices.

[16]Evanoff and Israilevich (1991) found significant regional differences in bank production techniques *and* levels of efficiency.

[17]Surprisingly, Ferrier and Lovell (1990) find exactly the opposite in their analysis of banks.

[18]More formally, the scale elasticity measure is the percentage change in cost relative to the percentage change in output, or $\partial lnC/\partial lnQ$. One major issue in bank cost studies is determining what constitutes output. Although defining output is difficult in any service oriented industry, there seems to be more controversy with respect to banking. However, of the measures used to date, the findings tend to be similar regardless of the measure employed; see Humphrey (1990).

[19]In 1989, over three-quarters of U.S. banks had less than $100 million in assets; see FDIC (1989). However, banks over 1000 times this size also existed.

[20]It is possible that scale estimates could be biased as a result of misspecifying the cost relationship. For example, the standard assumption of efficient utilization of factor inputs, if incorrect, could produce misleading findings concerning scale (dis)advantages. However, Berger and Humphrey (1990), Evanoff and Israilevich (1990b) and Evanoff, Israilevich, and Merris (1990) found scale estimates were *not* substantially different when input inefficiency was accounted for.

[21]The distinction between the scale elasticity and inefficiency measures has been emphasized in Shaffer (1988) and Shaffer and David (1991). Using data from a previous study, Shaffer and David show that a scale elasticity of .99 could result in a 25% cost savings if production was shifted from small to large banks.

[22]Actually the scale inefficiency will be determined by the difference in average cost between the efficient and inefficient firm. The elasticity at the output level corresponding to the inefficient firm gives us information about cost changes at slightly larger or slightly smaller output levels. Neither of these levels is relevant for determining inefficiency since we never produce at these levels. For efficiency analysis, production takes place either at the efficient or the inefficient firm; therefore only the corresponding two average cost values are relevant. Whereas the elasticity measure gives percentage changes in cost induced by incremental changes in output, in the banking studies analyzed in Table 3 the difference in output between the efficient and inefficient firm is not incremental.

[23]Caution should be taken in deriving policy implications from findings concerning scale efficiency alone. There may be alternative factors which partially offset these potential gains. In fact, in viewing bank data, Humphrey (1990) finds the average cost across all bank size groups to be amazingly similar. That, combined with the potential efficiency gains from scale economies discussed here, suggest that there may be some factors counteracting these potential efficiencies. However, with respect to scale efficiency alone, there would appear to be significant potential gains for banking.

[24]Some studies have, however, found significant advantages resulting from joint production; for example, Gilligan, Smirlock, and Marshall (1984), and Evanoff and Israilevich (1990b). However, the finding of relatively small or no scope economies is most typical. The methodologies utilized to generate estimates of scope economies have been critiqued in Pulley and Humphrey (1990). This is obviously a rich area for future research.

[25]Although Aly et al.(1990) find evidence of greater allocative inefficiency than most of the studies reviewed, the major exception to the norm is the study by Ferrier and Lovell (1990). Using a parametric approach the authors found significant allocative inefficiency (over 17 percent). However, as mentioned earlier, the reliability of these techniques decreases significantly when non-homogeneous decision making units are considered. The data for this study included mutual savings banks, credit unions, savings and loan associations, and "noncommercial" institutions. Nearly a third of the sample was made up of noncommercial banks. Given that the technology for these institutions may differ from that of commercial banks, one would expect these observations to have a substantial influence on the error structure of the estimates. The authors themselves even state that some of these observations *do* significantly influence their results (Ferrier and Lovell, p. 243). Since the distribution of the errors is the major determinant of the efficiency measure, this may bias the results concerning commercial bank efficiency. The study also found that the allocative efficiency resulted from an over-utilization of labor relative to the other factors. This is precisely the opposite of what one, intuitively, would expect in banking (see Evanoff and Israilevich 1990b). Finally, the measures used for factor prices may bias the results toward finding allocative inefficiency resulting from over-utilization of labor (Berger and Humphrey 1990, p. 21).

[26]This finding has implications for the bank expense preference literature, for example, Mester (1989). Typically it is assumed that managers of the expensing bank prefer one input to others—usually labor. The results presented here suggest that a more restricted form of expense preference, a preference for all the inputs to the same degree, may best describe the situation in banking.

[27]However, this excludes any inefficiencies resulting from scope disadvantages which cannot be empirically captured.

[28]The evidence on this, however, is not conclusive. Aly, et al. (1990) found no significant efficiency difference across unit and branch banks.

[29]This does not imply that there will no longer be small banks. While most of the bank cost literature has assumed homogeneous outputs, recent research suggests that banks frequently find a market niche in an attempt to differentiate themselves from others. Efficient banks which are able to fill a needed market niche should continue to prosper in a deregulated environment. See Amel and Rhoades (1988).

REFERENCES

Afriat, Sydney, "Efficiency estimation of production functions," *International Economic Review*, 13, 1972, pp. 568-598.

Aigner, Dennis, C.A. Knox Lovell, and Peter Schmidt, "Formulation and estimation of stochastic frontier production function models," *Journal of Econometrics*, 6, 1977, pp. 21-37.

Aly, Hassan Y., Richard Grabowski, Carl Pasurka, and Nanda Rangan, "Technical, scale, and allocative efficiencies in U.S. banking: an empirical investigation," *The Review of Economics and Statistics*, 72, 1990, pp. 211-218.

Amel, Dean F., and Stephen A. Rhoades, "Strategic groups in banking," *The Review of Economics and Statistics*, 70, 1988, pp. 685-689.

Atkinson, Scott E., and Robert Halvorsen, "Parametric efficiency tests, economies of scale, and input demand in U.S. electric power generation," *International Economic Review*, 25, 1984, pp. 647-662.

Atkinson, Scott E., and Robert Halvorsen, "A test of relative and absolute price efficiency in regulated utilities," *The Review of Economics and Statistics*, 62, 1980, pp. 185-196.

Bauer, Paul W., "Recent developments in the econometric estimation of frontiers," *Journal of Econometrics*, 46, 1990, pp. 39-56.

Benston, George, Gerald A. Hanweck, and David B. Humphrey, "Scale economies in banking," *Journal of Money, Credit, and Banking*, 14, 1982, pp. 435-456.

Berger, Allen N., Gerald A. Hanweck, and David B. Humphrey, "Competitive viability in banking: Scale, scope, and product mix economies," *Journal of Monetary Economics*, 16, 1987, pp. 501-520.

Berger, Allen N., and David B. Humphrey, "The dominance of inefficiencies over scale and product mix economies in banking," forthcoming in *Journal of Monetary Economics*, 28, 1991. Also in *Finance and Economics Discussion Series*, 107, Board of Governors of the Federal Reserve System, 1990.

Cebenoyan, A. Sinan, "Multiproduct cost functions and scale economies in banking," *The Financial Review*, 23, 1988, pp. 499-512.

Cebenoyan, A. Sinan, "Scope economies in banking: the hybrid box-cox function," *The Financial Review*, 25, 1990, pp. 115-125.

Clark, Jeffrey A., "Estimation of economies of scale in banking using a generalized functional form," *Journal of Money, Credit, and Banking*, 16, 1984, pp. 53-67.

Clark, Jeffrey A., "Economies of scale and scope at depository financial institutions: A review of the literature," *Economic Review*, Federal Reserve Bank of Kansas City, 1988, p. 16-33.

Drucker, Peter F., "Don't change corporate culture—use it," *The Wall Street Journal*, March 28, 1991, p. A 14.

Elyasiani, Elyas, and Seyed M. Mehdian, "A nonparametric approach to measurement of efficiency and technological change: The case of large U.S. commercial banks," *Journal of Financial Services Research*, 4, 1990a, pp. 157-168.

Elyasiani, Elyas, and Seyed M. Mehdian, "Efficiency in the commercial banking industry, a production frontier approach," *Applied Economics*, 22, 1990b, pp. 539-551.

Evanoff, Douglas D., "Branch banking and service assessibility," *Journal of Money, Credit, and Banking*, 1988, pp. 191-202.

Evanoff, Douglas D., and Philip R. Israilevich, "Cost economies and allocative efficiency in large U.S. commercial banks," *Proceedings of a Conference on Bank Structure and Competition*, 26, 1990a, pp. 152-169.

Evanoff, Douglas D., and Philip R. Israilevich, "Deregulation, cost economies and allocative efficiency of large commercial banks," *Issues in Financial Regulation*, Federal Reserve Bank of Chicago Working Paper 90-19, 1990b.

Evanoff, Douglas D., and Philip R. Israilevich, "Regional differences in bank efficiency and technology," *The Annals of Regional Science*, 25, 1991, pp. 41-54.

Evanoff, Douglas D., Philip R. Israilevich, and Randall C. Merris, "Relative efficiency, technical change, and economies of scale for large commercial banks," *Journal of Regulatory Economics*, 2, 1990, pp. 281-298.

Färe, R.J., Shawna Grosskopf, C.A. Knox Lovell, *The measurement of efficiency of production*, Boston, Kluwer Academic Publishers, 1985.

Färe, R.J., and C.A. Knox Lovell, "Measuring the technical efficiency of production," *Journal of Economic Theory*, 19, 1978, pp. 150-162.

Farrell, M.J., "The measurement of productive efficiency," *Journal of Royal Statistical Analysis, A*, 120, 1957, pp. 253-281.

FDIC, "Statistics on banking," Federal Deposit Insurance Corporation, Washington, GPO, 1989.

Ferrier, Gary D., and C.A. Knox Lovell, "Measuring cost efficiency in banking: Econometric and linear programming evidence," *Journal of Econometrics*, 46, 1990, pp. 229-245

Gilbert, R. Alton, "Bank market structure and competition," *Journal of Money, Credit, and Banking*, 16, 1984, pp. 617-644.

Gilligan, Thomas W., and Michael L. Smirlock, "An empirical study of joint production and scale

economies in commercial banking," *Journal of Banking and Finance*, 8, 1984, pp. 67-77.

Gilligan, Thomas W., Michael L. Smirlock, and William Marshall, "Scale and scope economies in the multi-product banking firm," *Journal of Monetary Economics*, 13, 1984, pp. 393-405.

Humphrey, David B., "Why do estimates of bank scale economies differ?," *Economic Review*, Federal Reserve Bank of Richmond, 1990, pp. 38-50.

Hunter, William C., and Stephen G. Timme, "Technical change, organization form, and the structure of bank production," *Journal of Money, Credit, and Banking*, 18, 1986, pp. 152-166.

Hunter, William C., Stephen G. Timme, and Won Keun Yang, "An examination of cost subadditivity and multiproduct production in large U.S. banks," *Journal of Money, Credit, and Banking*, 22, 1990, pp. 504-525.

Kolari, James, and Asghar Zardkoohi, "Bank cost, structure, and performance," Lexington, D.C., Heath Publishers.

Kopp, Raymond, and W. Erwin Diewert, "The decomposition of frontier cost function deviations into measures of technical and allocative efficiency," *Journal of Econometrics*, 18, 1982, pp. 319-331.

Lau, L. J., and P. A. Yotopoulos, "A test for relative efficiency and application to Indian agriculture," *American Economic Review*, 61, 1971, pp. 94-109.

Lawrence, Colin, and Robert Shay, "Technology and financial intermediation in a multiproduct banking firm: an econometric study of U.S. banks, 1979-82," in Colin Lawrence and Robert Shay (ed.), *Technological Innovation, Regulation, and the Monetary Economy*, Cambridge, Ballinger, 1986, pp. 53-92.

Lecompte, Richard, L. B., and Stephen D. Smith, "Changes in the cost of intermediation: The case of savings and loans," *Journal of Finance*, 45, 1990, pp. 1337-1345.

Mester, Loretta J., "A multiproduct cost study of savings and loans," *Journal of Finance*, 42, 1987, pp. 423-445.

Mester, Loretta J., "Testing for expense preference behavior: Mutual and stock savings and loans," *Rand Journal of Economics*, 20, 1989, 483-498.

Moynihan, Jonathan P., "Banking in the 90s—where will the profits come from?," *Proceedings of a Conference on Bank Structure and Competition*, Federal Reserve Bank of Chicago, 27, 1991.

Noulas, Athanasios G., Subhash C. Ray, and Stephen M. Miller, "Returns to scale and input substitution for large U.S. banks," *Journal of Money, Credit, and Banking*, 22, 1990, pp. 94-108.

Pulley, Lawrence B., and David B. Humphrey, "Correcting the instability of bank scope economies from the translog model: A composite function approach," paper presented at the Financial Management Association meetings, Orlando Florida, October, 1990.

Rangan, Nanda, Richard Grabowski, Hassan Aly, and Carl Pasurka, "The technical efficiency of U.S. banks," *Economic Letters*, 28, 1988, pp. 169-75.

Rhoades, Stephen A., "Mergers and acquisitions by commercial banks," *Staff Studies*, 142, Board of Governors of the Federal Reserve System, 1985.

Shaffer, Sherrill, "Scale economies in multiproduct firms," *Bulletin of Economic Research*, 1, 1984, pp. 51-58.

Shaffer, Sherrill, "A revenue-restricted cost study of 100 large banks," Federal Reserve Bank of New York, unpublished research paper, 1988.

Shaffer, Sherrill, and Edmond David, "Economies of superscale in commercial banking," *Applied Economics*, 23, 1991, pp. 283-293.

Sherman, H. David, and Franklin Gold, "Bank branch operating efficiency," *Journal of Banking and Finance*, 9, 1985, pp. 297-315.

Zieschang, Kimberly D., "A note on the decomposition of cost efficiency into technical and allocative components," *Journal of Econometrics*, 23, 1983, pp. 401-405.