

# **Optimal Industrial Structure in Banking**

**Loretta J. Mester<sup>1</sup>**

**Federal Reserve Bank of Philadelphia  
and  
Finance Department, The Wharton School, University of Pennsylvania**

**Prepared for the *Handbook of Financial Intermediation*  
(Elsevier, forthcoming in June 2008)**

**July 25, 2005**

## **Abstract**

This chapter discusses the research agenda on optimal bank productive efficiency and industrial structure. One goal of this agenda is to answer some fundamental questions in financial industry restructuring, such as what motivates bank managers to engage in mergers and acquisitions, and to evaluate the costs and benefits of consolidation, which is essentially an empirical question. The chapter reviews the recent literature, including techniques for modeling bank production and the empirical results on scale economies, scope economies, and efficiency in banking.

## **1. Introduction and Motivation**

The banking industry has been undergoing a significant restructuring over the last several years. Since the mid-1980s the number of commercial banks has fallen by over 6,000 (from 14,407 in 1985 to 7630 in 2004) as a result of failures and especially mergers. According to FDIC statistics, there were over 9,400 mergers, an average of 523 per year, between 1985 and 2002.<sup>2</sup> The average asset size of banks has also increased, as assets are being redistributed from smaller banks to larger ones. In real terms, the average asset size of U.S. banks has more than tripled since 1985 and in 2004 was over \$1 billion. Citigroup, Inc., the largest bank holding company in the U.S. (as of May 2005) has nearly \$1.5 trillion in assets. Another result of consolidation is that by some measures, banking is becoming more concentrated. According to data from commercial bank reports of condition and income, the largest 10 banks in the U.S. were holding almost half of the U.S. banking industry's assets in 2004, compared to 25 percent in 1985. Banks with assets over \$5 billion were holding about 79 percent of total U.S. banking industry assets in

---

<sup>1</sup> The views expressed here are those of the author and do not necessarily represent the views of the Federal Reserve Bank of Philadelphia or of the Federal Reserve System. This paper is available free of charge at: [www.philadelphiafed.org/econ/wps/](http://www.philadelphiafed.org/econ/wps/).

2004, compared with 47 percent in 1985 and banks with more than \$10 billion in assets were holding almost 75 percent of industry assets in 2004, compared with about 35 percent in 1985.

Consolidation is a global phenomenon: there has been a considerable amount of merger activity not only in the U.S. but in other countries. The Group of Ten study (2001) of consolidation in 13 countries in the 1990s indicates that of 7,304 financial mergers, 61 percent involved banks, and the number of banks fell in almost every country.<sup>3</sup>

The consolidation in the financial services industry has raised some conundrums.

Conundrum 1: The consolidation in the banking industry has created some very large banks; indeed, as of December 31, 2004, there were three bank holding companies in the U.S. with over \$1 trillion in assets. Bank managers say that one of their motivations in consolidating is to capture scale economies – i.e., efficiencies gained from operating at a large scale – but much of the literature suggests these economies are exhausted at relatively small sizes.

Conundrum 2: The Gramm-Leach-Bliley Act repealed the Glass-Steagall prohibitions against mixing commercial banking with investment banking and allowed commercial banks into other nonbank activities (like insurance). Under Gramm-Leach-Bliley, an institution must form a Bank Holding Company (BHC) and then convert the BHC into a Financial Holding Company (FHC) before engaging in nontraditional activities. As of July 2, 2004, there were 637 FHCs; most were created from long standing BHCs rather than de novo. Many are small; most are not engaged in nonbanking activities. Indeed, fewer commercial banks have moved into those areas than was anticipated when the Act was passed.

Conundrum 3: Official government statistics suggest that productivity in the banking industry rose at a slower rate from 1994 to 2002 than that in the rest of the corporate sector. This seems somewhat

---

<sup>2</sup> FDIC's Historical Statistics on Banking, Table CB02, Changes in Number of Institutions, FDIC-Insured Commercial Banks, January 14, 2005.

<sup>3</sup>Only in Belgium, Japan, and Australia did the number of banks rise in the 1990s – Japan because of a change in definition and Belgium by just two banks. Although the number of banks in Australia increased from 34 in 1990 to 44 in 1999, the report characterizes the banking industry in Australia as highly concentrated. The five largest banks in Australia hold 74 percent of deposits, while the five largest banks in the U.S. in 1999 held 27 percent of deposits. (As of 2004, the largest five banks in the U.S. held 38 percent of industry deposits and 37 percent of industry assets.)

surprising given the technological advances that have been made in banking over the last two decades.<sup>4</sup>

One goal of the research agenda on optimal bank productive efficiency and industrial structure is to answer some fundamental questions in financial industry restructuring, such as what motivates bank managers to engage in mergers and acquisitions, and to evaluate the costs and benefits of consolidation, which is essentially an empirical question. The most recent literature has begun to shed some light on these three puzzles via several advances in modeling bank production which are then brought to data. These advances include recognizing that the level of risk is an endogenous choice of bank managers and that financial capital is an input into bank production. Models that allow for managerial preferences that differ from cost minimization and profit maximization and that allow managers to trade off risk versus return have been developed and estimated. The models can be used to help in understanding the motivation and outcomes of banking industry consolidation.

Consolidation is a potential positive for the industry and economy to the extent that it eliminates inefficient banks and results in a healthier banking system via better diversification of risks. Positives potentially include the following: (1) More efficient scale or product mix. Scale or scope economies exist if the average cost of production declines as size or number of products increases. Average cost might decline as the size of the bank increases if there are significant fixed costs that can be spread over larger sized operations. Technological change has afforded banks new tools of financial engineering (e.g., derivatives, off-balance-sheet guarantees, and risk management) that may be more efficiently produced by larger institutions. Also, new delivery methods for banking services (e.g., on-line banking, phone centers, ATMs) favor larger banks that can spread the fixed costs of setting up such systems over larger volumes, implying lower average costs of production. With respect to product mix, if there are cost complementarities among products (e.g., joint use of inputs, such as marketing), then producing multiple

---

<sup>4</sup>According to Berger and Mester (2003), government agencies typically measure productivity by the ratio of an output index to an input index. Updating the statistics reported in Berger and Mester (2003), average annual growth in labor productivity (measured by output per employee hour) in commercial banking (NAICS Code 52211) was 1.58 percent over 1994-2002, compared to 4.09 percent in manufacturing, 2.34 percent in nonfarm businesses, and 2.79 percent in nonfinancial corporations. These data indicate banking productivity is rising at a slower pace than the productivity of the rest of the corporate sector.

products in the same bank can be efficient. (2) Better diversification over product lines and/or across geographic markets. The price of risk-taking falls via diversification, and there is some evidence that U.S. acquiring banks bid more for targets when the M&A would lead to significant diversification gains (Benston, Hunter, Wall, 1995). (3) Higher X-efficiency. Even firms that are operating at the efficient scale of operations and producing the efficient mix of products might not be doing so in a manner that minimizes costs, e.g., managers may be wasting some of the firm's inputs or diverting some for their own benefit. Consolidation can help rid the industry of such X-inefficiency to the extent that more efficient firms take over less efficient firms and are able to extend efficient operations to the target. In many U.S. M&As, a larger, more efficient institution takes over a smaller, less efficient institution and acquiring banks are more cost efficient than target banks on average (Pilloff and Santomero, 1998).

But consolidation could also be a negative for the industry and economy. It could result in a less competitive banking system, concentrating market power in a handful of very large institutions, or reduce the supply of funds to small firms by driving community banks out of business. To the extent that banks can be "too-big-to-fail," consolidation might be motivated by banks' desire to exploit the under-priced federal safety net. Using 1990 data on U.S. bank holding companies, Hughes and Mester (1993) found evidence of such a "too-big-to-fail" size advantage: for banks with greater than \$6.5 billion in assets, an increase in size, holding default risk and asset quality constant, significantly lowers the uninsured deposit price. Consolidation might be motivated by a desire to maximize managers' objectives and therefore not be socially optimal. But even if consolidation is motivated by a desire to maximize shareholder value, it need not be socially optimal. While shareholder value can be raised via more efficient production, it can also be raised via higher prices if banks' market power rises via consolidation

Systemic risk problems might also increase as a result of consolidation. Adverse shocks to a large bank can be transmitted across the financial system, since a large bank has more linkages to other banks. As discussed in the G10 report (2001) on consolidation, evidence shows that interdependencies between large and complex banking organizations have increased in the last 10 years in the U.S. and Japan and are beginning to do so in Europe. These increases are correlated with consolidation (but a

causal link has not been established). According to the G10 report, the interdependencies most associated with consolidation include interbank loans, market activities such as over-the-counter derivatives, and payment and settlement systems.

Research suggests that consolidation in the latter half of the 1990s was not driven by the cleanup of failed or failing banks, since bank performance was very good (Berger and Mester, 2003); thus, it is a trend rather than merely a response to cyclical events.<sup>5</sup> Instead, changes in the banking environment appear to be important factors spurring consolidation. These include technological progress, improvements in financial condition, which allowed for more voluntary M&As, deregulation of geographic and product restrictions on banking, which allowed the industry to evolve into the structure that would have existed had the restrictions not been imposed, and excess capacity in the industry or particular markets. International consolidation (globalization) of markets also has been a factor. Transfer of securities, goods, and services in international markets creates demand for financial services in international markets, spurring cross-border M&As. Banks can also achieve the dual goals of risk diversification and new sources of funds by cross-border expansion. But these potential benefits must be weighed against the costs, which include having to deal with different regulatory regimes and corporate and national cultural differences.

The research on optimal bank productive efficiency and industrial structure can help in evaluating the extent to which consolidation yields cost and revenue benefits or, instead whether it is a way that agency problems within the firm are manifested, whether consolidation is attractive to managers because they gain from “building empires” and controlling larger banks, and whether large banks allow managers to consume “agency goods” such as reduced effort and risk avoidance. In helping us understand the motivation for consolidation, the research can also help guide policy regarding consolidation in the

---

<sup>5</sup> The mid-1980s to early 1990s was a time of relatively poor performance of U.S. banks. Performance problems with loans to less developed countries and in commercial real estate markets led to performance problems at U.S. banks and a “credit crunch” in the early 1990s. This was the first phase of the consolidation trend and the number of banks fell by almost 20 percent between 1984 and 1991. After the credit crunch period, the banking industry had much better performance. Profits and efficiency rose, the ratio of nonperforming loans to total loans fell, and risk-taking and deposit market concentration remained constant (see Berger and Mester, 2003).

industry.

The rest of this chapter is organized as follows. Section 2 discusses the concepts used in evaluating banking firm and industry productive efficiency. Section 3 discusses empirical implementation of the concepts. Section 4 discusses measurement issues that must be confronted when bringing the concepts to data. Section 5 discusses the main empirical findings in the literature related to each concept. Section 6 concludes.

## 2. Efficiency Concepts

In investigating the optimal structure of the banking industry and its efficiency, one must start with a concept of optimization. As a general definition, efficiency is a measure of deviation between actual performance and desired performance. Thus, efficiency must be measured relative to an *objective function*. A fundamental decision in measuring financial institution efficiency is which concept to use, and the choice will depend on the question being asked.

The concept chosen should be related to *economic* optimization in reaction to market prices and competition, rather than being based solely on the use of technology. We can ask the question, is the bank maximizing the amount of output it produces given its inputs or minimizing the amount of inputs it uses to produce a given level of output – i.e., is it operating on its production frontier – but that is a question about *technological* optimization. This is less interesting from an economic perspective, since it ignores values. It cannot account for allocative inefficiency in mis-responding to relative prices in choosing inputs and outputs, and it is difficult to compare firms that tend to specialize in different inputs or outputs, because there is no way to compare one input or output with another without the benefit of relative prices. There is also no way to determine whether the output being produced is optimal without value information on the outputs. Instead, we would like to investigate questions of *economic* optimization.<sup>6</sup>

For example, is the bank minimizing its costs of production given its choice of inputs, taking

input prices as given? Is the bank maximizing its profits given its choice of inputs and outputs, taking input and output prices as given? A bank might be operating on its production frontier (i.e., not wasting resources), and so be *technically* efficient, but it could still be *allocatively* inefficient if it is choosing the wrong mix of inputs given the relative prices of those inputs. Similarly, the bank could be technically and allocatively efficient in producing its chosen level of output, but it could be choosing the wrong level of output in order to maximize profits.

Figure 1 shows a simple two input, one output case of firm production. The figure shows an isoquant – the combinations of inputs  $x_1$  and  $x_2$  (say labor and capital) it takes to make output level  $y_0$ . Firm B is technically efficient, since it is operating on the isoquant. Firm A is inefficient, since it is operating interior to the isoquant. That is, Firm A is using more of inputs  $x_1$  and  $x_2$  to produce  $y_0$  than an efficient firm would use. But note that Firm B could do better as well. Firm B could lower its costs of producing  $y_0$  by using a different combination of the inputs, given their prices  $w_1$  and  $w_2$ . Namely, given the prices of the inputs, Firm B would minimize its cost of producing  $y_0$  by operating at point O. Firm B should use more  $x_1$  and less of  $x_2$ . Since we want to capture such allocative inefficiency, we want to focus on the economic concepts of cost-minimization and profit-maximization, which are based on economic optimization in reaction to market prices and competition, rather than based solely on the use of technology.

There are different aspects to economic optimization. Most of the literature focuses on cost minimization. But from a performance standpoint, one might also investigate whether the bank is producing the optimal outputs in terms of profitability and firm value. For this, one can study the profit function (and, less commonly, the revenue function). This is important to the extent that bank output quality is a significant choice variable for the bank. If revenue losses more than counteract cost savings, the choice is not profit maximizing. Profit efficiency includes revenue benefits from improving product mix and can reflect the benefits of improved diversification.

Newer studies acknowledge the fact that the objectives of firm management may differ from cost

---

<sup>6</sup> For further discussion see Berger and Mester (1997) and Mester (2003).

minimization and profit maximization and try to incorporate this into efficiency measurement. These papers focus on more market-based definitions of efficiency, e.g., operation on a risk-return frontier.

Three main types of efficiency are measured: scale, scope, X-efficiency. They are used to address questions pertaining to different aspects of bank structure.

What is the optimal scale of operations of the bank? This is pertinent to the issue of optimal structure in terms of number of firms in the industry; is banking a natural monopoly? *Scale economies* are usually measured with respect to bank costs and refer to how the bank's scale of operations (its size) is related to cost – what percentage increase in costs occurs with a 1 percent increase in scale. A firm is operating at constant returns to scale if, for a given mix of products, a proportionate increase in all its outputs would increase its costs by the same proportion; a firm is operating with scale economies if a proportionate increase in scale leads to a less than proportionate increase in cost; a firm is operating with scale diseconomies if a proportionate increase in scale leads to a more than proportionate increase in cost. For single-product firms, operating at the point of constant returns to scale implies operating at minimum average cost.

What is the efficient mix of outputs in banking? That is, what's the optimal combination of products to minimize cost (or maximize profits)? This is pertinent to the issue of universal banking and the mixing of commercial and investment banking in the aftermath of Gramm-Leach-Bliley. *Scope economies* are usually measured with respect to bank costs and refer to how the bank's choice of multiple product lines is related to cost. A firm producing multiple products enjoys scope economies if it is less costly to produce those products together in a single entity than it would be to separate production into specialized firms.<sup>7</sup> A potential source of such scope economies is the opportunity to cross-market new and existing products to customers. For example, the merger of Citibank with Travelers, which had bought Smith Barney (which had bought Salomon), brought together commercial banking, securities, and insurance products. On the other hand, the cost of integrating disparate computer systems in order to take

---

<sup>7</sup> Note, I have defined scale and scope economies relative to the costs of production, but they could just as well have been defined relative to the bank's revenues or profits.

advantage of such potential cross-marketing opportunities might mitigate any scope economies.

Given the technology, what percent of banks are using the best-practice methods of production, i.e., are operating on the efficient frontier? *X-efficiency* measures how productive the firm is in its use of inputs to create output. The concept refers to the dispersion of costs (profits, revenues) in any given size/product mix class. If all firms in an industry are producing the scale and combination of outputs that minimize the average cost of production, then the total cost of producing the industry's output is minimized, and the industry is producing the efficient combination and level of products, provided each firm is using its inputs efficiently. Firms that exhibit cost X-inefficiency are either wasting some of their inputs (technical inefficiency) or are using the wrong combination of inputs to produce outputs (allocative inefficiency), or both. Management ability (or lack thereof) may be a source of X-inefficiency, but managerial preferences might be another source, to the extent that managers can pursue objectives that differ from those of stockholders. For example, managers might derive utility,  $U$ , from having large staffs or other perquisites, as well as high profits, so that  $U=U(\pi,E)$ , where  $\pi$  is profits and  $E$  is expenditure on labor (or other inputs). Some studies of commercial banks and savings and loans have found evidence of such "expense-preference" behavior; others have found evidence of "empire building," i.e., pursuit of inefficient mergers to gain larger scale and presumably prestige (see Edwards, 1977; Mester, 1989a; Mester 1989b; Mester 1991; and Hughes, Lang, Mester, Moon, and Pagano, 2003).

How has the production technology shifted over time (*technological change*) and how has productivity changed over time? *Productivity* is a combination of a shift in the best-practice frontier and in dispersion from the frontier (X-inefficiency).

These concepts can be focused more specifically on the optimality of particular aspects of bank strategy. For example, Berlin and Mester (1998) provide evidence on whether relationship lending is efficient. Banks are able to smooth loan rate for their borrowers with which they have formed a long-term relationship. Berlin and Mester (1998) find that loan-rate smoothing in response to a shock to a small-business borrower's credit risk is not efficient, but in response to an interest-rate shock such loan-rate smoothing is efficient.

### 3. Empirical Implementation

**3.1. Bank Production.** To bring efficiency concepts to bear in investigating the optimal structure of the banking firm, one must begin with a theory of the banking firm – i.e., what do banks do. Most of the literature applies traditional microeconomic theory of firm production to banking firms – a bank is a factory producing financial services (like a factory makes widgets). The newer literature takes seriously the bank as a financial intermediary that differs from other types of firms. Factors important for banks that have generally been ignored in much of the literature include the bank's choice of risk and diversification of assets; asset quality and its feedback on the bank's input prices; and the bank's financial capital structure. The newer literature combines the theory of financial intermediation with the microeconomics of bank production (see Hughes, Lang, Mester, and Moon, 2000; Hughes 1999; Hughes, Lang, Mester, and Moon, 1999; and Hughes, Mester, and Moon, 2001).

In the standard application of efficiency analysis to banking, bank production decisions do not affect bank risk. The bank is assumed to take the entire price of its outputs and inputs as given. This rules out the possibility that scale-related improvements in diversification could lower the cost of borrowed funds and induce banks to alter their exposure to risk. In contrast, the newer research recognizes the bank's role as a monitor and producer of information and the bank as a manager of risk. The theory of the banking firm emphasizes the bank's role in producing information about its borrowers. Hence, output measures should attempt to proxy for these aspects of banking. One study, Mester (1992), directly accounted for the monitoring and screening role of banks in measuring bank output by treating loans purchased and originated loans as separate outputs entailing different types of screening, and treating loans held on balance sheet and loans sold as separate outputs entailing different types of monitoring.

The bank's choice of capital structure (funding choices regarding capital and debt) and its strategic decisions regarding asset quality vary with production decisions. Thus, part of the input and output prices a bank faces are not exogenous – the risk premium in these prices is partly endogenous as it

depends on the bank's production choices. This affects the modeling of banking production and therefore measurement of scale and scope economies (Hughes, Lang, Mester, and Moon, 2000; Hughes, 1999). But in standard efficiency studies, the bank is assumed to choose a production plan to minimize cost and maximize profits *given* the prices of inputs and outputs (including the required return on shareholders' equity). That is, the standard assumption is that the required return on debt and equity is independent of production decisions of the firm. The higher moments of cost and profit are assumed not to vary across banks. In newer research, banks are modeled as taking actions that will maximize their market value – since production decisions affect bank risk, they affect the discount rate applied to evaluating discounted present value. Production decisions that increase expected profit but also the discount rate applied to that profit may not increase the bank's market value. The optimal production choices depend not only on the expected profits they generate but also on the variability of the profit stream generated. The newer research tries to evaluate the tradeoff between expected return and riskiness of that return.

The newer theory also recognizes that bank managers may be making production decisions that are not value-maximizing because of agency problems between owners and managers. The researcher has data on the decisions managers are actually making, which need not be value-maximizing decisions. Measurement of scale and scope economies and X-efficiency should take this into account.

**3.2. Cost Minimization.** The second step in empirical implementation is to decide which optimization goal to investigate, e.g., cost, profits, revenue. The earliest literature assumed that the bank produced a single output. Once techniques were developed for measuring scale and scope economies at multiproduct firms (Baumol, Panzar, and Willig, 1982), these techniques were applied to financial institutions.

In a cost function, variable costs depend on the prices of variable inputs, the quantities of variable outputs, any fixed inputs or outputs, and environmental factors, as well as an error term. If the error term includes only random error and not the possibility of X-inefficiency, then the estimated cost function is an *average-practice cost function*, describing the average relationship between costs, outputs, and input prices. If the error term includes a term representing random error and a term representing X-

inefficiency, then the estimated cost function is a *best-practice frontier*, which indicates the cost of a bank producing using the best practices under ideal conditions. (Note, this does not necessarily represent the best possible practice, merely the best practice observed among banks in the sample. See Berger and Mester, 1997.) Such a cost function is often written in logarithmic form:

$$\ln C_i = \ln f(y_i, w_i, z_i, h_i) + u_i + v_i, \quad (1)$$

where  $C$  measures variable costs,  $w$  is the vector of prices of variable inputs,  $y$  is the vector of quantities of variable outputs,  $z$  indicates the quantities of any fixed netputs (inputs or outputs, such as physical plant, which cannot be changed quickly),  $h$  is a set of environmental or market variables that may affect performance (e.g., regulatory restrictions) but are not a choice for firm management,  $u_i$  denotes an inefficiency factor that may raise costs above the best-practice level, and  $v_i$  denotes the random error that incorporates measurement error and luck that may temporarily give firms high or low costs. The inefficiency factor  $u_i$  incorporates both allocative inefficiencies from failing to react optimally to relative prices of inputs,  $w$ , and technical inefficiencies from employing too much of the inputs to produce  $y$ .

The function  $f$  denotes some functional form and represents the best-practice frontier. The term,  $u_i + v_i$ , is treated as a composite error term:  $v_i$  is a two-sided error, since random measurement error or luck can be positive or negative, and  $u_i$  is a one-sided (positive) error, since inefficiency means higher costs. The various X-efficiency measurement techniques use different methods to identify the inefficiency term,  $u_i$ , as distinct from the random error term,  $v_i$ .

*Scale economies* measure the percentage change in costs per one percent increase in all the outputs, as given by the frontier. Consider composite output bundle  $y^0$ , and suppose  $y = ty^0$ . Then

$$SCALE = \frac{f}{\left(\frac{df}{dt}\right)} = \frac{f}{\sum_{i=1}^N \frac{\partial f}{\partial y_i} y_i} = \frac{1}{\sum_{i=1}^N \frac{\partial \ln f}{\partial \ln y_i}} = \frac{1}{\sum_{i=1}^N \frac{\partial \ln C}{\partial \ln y_i}}, \quad (2)$$

where  $N$  = number of outputs.

There are scale economies (i.e., increasing returns to scale) if  $SCALE > 1$ ; scale diseconomies (i.e., decreasing returns to scale) if  $SCALE < 1$ ; and constant returns to scale if  $SCALE = 1$ . Note that for

single-product firms, choosing  $y$  such that  $SCALE = 1$  minimizes the average cost of production.

*Scope economies* measure whether it is less costly for a multiproduct firm to produce the outputs together than for single-product firms to produce the products, as given by the frontier.

$$SCOPE(y_1, \dots, y_N) = \frac{[f(y_1, 0, \dots, 0) + f(0, y_2, 0, \dots, 0) + \dots + f(0, \dots, 0, y_N)] - f(y_1, y_2, \dots, y_N)}{f(y_1, y_2, \dots, y_N)} \quad (3)$$

Several criticisms have been levied at the scope economies measure. First, it requires evaluation of the cost function at zero output levels. This rules out certain functional forms, such as the translog, in which outputs appear in logarithmic form. Researchers have handled this either by replacing the zero output with a very small positive number or by selecting a functional form that permits zero output levels (e.g., the hybrid translog function which replaces  $\ln y_i$  in the translog cost function with  $y_i$  transformed by the Box-Cox metric, i.e.,  $[y_i^\lambda - 1]/\lambda$ , where  $\lambda$  is a parameter to be estimated).

A more telling criticism of the conventional measure of scope economies is that it requires the cost function to be evaluated at zero output levels even if all firms in the sample are producing positive levels of each output, as they often are in banking studies. So the scope measure involves extrapolation outside the sample. This problem is not resolved by functional forms such as the hybrid translog, which permit evaluation at zero output. Mester (1991) proposes a new measure, Within-Sample Scope Economies, which avoids extrapolation.

$$WSC(y_1, \dots, y_N) = \frac{\{[f(y_1 - (N-1)y_1^{\min}, y_2^{\min}, \dots, y_N^{\min}) + f(y_1^{\min}, y_2 - (N-1)y_2^{\min}, y_3^{\min}, \dots, y_N^{\min}) + \dots \\ \dots + f(y_1^{\min}, \dots, y_{N-1}^{\min}, y_N - (N-1)y_N^{\min})] - f(y_1, y_2, \dots, y_N)\}}{f(y_1, y_2, \dots, y_N)} \quad (4)$$

where  $y_i^{\min}$  is the minimum value of  $y_i$  in the sample. The specialized firms in the within-sample measure produce positive amounts of each output but tend to specialize in one or the other.

The cost *X-inefficiency* of any bank  $i$  would be measured relative to the best-practice frontier. Note that the best-practice frontier refers to the best practice observed in the industry and not true minimum cost, which is not observable. Conceptually, the cost inefficiency of bank  $i$  measures the

percentage increase in cost of bank  $i$ , adjusted for random error, relative to the estimated cost needed to produce bank  $i$ 's output vector if the firm were as efficient as the best-practice firm in the sample facing the same exogenous variables  $(w,y,z,v)$ .<sup>8</sup> It can be thought of as the proportion of costs or resources that are used inefficiently or wasted. Figure 2 shows an example. The estimated cost frontier is given by  $\ln f$ . Bank  $j$  is fully efficient. Its actual cost lies below the frontier due to random error. Bank  $i$  is inefficient. The difference in bank  $i$ 's cost and the frontier value at the same  $y$  is due to both random error,  $v_i$ , and inefficiency,  $u_i$ . Cost inefficiency would include both technical inefficiency (operating in the interior of the production possibilities frontier) and allocative inefficiency (operating at a point on the production possibilities frontier that is not cost-minimizing).

If time-series or panel data are available, then *productivity growth* can be measured. Productivity growth is a combination of technological change, which is given by shifts in the frontier over time, and changes in inefficiency, which are changes in dispersion around the frontier. Berger and Mester (2003) define cost productivity growth as the change in cost from period  $t$  to period  $t+k$  holding constant the exogenous environmental variables, which they term “business conditions,” at their period  $t$  levels. It is important to control for these business conditions to avoid attributing a change in costs that is not due to bank managers' decisions or skill to a change in productivity.

**3.3. Profit Maximization.** The bank should minimize the cost of producing a given output bundle, but that output bundle should be chosen to maximize profits. Standard profit efficiency measures how close a firm is to producing the maximum possible profit given a particular level of input prices and output prices (and fixed netputs and environmental variables). In contrast to the cost function, the standard profit function specifies variable profits in place of variable costs and takes variable output prices as given, rather than holding all output quantities statistically fixed at their observed, possibly inefficient, levels. That is, the dependent variable in the profit function allows for consideration of revenues that can be earned by varying outputs as well as inputs. Output prices are taken as exogenous, allowing for inefficiencies in the choice of outputs when responding to these prices or to any other

---

<sup>8</sup> To see this, note that, ignoring random error,  $u_i = \ln C_i - \ln f(y_i, w_i, z_i, h_i)$ .

arguments of the profit function.

The standard profit function, in log form, is:

$$\ln (\pi+\theta)_i = \ln g(p_i, w_i, z_i, h_i) - u_{\pi i} + v_{\pi i}, \quad (5)$$

where  $\pi$  is the variable profits of the firm;  $\theta$  is a constant added to every firm's profit so that the natural log is taken of a positive number;  $p$  is the vector of prices of the variable outputs;  $v_{\pi i}$  represents random error; and  $u_{\pi i}$  represents inefficiency that reduces profits.

Similar to cost X-inefficiency, profit X-inefficiency is defined as that amount of profit that is not being earned compared to the predicted maximum profit that could be earned if the firm were as efficient as the best-practice firm. Thus, it is the percentage of profits that is left on the table, so to speak. Similar to cost productivity growth, profit productivity growth is the change in profit from period  $t$  to period  $t+k$  holding constant the exogenous environmental variables ("business conditions") at their period  $t$  levels.

As discussed in Berger and Mester (1997), profit efficiency is a more comprehensive measure of performance than is cost efficiency, since it accounts for errors on the output side as well as those on the input side. It is based on the economic goal of profit maximization, which requires that the same amount of managerial attention be paid to raising a marginal dollar of revenue as to reducing a marginal dollar of costs. That is, a firm that spends \$1 additional to raise revenues by \$2, all else held equal, would appropriately be measured as being more profit efficient but might inappropriately be measured as being less cost efficient. Note that cost efficiency evaluates performance, holding output constant at its current level, which generally will not correspond to an optimum. A firm that is relatively cost efficient at its current output may or may not be cost efficient at its optimal output, which typically involves a different scale and mix of outputs. Standard profit efficiency embodies the cost inefficiency deviations from the optimal point, as well as revenue inefficiencies.<sup>9</sup>

---

<sup>9</sup> Berger and Mester (1997) discuss another type of profit efficiency, alternative profit efficiency. This concept is based on estimates of the alternative profit function, which substitutes output levels for output prices in the specification of the profit function. This function is estimated to provide additional information when the maintained assumptions underlying the standard profit function do not hold. It may provide useful information if there are unmeasured differences in output qualities across firms; outputs are not completely variable; output markets are not perfectly competitive; or output prices are not accurately measured.

**3.4. More Complicated Objectives.** As discussed earlier, the standard concepts of cost minimization or profit maximization may not be the only goals being pursued by the firms' managers, and some studies have incorporated more complicated objectives. Explicitly recognizing the tradeoff between return and risk, where risk is a choice variable of the firm, would seem to be an important consideration for financial institutions (see Hughes, 1999; and Hughes, Lang, Mester, and Moon, 2000; and Hughes, Mester, and Moon, 2001). For example, an increase in a bank's scale of operations may allow it to reduce its exposure to both credit and liquidity risk through diversification. All else equal, this could mean scale economies in risk management costs. But all else is not equal: by reducing the risk attached to any given production plan, better diversification can decrease the marginal cost of risk-taking and lead banks to take on more risk to earn a greater return. Not accounting for risk when specifying the production structure can obscure scale economies, since additional risk-taking is costly in terms of the additional resources needed to manage the risk and the higher risk premium that has to be paid to attract uninsured funding. When exposure to risk is influenced by production decisions, then cost minimization and profit maximization need not coincide with value maximization. Estimates of efficiency that are derived from cost and profit functions may be mismeasured, since they do not penalize suboptimal choices of risk and quality that then affect prices. Moreover, if the managers are able to make choices in their own interest rather than on behalf of the owners of the firm (the stockholders), i.e., if the market for corporate control does not discipline managers, then the choices of risk versus return need not be value-maximizing either. Recognition that managers make decisions introduces the possibility of agency problems that also need to be considered in measuring efficiency.

If firms take risk as well as profit into account when making production decisions, then the model of production against which efficiency is evaluated would need to include this. Hughes, Lang, Mester, Moon (1996, 2000) construct a model of firm production that incorporates the risk-return tradeoff. Managers' most preferred production plan maximizes a utility function that accounts for how the probability distribution of profit depends on the production plan. Duality theory is used to derive the most preferred input and profit demand equations from the expenditure function. These demand

functions are those that maximize the managers' utility function. The managers' demand for financial capital can also be estimated along with the input and profit demand equations.<sup>10</sup>

Hughes, Mester, Moon (2001) develop measures of efficiency based on the expected return-risk tradeoff implied by the production model.  $ER$  is the firm's predicted profit, as calculated from the estimated profit-share equation from the model, divided by the firm's equity level.  $RK$  is the standard error of predicted profit divided by equity. The authors show that  $ER$  and  $RK$  are systematically related to the market value of equity for the subsample of publicly traded banks, so they can be used to derive market return efficiency measures. A risk-return frontier is then estimated:

$$ER_i = \Gamma_0 + \Gamma_1 RK_i + \Gamma_2 RK_i^2 + v_i - u_i, \quad (6)$$

where  $v_i$  is a two-sided error term representing random error, and  $u_i$  is a one-sided error term representing inefficiency. An inefficiency measure based on this frontier would give the increase in expected return that would occur if the firm moved to the frontier, holding risk constant. That is, it identifies lost potential return given the firm's level of return risk. One can identify the group of banks that are most efficient (say, the quarter of banks with the lowest levels of measured inefficiency) as those that are value-maximizing banks.<sup>11</sup>

We can generalize the efficient frontier given in equation (6) so that it applies to more complicated objectives (see Hughes, Lang, Mester, Moon, 2000). If  $X_i$  denotes a measure of the financial performance of firm  $i$ , e.g., profit, or the market value of its assets, and  $G_i$  denotes a measure defining the

---

<sup>10</sup> The functional forms for the utility-maximizing input and profit equations can be derived from the Almost Ideal Demand System. These equations are conditioned on the level of financial capital. A second stage can be added to the utility maximization problem to determine the bank managers' choice of financial capital, and this demand function can be estimated along with the input and profit demand equations.

<sup>11</sup> Hughes, Lang, Mester, and Moon (1996) present two other efficiency measures. Instead of holding risk constant and comparing the bank's expected return to the expected return it would have if it were on the frontier and had the same level of risk, these measures compare the bank's expected return and risk with the expected return and risk it would have if it moved to the frontier along the shortest path to the frontier. This shortest path is along the ray that is orthogonal to the frontier. These measures have a drawback in that they cannot account for random error's effect on the placement of the bank relative to the frontier.

Hughes, Mester, and Moon (2001) present two additional efficiency measures. For publicly traded bank holding companies they derive an efficiency measure based on estimating a frontier that relates the market value of assets to the book value of assets, and they derive another efficiency measure based on estimating a frontier that relates the market value of equity to the book value of equity. These measures indicate the bank holding company's lost potential market value of equity or assets based on the book value of equity or assets, respectively.

peer group used to compare firm  $i$ 's financial performance, e.g., risk or the market value of assets, the general form of the frontier, which gives the highest potential value of  $X_i$  given  $G_i$  is:

$$X_i = \alpha_0 + \alpha_1 G_i + \alpha_2 (G_i)^2 + v_i - u_i \quad (7)$$

where  $v_i$  is a two-sided random error term with zero mean, and  $u_i$  is a one-sided error term representing inefficiency. (Note that more flexible function forms than the quadratic could be specified.) For example, financial performance,  $X$ , might be measured by predicted profit from an estimated model and  $G$  might be measured by risk, e.g., the firm's interest-rate beta, or by size, e.g., its equity or asset level. Note that for any  $G$ , the optimality of the choice of  $G$  is not taken into account when measuring efficiency. That is, if  $G$  is risk, then a firm's performance would be compared only to those taking on the same level of risk. The firm would not be penalized for a suboptimal choice of risk that lowered performance.

*Expense-preference* is one particular form of X-inefficiency, in which firm managers are assumed to derive utility from choosing a greater than efficient (i.e., cost-minimizing or profit-maximizing) level of one or more of the firm's inputs, usually labor. That is, the managerial utility function is  $U=U(\pi, E)$ , where  $E$  represents expenditures on the input.

Tests for expense preference are based on estimating input demand functions or cost functions. The functional forms are derived explicitly from the utility function, which depends on the underlying production function of the firm. Edwards (1977) derived the demand for labor equation for a firm using a Cobb-Douglas production function and exhibiting expense preference for labor. Mester (1989b) generalizes expense-preference tests to allow for less restrictive production structures and the presence of expense-preference toward any input, not just labor. Note that the derived tests in both of these studies cannot give firm-specific measures of inefficiency. Rather they are tests of whether a group of firms is showing expense-preference toward any input.

#### **4. Measurement**

Even after the appropriate concept or goal against which efficiency is to be evaluated is chosen,

certain issues need to be confronted before the estimates can be obtained. These include estimation technique; specification of the functional form of the frontier; variables to include in the frontier; and data measurement issues.

**4.1 Estimation Techniques.** Different methods have been developed to identify the inefficiency component from the random noise component in frontier estimation. Common frontier efficiency estimation techniques are data envelopment analysis (DEA), free disposable hull analysis (FDH), the stochastic frontier approach, the thick frontier approach, and the distribution-free approach. The first two of these are nonparametric techniques, and the latter three are parametric methods (see Berger and Mester, 1997, for further discussion of these techniques).

My preference is for the parametric techniques. The nonparametric methods generally ignore prices and can, therefore, account only for technical inefficiency in using too many inputs or producing too few outputs (as discussed above). Another drawback is that they usually do not allow for random error in the data, assuming away measurement error and luck as factors affecting outcomes (although some progress is being made in this regard by using bootstrapping methods). In effect, they disentangle efficiency differences from random error by assuming that random error is zero! To see the effect of measurement error, consider Figure 3. The true data for a set of banks are given by the squares. The true frontier for this set of banks is indicated by the brown line. The measured data for these banks are indicated by the circles. The frontier determined by DEA using the measured data is given by the black line. Now consider Banks B and C. The researcher using DEA and ignoring measurement error would conclude that Bank C is not on the frontier and that Bank B is more efficient than Bank C. But the data are measured with error, and Bank C is actually more efficient than Bank B. The researcher would not know the true data but would need to allow for the possibility that the data are measured with error to avoid erroneous conclusions.

In the parametric methods, a bank is labeled inefficient if it is behaving less than optimally with respect to the specified goal – e.g., costs are higher or profits are lower – than the frontier value. The estimation methods differ in the way  $u_i$  is disentangled from the composite error term  $u_i + v_i$ . A drawback

of the parametric methods is that assumptions must be made about the shape of the frontier and the distribution of the inefficiency term. However, sufficient flexibility can usually be introduced so that the stochastic methods dominate the nonparametric methods in my opinion.

In the *stochastic frontier approach*, the inefficiency and random error components of the composite error term are disentangled by making explicit assumptions about their distributions. The random error term,  $v_i$ , is assumed to be two-sided (usually normally distributed), and the inefficiency term,  $u_i$ , is assumed to be one-sided (usually half-normally distributed). The parameters of the two distributions are estimated and can be used to obtain estimates of firm-specific inefficiency. The estimated mean of the conditional distribution of  $u_i$  given  $u_i + v_i$ , i.e.,  $\hat{u}_i \equiv \hat{E}(u_i | (u_i + v_i))$  is usually used to measure inefficiency. The distributional assumptions of the stochastic frontier approach are fairly arbitrary, and sometimes the residuals are not skewed in the direction predicted by the assumptions of the stochastic frontier approach, so that estimates are not obtainable.

If panel data are available, some of these maintained distributional assumptions can be relaxed, and the *distribution-free approach* may be used. This method assumes that there is a core efficiency or average efficiency for each firm over time. The core inefficiency is distinguished from random error (including any temporary fluctuations in inefficiency) by assuming that core inefficiency is persistent over time, while random errors tend to average out over time. In particular, a cost or profit function is estimated for each period of a panel data set. The residual in each separate regression is composed of both inefficiency,  $u_i$ , and random error,  $v_i$ , but the random component,  $v_i$ , is assumed to average out over time, so that an estimate of the inefficiency term,  $\hat{u} \equiv$  the average of a firm's residuals from all of the regressions  $= \text{average}(u_i + v_i) = \text{average}(u_i)$ .<sup>12</sup> The reasonableness of the maintained assumptions about the error term components depends on the length of the period studied. If too short a period is chosen, the random errors might not average out, in which case random error would be attributed to inefficiency (although truncation can help). If too long a period is chosen, the firm's core efficiency becomes less

meaningful because of changes in management and other events, i.e., it might not be constant over the time period.

**4.2 Functional Form, Variable Selection, and Variable Measurement.** The next step in the parametric estimation methods is choice of functional form for the frontier, including variable selection and measurement. The most popular form in the literature for cost and profit functions is the translog. The Fourier-flexible functional form augments the translog by including Fourier trigonometric terms, which makes it more flexible than the translog. Berger and Mester (1997) found that there was only a small difference in average efficiency and very little difference in efficiency dispersion or rank between cost or profit efficiency estimates based on the translog functional form and those based on the Fourier-flexible functional form. While formal statistical tests indicated that the coefficients on the Fourier terms were jointly significant at the 1 percent level, the average improvement in goodness of fit was small and was not significant from an economic point of view.

Once the objective and functional form are selected, the next decision is the variables to include in the function and proxies for those variables. Ideally, the frontier to be estimated should be derived from first principles. For example, if the objective is cost minimization, the cost function should be derived based on the specified production technology. Variables to include in the cost function would be those indicated by the theory of duality – output levels, input prices, netputs (factors that the firm cannot vary over the short run, which are measured in levels), and environmental variables (to account for differences across the firms' environments or markets, which may affect performance but are not a choice for firm management). For example, Hughes, Lang, Mester, and Moon (2000) derive the profit and input demand functions by applying Shephard's Lemma to the managerial expenditure function (based on the Almost Ideal Demand System), which is dual to the managerial utility maximization problem, in which managers trade off risk and return. These equations include revenue terms, the tax rate, and risk terms, which would not be included in the functions were the managers maximizing profits. Hence, the

---

<sup>12</sup> For banks with very low or very high  $\hat{u}$ , an adjustment (called truncation) is made to assign less extreme values of  $\hat{u}$  to these banks, since extreme values may indicate that random error,  $v_i$ , has not been completely purged by

coefficients on these terms offer a test of profit maximization versus utility maximization.<sup>13</sup> Other models would lead to other specifications.

In any of the estimation techniques, X-efficiency is essentially the residual. This means that omitted variables (or extraneous variables) can have large effects on measured efficiency. Specification of included variables is important, since the methodology depends on comparing the firm's cost or profit or market value, etc., to those of a best-practice firm operating at the same level of the exogenous variables included in the frontier. That is, the exogenous variables determine the reference set for the firm whose efficiency is being measured. If something extraneous is included in the frontier specification, then one might mislabel a firm as efficient because the estimation would be comparing firms in too narrow a reference set and not the entire set of relevant firms. For example, if two firms differ only in that one's CEO is blond and one is a brunette – which I'm assuming is unrelated to efficiency! – then we would want to consider these two firms in the same reference set and compare their costs to one another. If we included CEO hair color in the cost function as a dummy variable, we would preclude such a comparison. We might want to include in the specification of the frontier variables that account for differences in the environment in which the firm operates that are exogenous to the firm's decision-making but that may affect performance, e.g., we might want to include variables that account for demand, such as income growth in the firm's market, or whether the firm is located in an urban or rural market. Then in measuring efficiency, the urban firms would be compared to other urban firms and the rural firms to rural firms. But note that the manager's potentially inefficient choice of where to set up shop – in a rural or an urban market – would not be penalized. The alternative is to leave the variable out of the frontier specification but then determine whether the efficiency estimates are correlated with the variable. For example, Mester, 1993; Mester, 1996; and Mester, 1997; and Berger and Mester, 1997 have looked at correlations between efficiency measures and various exogenous factors. Judgment has to be used about the better way to proceed, including the variables as part of the frontier or excluding them and

---

averaging.

looking at correlations.

**4.3 Special Issues in Banking.** Judgment also has to be used when applying efficiency techniques to certain industries. The special issues that arise in applying the techniques to the banking industry are suggestive of some of the problems and issues that can arise in efficiency estimation in general. In banking, an important issue has been how to measure outputs and inputs. There has been some disagreement in the literature over what a commercial bank is actually producing. Two general approaches have been taken: the “production” approach and the “intermediation” approach (also called the “asset” approach).

The production approach focuses on the bank’s operating costs, i.e., the costs of labor (employees) and physical capital (plant and equipment). The bank’s outputs are measured by the number of each type of account, such as commercial and industrial loans, mortgages, deposits, etc. because it is thought that most of the operating costs are incurred by processing account documents and debiting and crediting accounts; inputs are labor and physical capital.

The “intermediation” approach considers a financial firm’s production process to be one of financial intermediation, i.e., the borrowing of funds and the subsequent lending of those funds. Thus, the focus is on total costs, including both interest and operating expenses. Outputs are measured by the dollar volume of each of the bank’s different types of loans, and inputs are labor, physical capital, deposits and other borrowed funds, and, in some studies, financial capital.<sup>14</sup> The studies on X-efficiency in banking have

---

<sup>13</sup> Hughes, Lang, Mester, and Moon (2000) reject the hypothesis of profit maximization using 1989-1990 data on U.S. banks that reported at least \$1 billion in assets as of the last quarter of 1998.

<sup>14</sup> A slight variation on the intermediation approach, which has been used in some studies, is to distinguish between transactions deposits, which are treated as an output, since they can serve as a measure of the amount of transactions services the bank produces, and purchased or borrowed funds (such as federal funds or large CDs purchased from another bank), which are treated as inputs, since the bank does not produce services in obtaining these funds. The strict intermediation approach would consider the transactions services produced by the bank as an intermediate output, something that must be produced along the way toward the bank’s final output of earning assets. Hughes and Mester (1993) empirically tested whether deposits should be treated as an input or output and found support that they should be treated as an input in their study.

Another approach that has been taken less often is the “value-added” approach, which considers all liabilities and assets of the bank to have at least some of the characteristics of an output. Still another approach, taken in Mester (1992), is to consider the bank’s output to be its loan origination and loan monitoring services, since these outputs are more closely related to the theory of financial intermediation.

tended to use the intermediation approach.<sup>15</sup>

Theoretically, to compare one firm's efficiency to another's, we would like to compare each firm's cost of producing the *same* outputs. For banks, significant characteristics are loan quality, which reflects the amount of monitoring the bank does to keep the loan performing, and the riskiness of the bank's portfolio. Unless these characteristics are controlled for, one might conclude a bank was producing in a very efficient manner if it were spending far less to produce a given output level, but its output might be highly risky and of a lower quality than that of another bank. It would be wrong to say a bank was efficient if it were scrimping on the credit evaluation needed to produce sound loans. Thus, recent studies have included quality and nonperforming loans in the specifications of cost and profit functions. Hughes, Lang, Mester, and Moon (2000) derive the risk-return tradeoff explicitly from a utility maximization model rather than just augmenting the cost and profit functions with risk and quality measures. See Hughes (1999) for further discussion.

Unfortunately, there are likely to be unmeasured differences in quality because the banking data do not fully capture the heterogeneity in bank output. The amount of service *flow* associated with financial products is by necessity usually assumed to be proportionate to the dollar value of the *stock* of assets or liabilities on the balance sheet, which can result in significant mismeasurement. For example, commercial loans can vary in size, repayment schedule, risk, transparency of information, type of collateral, covenants to be enforced, etc. These differences are likely to affect the costs to the bank of loan origination, ongoing monitoring and control, and financing expense. Unmeasured differences in product quality may be incorrectly measured as differences in cost inefficiency.

Another issue raised in recent papers in the bank efficiency literature is the treatment of financial capital.<sup>16</sup> As discussed in Berger and Mester (1997), a bank's insolvency risk depends not only on the riskiness of its portfolio but on the amount of financial capital it has to absorb losses. Insolvency risk

---

<sup>15</sup> As discussed earlier, Mester (1992) took a different approach and specified outputs that were more directly related to the monitoring and screening services performed by the bank: loans originated, loans purchased, loans originated or purchased earlier and held on balance sheet, and loans sold.

<sup>16</sup> The discussion of the role of financial capital is taken mainly from Berger and Mester (1997).

affects bank costs and profits via the risk premium the bank has to pay for uninsured debt, through the intensity of risk management activities the bank undertakes and (as discussed in Hughes, 1999, and Hughes, Lang, Mester, and Moon, 2000) through the discount rate applied to future profits. Thus, the bank's financial capital should be considered when studying efficiency. To some extent, controlling for the interest rates paid on uninsured debt helps account for differences in risk, but these rates are imperfectly measured.

Even apart from risk, a bank's capital level directly affects costs by providing an alternative to deposits as a funding source for loans. In most studies, interest paid on debt (deposits) is counted as a cost, but dividends paid are not. On the other hand, raising equity typically involves higher costs than raising deposits. If the first effect dominates, measured costs will be higher for banks using a higher proportion of debt financing; if the second effect dominates, measured costs will be lower for these banks.

Studies that have considered financial capital include the level of capital rather than its price. Including the price assumes that banks on the frontier are selecting the cost-minimizing level of capital. This might not be the case because of regulations that set a minimum capital-to-asset ratio or because of risk-aversion on the part of bank managers. See Hughes and Mester (1993) for further discussion.

To summarize, the review above discusses some of the steps that need to be followed in implementing efficiency measurement. The main steps involve choosing the efficiency concept, i.e., firm objective function (this includes specification of the production function of the firm), estimation technique, functional form, and variables and their proxies.

## **5. Empirical Findings in the Literature**

There is a vast literature on efficiency at commercial banks, and there have been several comprehensive reviews of the literature, e.g., Berger, Hunter, and Timme (1993), Berger and Humphrey (1997), and Berger (2003). Here, I focus on several overall impressions that can be drawn from the literature rather than a comprehensive review.

**5.1 Scale Economies.** The evidence on scale economies has been changing over time as more

complicated and realistic models have been applied to the data. Early studies using data from the 1980s failed to find scale economies beyond a very small bank size – up to about \$100 million in assets. Later studies using data from the 1990s have found scale economies in a range of up to about \$10 billion. And the latest studies (e.g., Berger and Mester, 1997; Hughes, Mester, Moon, 2001; Bossone and Lee, 2004) which incorporate banks' risk preferences and financial capital into bank production models find scale economies for the very largest banks in the sample, up to at least \$25 billion in assets. For example, Berger and Mester (1997) incorporated asset quality and financial capital into the cost function, and using the sample of almost 6,000 U.S. commercial banks that were in continuous existence over the six-year period 1990-1995, found significant cost scale economies for banks in each size class, with estimates suggesting that the typical bank would have to be two to three times larger in order to maximize cost scale efficiency for its product mix and input prices.

The difference in results between the earlier and later studies may partly reflect improvements in the technologies used for bank intermediation and the relaxation of geographic restrictions on competition. Improvements in information processing, automated loan systems, and credit scoring may have reduced costs of extending loans more for large banks than for smaller banks. The removal of geographic branching restrictions may have made it less costly to become large.

There also may be some measurement issues involved. Studies that have focused on smaller banks and studies that have focused on larger banks have tended to find scale economies exhausted at different sizes. For example, studies that used only banks with under \$1 billion in assets (and used the standard approach which did not incorporate risk or financial capital) usually found average costs to be minimized between about \$75 million and \$300 million in assets, while studies that used only banks with over \$1 billion in assets usually found the minimum average cost point to be between \$2 billion and \$10 billion in assets (see Berger, Hunter, and Timme, 1993). This suggests that a single function may not be able to incorporate both large and small bank technologies or that some important factor that varies with bank size is excluded from the model. There is conflicting evidence on this point. McAllister and McManus (1993) found that the translog is not a good global approximation to banks of all sizes. Berger

and Mester (1997) found that while the coefficients on the Fourier terms in the Fourier-flexible functional form were jointly significantly different from zero, the improvement in the goodness of fit of the Fourier over the translog was small and not economically significant. Both functional forms yield essentially the same average level and dispersion of measured efficiency and both ranked the individual banks in almost the same order.

But the later studies' finding of significant scale economies likely also reflects improvements in the methods used to measure scale economies – in particular, accounting for the bank's choice of risk and financial capital. As discussed in Hughes, Mester, and Moon (2001), the standard model ignores the fact that bank risk is endogenous. A larger scale of operations may allow the bank to be better diversified. Better diversification can lead to reduced liquidity risk on the liability side of the balance sheet and reduced credit risk on the asset side of the balance sheet, which can mean reduced costs of risk management. The bank might be able to economize on financial capital, a relatively expensive source of funds, to the extent that diversification lowers banks' insolvency risk. Also, the cost of funds might decline as banks grow in size if large depositors and other creditors perceive that regulators consider some banks are "too-big-to-fail."<sup>17</sup>

Better diversification leading to reduced marginal cost of risk-taking and reduced marginal cost of risk management, all else equal, is the usual diversification effect. But all else is not necessarily equal, because risk-taking is endogenous. Banks might respond to the lower cost of risk management by taking on more risk. In turn, banks may have to spend more to manage the increased risk. This risk-taking effect may offset the diversification effect, and the potential economies that follow from scale-related diversification may be obscured. Thus, to unmask scale economies due to better diversification it is important to incorporate risk into the analysis. It is also important to account for the fact that bank managers need not be holding the level of financial capital that minimizes costs. As discussed in Hughes and Mester (1998), financial capital is the bank's own bet on its management of risk, so it provides a

---

<sup>17</sup> Hughes and Mester (1993) find evidence of "too-big-to-fail": for large banks, an increase in size, holding default risk and asset quality constant is associated with a significantly lower price of uninsured deposits.

credible signal to depositors and creditors of the resources allocated to preserve capital and reduce insolvency risk. As a bank's scale increases, its loan portfolio and deposit base become more diversified. Diversification reduces the cost of the signaling, since the same degree of protection against financial distress can be attained at a lower capital-to-asset ratio. Larger scale also reduces the level of the signal required, to the extent that outsiders infer the bank's level of diversification from the bank's scale of operations, which is observable.

Using 1989 and 1990 data on U.S. banks with assets over \$1 billion, Hughes and Mester (1998) find evidence that: financial capital is a signal of risk, banks do not hold the cost-minimizing level of capital, the level of capitalization increases less than proportionally with assets, and there are significant scale economies at even the largest banks in the sample (which is \$74 billion).

Hughes, Mester, and Moon (2001) undertake a systematic study of bank cost models and find that estimated scale economies depend critically on how banks' capital structure and risk-taking are modeled. Using 1994 data on highest-level bank holding companies in the U.S., they find that a standard cost function that omits equity capital and a standard cost function that incorporates capital structure and the cost of capital both generally yield estimates of constant returns to scale across bank holding companies in the sample. However, regressing the bank-specific scale economies measures on variables accounting for sources of risk-taking and diversification, they show that better diversification is associated with larger scale economies, while increased risk-taking is related to smaller scale economies. They also find that a proportional variation in size and diversification, controlling for sources of risk-taking, yields a statistically and economically significant increase in scale economies, and that by the criterion of cost minimization, smaller banks overutilize capital, while larger banks underutilize capital. These results suggest that scale economies might be masked by the banks' endogenous choice of risk, which needs to be modeled.

Hughes, Mester, and Moon (2001) verify this by estimating the managers' most-preferred production model that includes equity capital, in addition to debt, and models bank managers as maximizing utility as a function of expected profits and risk. This allows banks to be value-maximizers

rather than profit-maximizers and allows the bank's production choices to reflect risk management concerns. Calculating the change in cost as output is expanded so as to maximize utility, they find that banks have large scale economies that increase with size. Since agency problems between owners and managers might mean utility-maximizing managers might not choose value-maximizing production plans, the value-maximizing banks are identified as those that make efficient risk-return tradeoffs. Restricting attention to the most efficient quarter of banks in each of five size groups and calculating the change in cost as output is expanded so as to maximize utility, they again find that banks have large scale economies that increase with size. By incorporating capital structure and risk-taking into models of bank production they have uncovered the scale economies that are often cited by merging banks but that can be obscured in the standard models, which ignore the endogeneity of the bank's choice of risk.

Hughes, Lang, Mester, and Moon (2000) also measure scale economies along the value-maximizing expansion path using 1990 data on banks with greater than \$1 billion in assets. Banks in all size quartiles were found to be operating with significant scale economies.

Bossone and Lee (2004) apply the methods of Hughes and Mester (1998), and Hughes, Mester, and Moon (2001) to study the relationship between productive efficiency and the size of the financial system. Using a sample of 875 commercial banks from 75 countries, they estimate a cost function and measure scale economies, allowing for the banks' endogenous choice of risk and financial capital. (For comparison, they also estimate the standard measure of scale economies, which does not incorporate risk and financial capital). Size of the financial system is proxied by three measures: absolute size, which is the sum of domestic credit, domestic deposits, foreign assets, and foreign liabilities of the banking system; relative size or financial depth, which is the ratio of absolute size to the level of GDP; and financial market size, which is  $\text{stock market capitalization to GDP} \times \text{stock market total value traded to GDP} \times \text{stock market turnover to GDP}$ . They find the presence of significant scale economies that are increasing with the size of the financial system, for each of the three measures of size. (Similar to the results in Hughes, Mester, and Moon (2001), these scale economies are not uncovered using the standard cost function, which doesn't incorporate risk and financial capital.) They also find that small banks in larger financial systems are more cost efficient

than small banks in small systems, and that scale economies are less variable across bank size, holding the financial system size constant, than they are across financial system size, holding bank size constant. They interpret their findings as evidence of what they call “systemic scale economies,” that is, economies derived from operating in a larger financial system. For example, it might be less costly for a bank operating in a large financial system if a larger payment system charges lower fees to banks using its services or if a larger financial system makes it easier to diversify across products or geography, thereby allowing banks to save on capital costs.

The papers discussed suggest that scale can confer economic benefits. But the degree of benefits can vary across the type of expansion. Hughes, Lang, Mester, and Moon (1999) find that the economic benefits of consolidation are strongest for those banks engaged in interstate expansion and, in particular, interstate expansion that diversifies banks’ macroeconomic risk. Hughes, Lang, Mester Moon, and Pagano (2003) find evidence that an increase in assets by internal growth (in contrast to acquisition) is associated with better performance at most banks, consistent with the existence of scale economies. They also find that at banks without entrenched management, both asset acquisitions (e.g., via merger) and asset sales are associated with improved performance, but at banks with entrenched managers, asset sales are associated with smaller improvements, and asset acquisitions are associated with worse performance.<sup>18</sup> This suggests that while there are value-enhancing incentives to merge, they may be subordinated to the incentives to build larger institutions from which entrenched managers can gain perquisites.

**5.2 Scope Economies.** Most studies have not found strong evidence of scope economies, either between traditional commercial banking products or between on-balance-sheet and off-balance-sheet bank products. This is not to say that deregulation that permits banks to expand the types of products they can offer could not enable banks to take advantage of potential scope economies. Still, it is difficult to find evidence of strong scope economies in the literature, with a few exceptions. Mester (1991) found evidence

---

<sup>18</sup> Entrenchment is found to be related to higher levels of managerial ownership, better investment opportunities, higher inefficiency, and smaller asset size.

of diseconomies of scope for mutual savings and loans using 1982 data on California S&Ls, but Mester (1993) using 1991 data on U.S. S&Ls found scope economies between traditional outputs – these results are consistent with the hypothesis that the removal of interest-rate ceilings in 1986 reduced the ability of mutual S&L managers to pursue their own goals.

Mester (1992) measures outputs based on an information-theoretic approach – loan origination, monitoring, selling, and buying. These outputs involve different levels of credit evaluation and loan monitoring. Loans originated or purchased before the current date  $t$  and loans originated at  $t$  and held are the traditional outputs of a bank. Loans bought at time  $t$  and loans originated at time  $t$  and sold are less traditional. She finds diseconomies of scope between the traditional banking services and nontraditional services.

Berger, Hancock, and Humphrey (1993), using data on U.S. banks from 1984-1989, find evidence of scope economies based on the profit function. The profit function measure takes into account not only cost gains from joint production but also revenue gains perhaps derived from cross-selling. They test whether the optimal quantity of every output is positive for all the price vectors observed in the data and find that for most firms, this is true. This contrasts with Berger, Humphrey, and Pulley (1996), who estimated a revenue function using 1978-1990 data on U.S. banks and found no evidence of revenue scope economies between loans and deposits.

**5.3 X-Efficiency.** Research on cost X-efficiency in banking generally finds large inefficiencies on the order of 20 to 25 percent or more of total banking industry costs when the stochastic methods are used.<sup>19</sup> That is, achievement of X-efficiency (elimination of X-inefficiency) at the average bank could produce about a 20 to 25 percent cost savings, making this an important source of inefficiency in banking. Although 20 percent seems quite large, and perhaps too large to sustain in a reasonably competitive industry, I note that similar levels of inefficiency are found in studies of manufacturing and other industries. The conclusion from the earlier literature that found constant returns to scale but high levels of X-

---

<sup>19</sup> When the nonparametric DEA method is used, there is a greater range of findings from 10 percent to 50 percent.

inefficiency was that managerial inefficiencies outweighed the inefficiencies related to scale and scope. However, with the latest studies finding significant scale economies, this conclusion need not be the case.

Berger and Mester (1997) estimate both cost X-efficiencies and profit X-efficiencies. The mean cost efficiency from their preferred model is 0.868 suggesting that about 13.2 percent of cost is wasted on average relative to a best-practice firm. The mean profit efficiency is much lower, suggesting that 50 percent of potential profits that could be earned by a best-practice bank are lost to inefficiency. They also find that there is considerably more variation in profit inefficiency among the banks than in cost inefficiency, with many banks achieving higher or lower profit efficiency than the average. As with scale economies measurement, how financial capital is modeled affects estimates of X-efficiency. They find that profit X-efficiency was much lower when equity capital was excluded in the profit function. Instead of 50 percent inefficiency, the estimates indicate 90 percent inefficiency.

Berger and Mester (1997) also investigate the relationship between their X-efficiency estimates and various aspects of the banks, their markets, and their regulation that are potential correlates of efficiency that are at least partially exogenous. The characteristics investigated fall into six categories: bank size, organizational form and corporate governance, other bank characteristics, market characteristics, state geographic restrictions on competition, and primary federal regulator. Both multiple regression and single variable regressions were estimated. A few robust relationships were uncovered. Large and small banks appear to be equally cost X-efficient but large banks are less profit X-efficient suggesting it is harder to efficiently generate revenues as a bank grows in size. Higher risk, as measured by the standard deviation of return on assets, is associated with lower X-efficiency. Greater market power is associated with lower cost X-efficiency and greater profit X-efficiency. But the basic conclusion from this analysis was that the correlates of efficiency are still largely unknown: 25 explanatory variables explain only about 7 percent of the variance of measured cost efficiency and 35 percent of the variables of measured profit efficiency.

Hughes, Lang, Mester, Moon, and Pagano (2002) measure market-value inefficiency by the bank's shortfall ratio, which gives the shortfall of a bank's market value from its highest potential market value as a proportion of the bank's book-value investment in its assets, net of goodwill. The measure is

derived by stochastic frontier techniques to fit the frontier of market value on book value (i.e., replacement cost) of assets. They find an average shortfall of 19 percent.

Koetter (2004) studies the efficiency of German banks over the period 1995-2001 using the managerial utility-maximization model of Hughes, Lang, Mester, and Moon (2000). He finds average inefficiency measured relative to the risk-return frontier to be quite low, around 5 percent.

**5.4 Productivity.** There are fewer studies of productivity in banking. Using data on banks from the late 1970s and 1980s, most studies find negative cost productivity growth, on the order of  $-1$  percent per year. Using panel data on 661 top-tier bank holding companies continuously in existence during 1991-1997, Stiroh (2000) found small cost productivity improvements of between 0.05 percent and 0.47 percent annually depending on the definitions of output and method of measurement. But the literature suggests that bank size matters. Some studies find increased productivity growth (in terms of costs or profits) in the early 1980s for large banks (due to shifts in the best-practice frontier) but not for small banks. For example, Humphrey and Pulley (1997) found that profits of larger banks in the sample (with assets over \$500 million) increased by 12 percent between the 1977-1981 period and the 1981-1984 period. Decomposing this change, they found that it results from a shift in the profit function and changes in business condition, particularly deposit deregulation. Only business conditions accounted for the risk in large banks' profits from 1981-1984 to 1985-1998. For smaller banks (assets of \$100 million to \$500 million), there was little increase in profits between 1977-1981 and 1971-1984. Wheelock and Wilson (1999) used linear programming techniques (DEA) and decomposed the change in productivity into the change in efficiency and the shift in efficient frontier. They found that banks on the frontier improved over the period 1984-1993 but productivity declined, on average, during this period because of reductions in efficiency. Smaller banks (assets below \$300 million), in particular, were unable to adapt to changes in technology, regulation, and competitive condition and fell further away from the efficient frontier.<sup>20</sup>

Berger and Mester (2003) look at both cost and profit productivity, where productivity is measured as a combination of technological change and changes in inefficiency, holding constant the

exogenous environmental variables. They find that during 1991-1997, cost productivity in the banking industry worsened while profit productivity improved substantially and concluded this was because revenue-based productivity changes are not accounted for in measuring cost productivity. Banks have been offering wider varieties of financial services and have been providing additional convenience, which may have raised costs but also raised revenues by more than the cost increases. They also found that banks involved in merger activity might be responsible for their main findings. The merging banks had greater cost productivity deterioration and profit productivity improvements than other banks. Merging banks may have also improved their profit performance, on average, by shifting their portfolios into investments with higher risk and higher expected return to take advantage of the diversification gains from mergers, as suggested by the work of Hughes, Lang, Mester, and Moon (1996) and Hughes, Mester, and Moon (2001).

## **6. Conclusions**

One goal of the research agenda on optimal bank productive efficiency is to answer some fundamental questions about financial industry restructuring. The results from this literature shed light on the consolidation trend in the commercial banking industry and suggest some answers to the three conundrums posed in the introduction.

Conundrum 1: In contrast to the earlier literature, new bank production models that incorporate banks' choice of risk and financial capital and that explicitly consider how banks' production decisions influence their riskiness have uncovered scale economies at very large banks. This is consistent with the consolidation trend, which is creating very large banks and helps resolve the inconsistency between the earlier literature's finding of constant returns to scale and the reality of consolidation. Diversification benefits appear to be a source of these scale economies. The cost of risk-taking decreases with size. If banks respond to the reduced price by taking on more risk, then the standard models would not be able to uncover scale economies.

---

<sup>20</sup> Berger and Mester (2003) discuss several other studies of bank productivity.

Conundrum 2: There is little evidence of scope economies in the literature, which may explain why banks have not responded to the Gramm-Leach-Bliley Act's relaxation of the barriers to offering nontraditional activities along with traditional commercial bank activities. This can only be a tentative conclusion, however, since the literature on this topic is thin. Partly this reflects a lack of data on institutions that are mixing these products, since the restrictions on product mix have only recently been repealed.

Conundrum 3: It is true that banks experienced a worsening of cost productivity in the 1990s. This might seem at odds with the technological changes that have occurred in banking. But a focus on cost productivity is misleading. At the same time cost productivity worsened, banks experienced an increase in profit productivity. This is consistent with banks' offering wider varieties of financial services and additional convenience, which may have raised costs but raised revenues more. Merging banks had greater cost productivity deterioration and profit productivity improvements than other banks. Their better profit productivity gains might also reflect their ability to take advantage of diversification benefits.

Figure 1

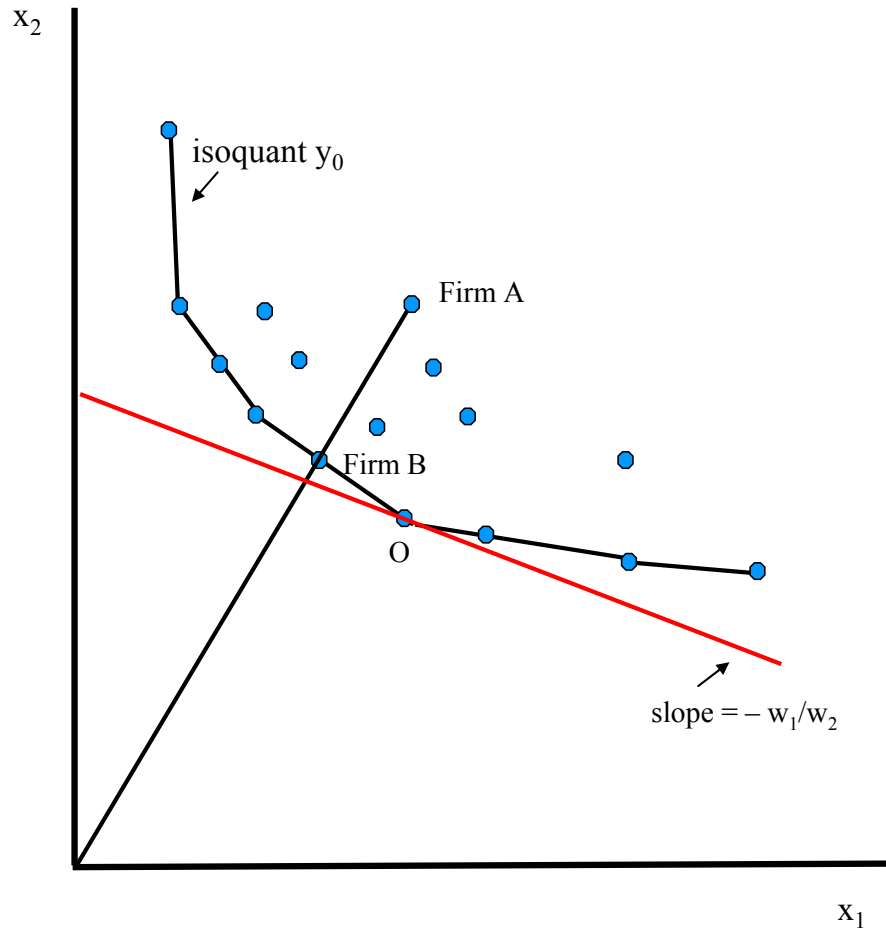


Figure 2

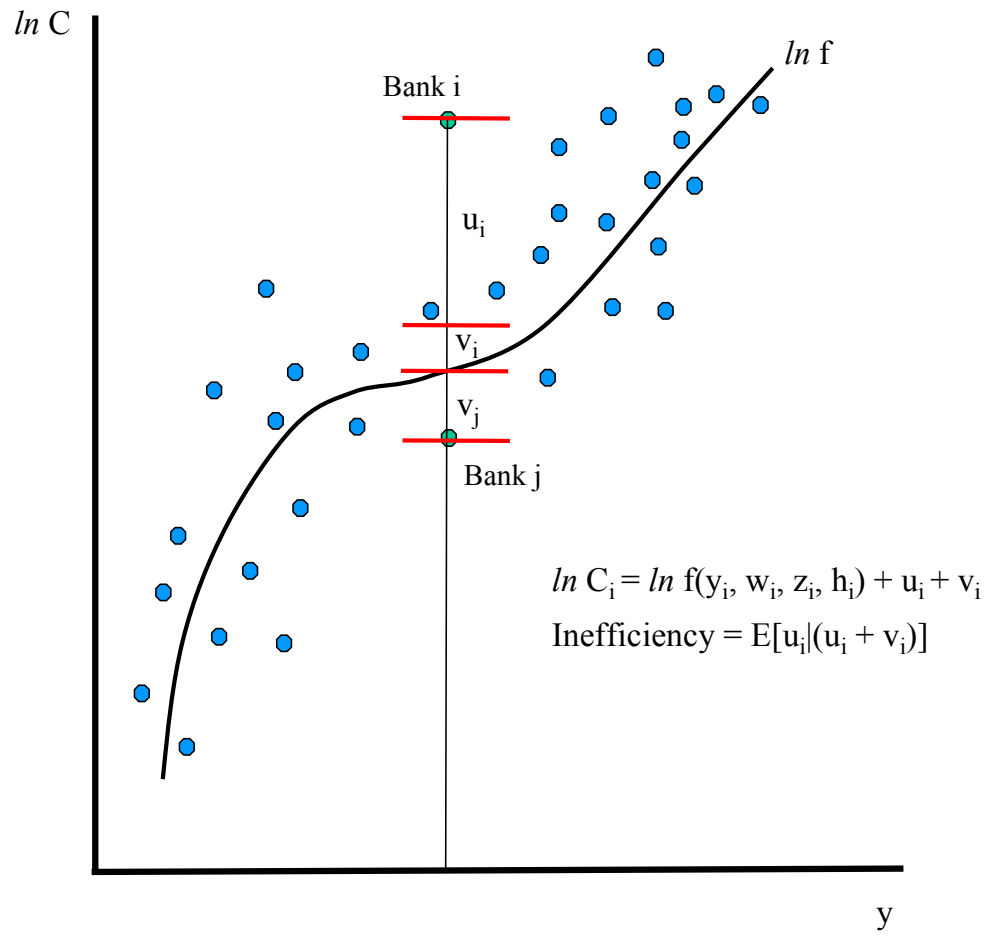
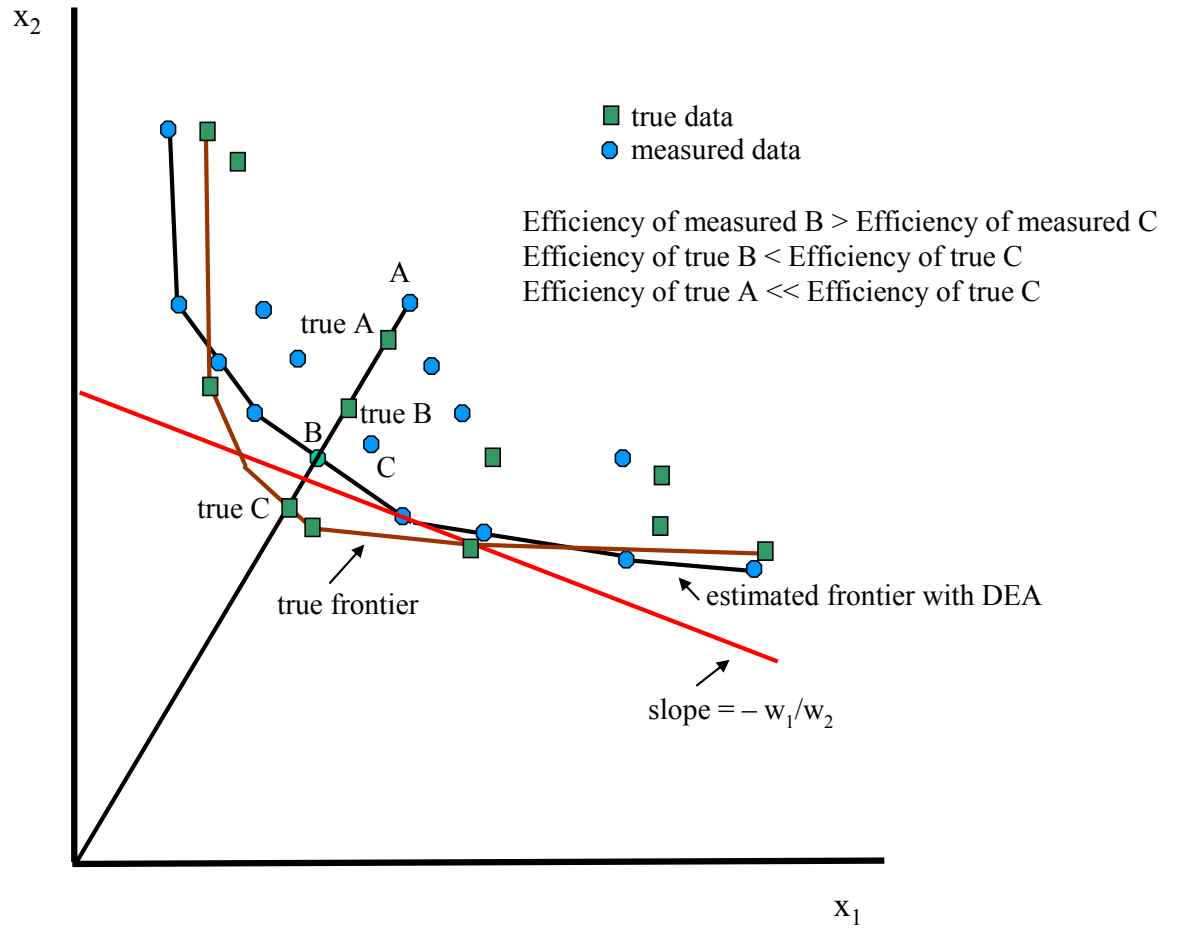


Figure 3



## References

- Akhavein, Jalal D., Allen N. Berger, and David B. Humphrey, "The Effects of Megamergers on Efficiency and Prices: Evidence From a Bank Profit Function," *Review of Industrial Organization*, 12 (February 1997), pp. 95-139.
- Baumol, William J., John C. Panzar, Robert D. Willig, *Contestable Markets and the Theory of Industry Structure*, New York: Harcourt Brace Jovanovich, 1982.
- Benston, George J., William C. Hunter, and Larry D. Wall, "Motivations for Bank Mergers and Acquisitions: Enhancing the Deposit Insurance Put Option versus Earnings Diversification," *Journal of Money, Credit and Banking*, 27 (August 1995), pp. 777-788.
- Berger, Allen N., "The Economic Effects of Technological Progress: Evidence from the Banking Industry," *Journal of Money, Credit, and Banking*, 35 (April 2003), pp. 141-176.
- Berger, Allen N., Diana Hancock, and David B. Humphrey, "Bank Efficiency Derived from the Profit Function," *Journal of Banking and Finance*, 17 (1993), pp. 317-347.
- Berger, Allen N., and David B. Humphrey, "Efficiency of Financial Institutions: International Survey and Directions for Future Research," *European Journal of Operational Research* (1997), pp. 175-212.
- Berger, Allen N., David B. Humphrey, and Lawrence B. Pulley, "Do Consumers Pay for One-Stop Banking? Evidence from an Alternative Revenue Function," *Journal of Banking and Finance*, 20 (1996), pp. 1601-1621.
- Berger, Allen N., William C. Hunter, and Stephen G. Timme, "The Efficiency of Financial Institutions: A Review and Preview of Research Past, Present, and Future," *Journal of Banking and Finance*, 17 (1993), pp. 221-249.
- Berger, Allen N., and Loretta J. Mester, "Inside the Black Box: What Explains Differences in the Efficiencies of Financial Institutions?" *Journal of Banking and Finance* 21 (July 1997), pp. 895-947.
- Berger, Allen N., and Loretta J. Mester, "Explaining the Dramatic Changes in Performance of U.S. Banks: Technological Change, Deregulation, and Dynamic Changes in Competition," *Journal of*

- Financial Intermediation*, 12 (2003), pp. 57-95.
- Berlin, Mitchell, and Loretta J. Mester, "On the Profitability and Cost of Relationship Lending," *Journal of Banking and Finance*, 22 (August 1998), pp. 873-897.
- Bossone, Biagio, and Jong-Kun Lee, "In Finance, Size Matters: The 'Systemic Scale Economies' Hypothesis," *IMF Staff Papers*, 51 (April 2004), pp. 19-46.
- Edwards, Franklin R., "Managerial Objectives in Regulated Industries: Expense-Preference Behavior in Banking," *Journal of Political Economy*, 85 (February 1977), pp. 147-162.
- FDIC's Historical Statistics on Banking, Table CB02, Changes in Number of Institutions, FDIC-Insured Commercial Banks, January 14, 2005.
- Group of Ten, "Report on Consolidation in the Financial Sector," January 2001 ([www.bis.org](http://www.bis.org)).
- Hughes, Joseph P., "Incorporating Risk Into the Analysis of Production," Presidential Address to the Atlantic Economic Society, *Atlantic Economic Journal* 27, (1999), pp. 1-23.
- Hughes, Joseph P., William W. Lang, Loretta J. Mester, and Choon-Geol Moon, "Efficient Banking Under Interstate Branching," *Journal of Money, Credit, and Banking*, 28 (November 1996), pp. 1043-1071.
- Hughes, Joseph P., William Lang, Loretta J. Mester, and Choon-Geol Moon, "The Dollars and Sense of Bank Consolidation," *Journal of Banking and Finance*, 23 (February 1999), pp. 291-324.
- Hughes, Joseph P., William W. Lang, Loretta J. Mester, and Choon-Geol Moon, "Recovering Risky Technologies Using the Almost Ideal Demand System: An Application to U.S. Banks," *Journal of Financial Services Research*, 18 (October 2000), pp. 5-27.
- Hughes, Joseph P., and Loretta J. Mester, "A Quality and Risk-Adjusted Cost Function for Banks: Evidence on the 'Too-Big-To-Fail' Doctrine," *Journal of Productivity Analysis*, 4 (September 1993), pp. 293-315.
- Hughes, Joseph P., and Loretta J. Mester, "Bank Capitalization and Cost: Evidence of Scale Economies in Risk Management and Signaling," *The Review of Economics and Statistics*, 80 (May 1998), pp. 314-325.

- Hughes, Joseph P., William W. Lang, Loretta J. Mester, Choon-Geol Moon, and Michael S. Pagano, "Do Banks Sacrifice Value to Build Empires? Managerial Incentives, Industry Consolidation, and Financial Performance," *Journal of Banking and Finance*, 27 (2003), pp. 417-447.
- Hughes, Joseph P., Loretta J. Mester, and Choon-Geol Moon, "Are Scale Economies in Banking Elusive or Illusive? Evidence Obtained by Incorporating Capital Structure and Risk-Taking into Models of Bank Production Checking Accounts and Bank Monitoring," *Journal of Banking and Finance*, 25 (December 2001), pp. 2169-2208.
- Humphrey, David B., and Lawrence B. Pulley, "Banks' Responses to Deregulation: Profits, Technology, and Efficiency," *Journal of Money, Credit, and Banking*, 29 (1997), pp. 73-93.
- Koetter, Michael, "The Stability of Efficiency Rankings When Risk-Preferences Are Different," Tjalling C. Koppmans Research Institute Discussion Paper No. 04-08, Utrecht School of Economics, University of Utrecht, January 2004.
- McAllister, Patrick H. and Douglas McManus, "Resolving the Scale Efficiency Puzzle in Banking," *Journal of Banking and Finance*, 17 (1993), pp. 389-405.
- Mester, Loretta J., "Owners versus Managers: Who Controls the Bank?" *Business Review*, Federal Reserve Bank of Philadelphia (May/June 1989a), pp. 13-23.
- Mester, Loretta J., "Testing for Expense Preference Behavior: Mutual versus Stock Savings and Loans," *RAND Journal of Economics*, 20 (Winter 1989b), pp. 483-498.
- Mester, Loretta J., "Agency Costs Among Savings and Loans," *Journal of Financial Intermediation*, 1 (June 1991), pp. 257-278.
- Mester, Loretta J., "Traditional and Nontraditional Banking: An Information-Theoretic Approach," *Journal of Banking and Finance*, 16 (June 1992), pp. 545-566.
- Mester, Loretta J., "Measuring Efficiency at U.S. Banks: Accounting for Heterogeneity Is Important," *European Journal of Operational Research*, 98 (April 1997), pp. 230-242.
- Mester, Loretta J., "A Study of Bank Efficiency Taking Into Account Risk-Preferences," *Journal of Banking and Finance*, 20 (July 1996), pp. 1025-1045.

- Mester, Loretta J., "Efficiency in the Savings and Loan Industry," *Journal of Banking and Finance*, 17 (April 1993), pp. 267-286.
- Mester, Loretta J., "Applying Efficiency Measurement Techniques to Central Banks," Working Paper No. 03-13, Federal Reserve Bank of Philadelphia, July 2003.
- Pilloff, Steven J., and Anthony M. Santomero, "The Value Effects of Bank Mergers and Acquisitions," in Y. Amihud and G. Miller, eds., *Bank Mergers and Acquisitions* (Boston: Kluwer Academic Publishers), 1998.
- Rhoades, Stephen A., "Bank Mergers and Banking Structure in the United States, 1980-98," Staff Study 174, Board of Governors of the Federal Reserve System, August 2000.
- Stiroh, Kevin J., "How Did Bank Holding Companies Prosper in the 1990s?" *Journal of Banking and Finance*, 24 (2000), pp. 1703-1745.
- Vander Vennet, Rudi, "The Effect of Mergers and Acquisitions on the Efficiency and Profitability of EC Credit Institutions," *Journal of Banking and Finance*, 20 (November 1996), pp. 1531-1558.
- Wheelock, David C., and Paul W. Wilson, "Technical Progress, Inefficiency, and Productivity Change in U.S. Banking, 1984-1993," *Journal of Money, Credit, and Banking*, 31 (1999), pp. 212-234.