FEDERAL RESERVE BANK OF PHILADELPHIA

Ten Independence Mall
Philadelphia, Pennsylvania 19106-1574
(215) 574-6428, www.phil.frb.org

Working Papers

Research Department

WORKING PAPER NO. 00-4

ARE SCALE ECONOMIES IN BANKING ELUSIVE OR ILLUSIVE? EVIDENCE OBTAINED BY INCORPORATING CAPITAL STRUCTURE AND RISK-TAKING INTO MODELS OF BANK PRODUCTION

Joseph P. Hughes
Department of Economics, Rutgers University

Loretta J. Mester
Research Department, Federal Reserve Bank of Philadelphia
and
The Wharton School, University of Pennsylvania

Choon-Geol Moon
Department of Economics, College of Business and Economics, Hanyang University

This draft: August 2000
First Draft: December 1999

WORKING PAPER NO. 00-4

ARE SCALE ECONOMIES IN BANKING ELUSIVE OR ILLUSIVE? EVIDENCE OBTAINED BY INCORPORATING CAPITAL STRUCTURE AND RISK-TAKING INTO MODELS OF BANK PRODUCTION*

Joseph P. Hughes
Department of Economics, Rutgers University

Loretta J. Mester
Research Department, Federal Reserve Bank of Philadelphia
and
Finance Department, The Wharton School, University of Pennsylvania

Choon-Geol Moon
Department of Economics, College of Business and Economics, Hanyang University

This Draft: August 2000
First Draft: December 1999

*The authors thank William Lang for his helpful comments. The views expressed in this paper do not necessarily reflect those of the Federal Reserve Bank of Philadelphia or of the Federal Reserve System.

Correspondence to Hughes at Department of Economics, Rutgers University, New Brunswick, NJ 08901-1248; phone: 732-932-7517; email: jphughes@rci.rutgers.edu. To Mester at Research Department, Federal Reserve Bank of Philadelphia, Ten Independence Mall, Philadelphia, PA 19106-1574; phone: (215) 574-3807; fax: (215) 574-4364; email: Loretta.Mester@PHIL.frb.org. To Moon at Department of Economics, College of Business and Economics, Hanyang University, 17 Haengdang-Dong, Seongdong-Gu, Seoul 133-791, Korea; phone: 82-2-2290-1035; email: mooncg@unitel.co.kr.

JEL Codes: D20, D21, G21, L23; Key Words: banking, production, risk, scale economies

Are Scale Economies in Banking Elusive or Illusive? Evidence Obtained by Incorporating Capital Structure and Risk-Taking into Models of Bank Production*

Joseph P. Hughes
Department of Economics, Rutgers University

Loretta J. Mester
Research Department, Federal Reserve Bank of Philadelphia
and
Finance Department, The Wharton School, University of Pennsylvania

Choon-Geol Moon
Department of Economics, College of Business and Economics, Hanyang University

This Draft: May 2000
First Draft: December 1999

Abstract

This paper explores how to incorporate banks' capital structure and risk-taking into models of production. In doing so, the paper bridges the gulf between (1) the banking literature that studies moral hazard effects of bank regulation without considering the underlying microeconomics of production and (2) the literature that uses dual profit and cost functions to study the microeconomics of bank production without explicitly considering how banks' production decisions influence their riskiness.

Various production models that differ in how they account for capital structure and in the objectives they impute to bank managers—cost minimization versus value maximization—are estimated using U.S. data on highest-level bank holding companies. Modeling the bank's objective as value maximization conveniently incorporates both market-priced risk and expected cash flow into managers' ranking and choice of production plans.

Estimated scale economies are found to depend critically on how banks' capital structure and risk-taking is modeled. In particular, when equity capital, in addition to debt, is included in the production model and cost is computed from the value-maximizing expansion path rather than the cost-minimizing path, banks are found to have large scale economies that increase with size. Moreover, better diversification is associated with larger scale economies while increased risk-taking and inefficient risk-taking are associated with smaller scale economies.

Introduction

Textbooks usually claim that commercial banking enjoys scale economies that result from such phenomena as spreading the overhead and better diversification. When commercial banks merge, their managers usually cite these scale economies as a justification for the merger, and the current wave of bank mergers, which is creating extremely large banks, lends credence to their claims. However, most academic studies of bank production fail to find evidence of these scale economies. This raises a fundamental question: are scale economies in commercial banking elusive or illusive?[1]

We demonstrate that scale economies exist but they are *elusive*, and we show that they elude the standard analysis of production because it fails to account for risk and, in particular, the endogeneity of risk. If, for example, a larger scale of operations leads to better diversification that reduces liquidity risk and credit risk, it is also likely to reduce the marginal cost of risk management, *ceteris paribus*. But, other things do not necessarily remain equal. In particular, risk-taking is endogenous, and the reduced marginal cost of managing risk may give banks the incentive to engage in more risk-taking. While scale-related diversification reduces cost, *ceteris paribus*—the **diversification effect**—additional risk-taking may increase cost if banks have to spend more to manage increased risk—the **risk-taking effect**. Does the risk-taking effect mask scale economies that result from better diversification?[2] Using estimates of scale economies from a standard model, we identify a diversification effect in a second-stage regression and provide evidence that, controlling for risk-taking, better diversification alone can uncover the elusive scale economies. Thus, uncovering banks' scale economies requires incorporating risk into the analysis of production.

Risk is an essential ingredient in bank production. Banks specialize in risk assessment, risk monitoring, and risk diversification. The theory of financial intermediation hypothesizes that banks' unique capital structure gives rise to their comparative advantage in assessing and monitoring risk, which allows banks to produce a variety of information-intensive assets and financial services.[3] Banks lever their equity capital with demandable debt (demand deposits) that participates in the payments system. Access to the deposit information of their customers gives banks a cost advantage over nonbank lenders in evaluating credit and monitoring loans while it reduces the debt contract's inherent moral hazard problem of risk-shifting.[4] In addition, banks' unique capital structure occasions substantial regulation that produces well-

---

[1]For a review of the scale economies literature, see Hughes (1999a).

[2]It is possible that a bank taking on more risk engages in less credit assessment and less monitoring, in which case costs need not rise. In this case, the effect of taking on more risk would not offset the diversification effect and so not obscure potential scale economies.

[3]See Bhattacharya and Thakor (1993) for a review of this literature.

[4]Calomiris and Kahn (1991) and Flannery (1994) analyze the incentive advantage of demandable debt, and Mester, Nakamura, and Renault (1998) offer empirical evidence of the informational advantage of banks over nonbank lenders.

documented, contrasting incentives for risk-taking. On the one hand, the potential for costly episodes of financial distress, which might involve liquidity crises, regulatory intervention, and, in extreme cases, revocation of the valuable charter, gives banks an incentive to reduce risk-taking. On the other hand, mispriced safety-net protections, such as underpriced deposit insurance and discount-window borrowing, provide an incentive to increase risk-taking.[5]

While many investigations have linked banks' market value to their capital structure and to their responses to the opposing incentives for risk-taking, they have generally ignored how risk-taking is linked to production decisions. In contrast, investigations that have employed dual profit and cost functions to study production decisions and scale economies have generally ignored capital structure and endogenous risk, which has left them unable to analyze such risk-related phenomena as diversification and moral hazard and to detect the alleged scale economies that follow from better diversification.[6]

We attempt to bridge these two literatures by incorporating **endogenous risk-taking** into a model of bank production, and we ask how incorporating risk affects the model's estimates of scale economies. Incorporating capital structure is a key component of this strategy, but an equally important component generalizes the managerial objective function. There is an extensive literature on the contrasting incentives for risk-taking created by banking's unusual distress costs and by its safety net-subsidies. To study how banks' capital structure and, in general, their production decisions are influenced by these contrasting risk-taking incentives, models of banks' decision-making must be sufficiently general to account for managers' attitudes toward risk. Thus, we incorporate endogenous risk-taking into a model of production by **incorporating capital structure** and by **generalizing managerial objectives** to include value maximization as well as profit maximization. As Modigliani and Miller (1958) have noted, value maximization is a more appropriate goal to attribute to managers when production is characterized by uncertainty, because profit maximization does not account for production risk or the appropriate discount rate that is applied to the profit stream.

The rest of the paper is organized as follows. In section I, we explore how to incorporate capital structure in models of bank production. To resolve a persistent question in modeling bank production, we empirically test whether deposits are a net input or output. Then we estimate a minimum cost function that accounts for banks' mix of debt and equity. Using this minimum cost function, we compute a shadow price for equity, and using this shadow price, we compute the cost of equity capital. Most studies of bank cost functions exclude the cost of equity capital from their measure of costs. Hence, they model cash-flow

---

[5]See, for example Demsetz, Saidenberg, and Strahan (1996), Keeley (1990), Grossman (1992), Marcus (1984), and Merton (1977).

[6]See Hughes (1999a) for an extensive review of this literature.

costs rather than economic costs.[7]

In section II we compute a standard cost function that omits equity capital, and we compare its estimated cash-flow scale economies with those of the model that incorporates capital structure and the cost of capital. Both formulations yield essentially constant returns to scale.

In section III we ask if these two models' estimates of constant returns to scale reflect a risk-taking effect that masks cost economies due to better diversification. To distinguish the effects of risk-taking and diversification on scale economies, we regress banks' scale economies on variables that control for their asset size, sources of risk-taking, and their diversification. We show that better diversification is related to larger scale economies while increased risk-taking is related to smaller scale economies. Most important, we demonstrate that a proportional variation in size and diversification, controlling for sources of risk-taking, yields a statistically and economically significant increase in scale economies.

In section IV we ask whether the objective of cost minimization is sufficiently general to model the behavior of banks. We test whether the first-order conditions necessary for cost minimization are satisfied by the data. We find that many larger banks tend to underemploy equity capital relative to their cost-minimizing levels while most smaller banks tend to overemploy capital. This result is consistent with the too-big-to-fail incentive of larger banks to exploit safety-net subsidies and the incentive of smaller banks to protect their generally higher ratio of charter value to book value by taking on less risk.[8]

In section V, we model risk as an explicit component of banks' production decisions by allowing banks to be value maximizers rather than profit maximizers. That is, bank managers rank production plans not just by their expected profitability, but also by their riskiness—by higher moments of the plans' implied, subjective probability distributions of profit. Although agency problems might mean utility-maximizing managers might not choose value-maximizing plans, we show that the choices of efficient risk-takers in our sample will approximate value-maximizing choices. We then measure scale economies along the **value-maximizing expansion path** and compare this measure to those obtained from different formulations of the cost-minimizing path in previous sections. By explicitly allowing managers to choose production plans that do not necessarily maximize current expected profitability, we find the largest measured scale economies of all the formulations. Moreover, these economies increase with bank size—a result that suggests that even megamergers are exploiting scale economies. In a second-stage regression, we show that higher scale economies are associated with better diversification and lower scale economies are associated with increased risk-taking and inefficient risk-taking. In short, commercial banking's

---

[7]Studies that employ economic costs include McAllister and McManus (1993) and Clark (1996).

[8]While the FDIC Improvement Act of 1991 makes it more difficult for regulators to invoke too-big-to-fail to keep an ailing bank open, it does not close off this possibility for banks that would pose a systemic risk if allowed to fail.

alleged scale economies are not elusive when production models include risk in banks' production decisions.

I. Incorporating Capital Structure

The theory of financial intermediation identifies banks' unique capital structure (levering equity capital with demandable debt that is part of the economy's payments system) as the source of their comparative advantage in producing information-intensive loans and financial services. As noted above, commercial banks' comparative advantage involves both an informational advantage and an incentive advantage over nonbank lenders. As a source of loanable funds, debt resembles an input in the production of loans and financial services. As a payments service, demandable debt resembles an output, although all debt involves issuance and redemption activities. In either case, debt clearly is a component of banks' technology.

In contrast to the debate over debt's status in bank technology, equity capital's status is often ignored in models of bank technology even though the risk-incentives literature gives equity capital a prominent role in banks' decision-making. Banks' equity capital serves as a source of loanable funds, as a cushion to protect banks from loan losses and financial distress, and as a credible signal to less informed outside creditors of asset quality and the resources allocated to maintaining their quality.[9] Banks that fund assets with a lower capital-to-asset ratio need more debt financing and have a higher risk of insolvency, *ceteris paribus*. Equity capital is, thus, an important component of banks' technology, too.

Incorporating debt and equity in models of bank technology raises two important questions: are demand deposits to be modeled as an input or an output and how is the cost of equity capital to be taken into account? The first question's answer is often treated as a matter of taste, but the data can provide a theoretically reliable answer. We implement an empirical test and find that deposits' empirical influence on cost is theoretically consistent with that of an input. We formulate the second question's answer by conditioning the minimum cost on the level of equity capital and computing equity capital's shadow price from this conditional optimum.

Let bank technology be represented by the transformation function, $T(\mathbf{y}, \mathbf{x}, k) \leq 0$, where $\mathbf{y}$ denotes information-intensive loans and financial services; k, equity capital; $\mathbf{x}_d$, demandable debt and other types of debt; $\mathbf{x}_p$, labor and physical capital; and $\mathbf{x} = (\mathbf{x}_p, \mathbf{x}_d)$. Representing the price of the i-th type of input by w_i, the economic cost of producing the output vector $\mathbf{y}$ is given by $\mathbf{w}_p\mathbf{x}_p + \mathbf{w}_d\mathbf{x}_d + w_k k$; omitting the cost of equity capital, the cash-flow cost (C_{CF}) is represented by $\mathbf{w}_p\mathbf{x}_p + \mathbf{w}_d\mathbf{x}_d$.

---

[9]For a discussion of the signaling literature and how commercial banks signal their safety, see Lucas and McDonald (1992) and Hughes and Mester (1998).

A. Techniques for Conditioning Cost on the Capital Structure

The minimum **operating cost function** is defined by

(1) $$C_P(\mathbf{y}, \mathbf{w}_p, \mathbf{x}_d, k) = \min_{\mathbf{x}_p} (\mathbf{w}_p \mathbf{x}_p) \text{ s.t. } T(\mathbf{y}, \mathbf{x}, k) \leq 0, \mathbf{x}_d = \mathbf{x}_d^0, \text{ and } k = k^0.$$

The operating cost function accounts for capital structure by conditioning cost on the levels of debt and equity while excluding their expense from the cost function.

A cash-flow measure of cost includes the cost of debt but excludes the cost of equity capital. The minimum **cash-flow cost function** is defined by

(2) $$C_{CF}(\mathbf{y}, \mathbf{w}_p, \mathbf{w}_d, k) = \min_{\mathbf{x}_p, \mathbf{x}_d} (\mathbf{w}_p \mathbf{x}_p + \mathbf{w}_d \mathbf{x}_d) \text{ s.t. } T(\mathbf{y}, \mathbf{x}, k) \leq 0 \text{ and } k = k^0.$$

The level of debt minimizes cost while cost is conditioned on the level of equity capital. Hence, the level of equity capital does not have to minimize cost. This formulation accounts for capitalization but does not require a price for equity capital.

In contrast, the minimum **economic cost function** is conditioned on the price of equity capital rather than the quantity and, hence, the level of equity capital minimizes cost:

(3) $$C(\mathbf{y}, \mathbf{w}_p, \mathbf{w}_d, w_k) = \min_{\mathbf{x}_p, \mathbf{x}_d, k} (\mathbf{w}_p \mathbf{x}_p + \mathbf{w}_d \mathbf{x}_d + w_k k) \text{ s.t. } T(\mathbf{y}, \mathbf{x}, k) \leq 0.$$

While these three formulations of cost incorporate equity capital's influence on production, many bank cost studies omit any role for equity capital in defining cash-flow cost:

(4) $$C_{CF'}(\mathbf{y}, \mathbf{w}_p, \mathbf{w}_d) = \min_{\mathbf{x}_p, \mathbf{x}_d} (\mathbf{w}_p \mathbf{x}_p + \mathbf{w}_d \mathbf{x}_d) \text{ s.t. } T(\mathbf{y}, \mathbf{x}) \leq 0.$$

The differences among these four formulations of cost are important. In the cash-flow cost functions (2) and (4), (2) controls for the level of equity capital while (4) does not. The formulation in (4) is misspecified because a change in equity capital affects the cash-flow measure of cost. Consider two banks that differ only in their capital-to-asset ratio. The less capitalized bank is substituting debt for equity, and consequently, its cash-flow cost will exceed that of the more capitalized bank. By not controlling for the level of capitalization, the cost function (4) makes the less capitalized bank's production appear more costly. Of course, the level of capitalization also affects risk and, hence, the resources required to manage risk and the required return on debt.

B. Are Demand Deposits Inputs or Outputs?

Many studies of bank cost functions classify demand deposits as an "output" without testing whether they are in fact an output. The intuition is that deposits involve transactions "services" that are costly to produce. Of course, all leveraged firms incur debt issuance and redemption costs. Only commercial banks rely heavily on *demandable* debt and are willing to assume its added costs of frequent

issuance and redemption. When these studies treat deposits as an output, they include the quantity of deposits in the cost function. But there is a complication: as debt, deposits involve a direct expense—the interest paid to depositors. Hence, some of these studies include the interest expense in the measure of costs and include the interest rate on deposits in the cost function. Thus, *deposits enter the cost function as a quantity (output) and as a price (an expense) while the measure of costs includes the interest expense of deposits.* If the quantity as well as the price of demand deposits is included in the cost function, the cost function is defined by

$$(5) \qquad C_{CF'}(\mathbf{y}, \mathbf{w}_p, w_d, x_d) = [\min_{\mathbf{x}_p} (\mathbf{w}_p \mathbf{x}_p)] + w_d^0 x_d^0 \text{ s.t. } T(\mathbf{y}, \mathbf{x}) \leq 0, w_d = w_d^0, \text{ and } x_d = x_d^0,$$

but the solution, $\mathbf{x}_p(\mathbf{y}, \mathbf{w}_p, x_d)$, is not influenced by the price of deposits, w_d, since the quantity of deposits is fixed. Hence, this cost function is equivalent to the sum of the minimum operating cost function (1) and the fixed costs of deposits, $w_d^0 x_d^0$.

The question of whether deposits are to be modeled as an output, an input, or as both an output and an input is not a matter of taste. It is a technological question that can be answered by testing whether the data are consistent with the different technological roles of outputs and inputs. Since the cost functions (2) - (4) are conditioned on the *prices of deposits* and, hence, imply that the levels of deposits minimize cost, they implicitly classify deposits as inputs. On the other hand, since the operating cost function (1) is conditioned on the *levels of deposits*, it is consistent with either role for deposits. In fact, it affords an empirical test of the status of deposits. This test asks how an increase in the level of deposits affects the variable cost, $\mathbf{w}_p \mathbf{x}_p$, of producing the output levels, $\mathbf{y}$. If deposits are outputs, then more variable inputs and, hence, variable expenditure will be required to produce $\mathbf{y}$ and the increased $\mathbf{x}_d$, which implies that $\partial C_P / \partial \mathbf{x}_d > \mathbf{0}$. If deposits are inputs, an increase in their level allows a reduction in the expenditure on variable inputs needed to produce $\mathbf{y}$, which implies that $\partial C_P / \partial \mathbf{x}_d < \mathbf{0}$.[10]

We implement this test by estimating a modified version of the operating cost function (1) using 1994 data on the highest-level bank holding companies (BHCs) in the United States. These are holding companies that are not owned by other companies. The sources of data and the definitions and constructions of the variables are discussed in Appendix 1. We amend the operating cost function described in (1) to account for **asset quality**, since asset quality influences risk and the cost of managing it. We use two proxies: an *ex ante* measure, the average contractual interest rate on loans, p, which, given the risk-free interest rate, r, captures an average risk-premium, and an *ex post* measure, the amount of nonperforming loans, n. Hence, the transformation function amended to account for asset quality becomes $T(\mathbf{y}, n, p, \mathbf{x}, k) \leq 0$. In addition, we allow one type of debt, other borrowed funds, designated by $x_{d'}$, to be a

[10]See Hughes and Mester (1993) for the first application of this test to bank costs.

variable input while we control for the levels of the other two types of debt, insured and uninsured deposits, designated by $\boldsymbol{x_{d''}}$. The resulting operating cost function includes the price $w_{d'}$ and the quantities $\boldsymbol{x_{d''}}$ and k:

$$(6) \qquad C_P(\boldsymbol{y}, n, p, \boldsymbol{w}_p, w_{d'}, \boldsymbol{x_{d''}}, k) = \min_{\boldsymbol{x_p}, x_{d'}} (\boldsymbol{w_p x_p} + w_{d'} x_{d'}) \text{ s.t. } T(\boldsymbol{y}, n, p, \boldsymbol{x}, k) \leq 0, \boldsymbol{x_{d''}} = \boldsymbol{x^0_{d''}}, \text{ and } k = k^0.$$

We estimate (6) using the translog specification, $\ln C = \alpha_0 + \sum_i \alpha_i \ln z_i + (\frac{1}{2})\sum_i \sum_j \alpha_{ij} \ln z_i \ln z_j$, where $\boldsymbol{z} = (\boldsymbol{y}, n, p, \boldsymbol{w}_p, w_{d'}, \boldsymbol{x_{d''}}, k)$ and its associated share equations. We impose the usual restrictions. The test for input or output status is conducted on both insured and uninsured deposits.

Table 1 reports the results of these tests for five subsamples, which divide the sample by asset size.[11] The mean derivatives ($\partial C_p / \partial \boldsymbol{x_{d''}}$) with respect to uninsured and insured deposits are significantly negative with the exception of the largest group's insured deposits. If the influence of three outliers is removed, this group's mean is also negative. Thus, **the data strongly imply that deposits function as inputs in production**. In the analysis that follows, we shall model them as inputs. Having identified the role of deposits in production as that of inputs, we specify the cost function in terms of the operating cost function (1), which includes the levels but not the prices of deposits, or in terms of the cash-flow function (2), which includes the prices of deposits but not the levels. Appendix 2 explains why trying to include the price of deposits, representing deposits' role as an input, and the level of deposits, proxying for the transactions services output, can lead to biased estimates of scale economies.

C. The Shadow Price of Equity Capital

Having determined that deposits behave as inputs in our sample, we estimate the cash-flow cost function (2), which conditions cost on the prices of deposits rather than their levels. This assumes the levels of insured and uninsured deposits minimize cost. As in the estimation of the operating cost function, we include controls for asset quality:

$$(7) \qquad C_{CF}(\boldsymbol{y}, n, p, \boldsymbol{w_p}, \boldsymbol{w_d}, k) = \min_{\boldsymbol{x_p}, \boldsymbol{x_d}} (\boldsymbol{w}_p \boldsymbol{x}_p + \boldsymbol{w_d x_d}) \text{ s.t. } T(\boldsymbol{y}, n, p, \boldsymbol{x}, k) \leq 0 \text{ and } k = k^0.$$

This cash-flow cost function excludes the cost of equity capital but accounts for its level. If its level minimizes economic cost at the market price of capital, w_k, it solves the minimization problem that defines economic cost,

$$(8) \qquad C(\boldsymbol{y}, n, p, \boldsymbol{w_p}, \boldsymbol{w_d}, w_k) = \min_k C_{CF}(\boldsymbol{y}, n, p, \boldsymbol{w_p}, \boldsymbol{w_d}, k) + w_k k,$$

and, hence, satisfies the first-order condition,

---

[11]Note that the reported mean derivatives are computed as the mean of the derivatives calculated at each observation rather than the derivative evaluated at the mean of the data.

$$w_k = -\frac{\partial C_{CF}}{\partial k}. \tag{9}$$

Thus, $-\partial C_{CF}/\partial k$ gives the shadow price of equity capital; it equals the market price when the level of equity capital minimizes cost.

Table 2 reports the estimated mean shadow price for equity capital ($-\partial C_{CF}/\partial k$) for each of the five size groups and shows that it increases with asset size. If the level of equity capital is cost-minimizing—that is, the shadow price equals the market price—the positive relationship between asset size and the estimated shadow price suggests that larger banks have higher levels of market-priced risk, which increase their required return on equity.

But there are good reasons to believe that a bank's capitalization does not minimize cost, and the range of estimated shadow prices seems to confirm these reasons. In particular, the low shadow prices for smaller banks relative to plausible market prices imply that their shadow prices are less than their market prices, while the relatively high shadow prices for larger banks suggest that their shadow prices exceed their market prices. By the criterion of cost minimization, **smaller banks appear to overutilize capital while larger banks seem to underutilize it**. This pattern is consistent with smaller banks protecting their charter values by holding extra capital and larger banks taking extra risk to exploit safety-net subsidies. We shall return to this issue in section IV. In the next section, we use the shadow price of capital to compute economic cost economies.

II. Calculating Scale Economies from Minimum Cost Functions

The standard analysis that omits any role for equity capital in bank production typically finds constant returns to scale when measuring cost economies. Using the same definitions of inputs and outputs as above (described in Appendix 1), we estimate the cash-flow cost function (4), which omits equity capital. As shown in Table 3, we also find essentially constant returns to scale. We measure scale economies by the inverse cost elasticity of output,

$$scale\ economies = \frac{1}{\sum_i \frac{\partial \ln C}{\partial \ln y_i}}, \tag{10}$$

so that *scale economies* > 1 implies increasing returns to scale. The average scale economies for the full sample is slightly greater than 1 but not statistically different from 1 for large banks.

We turn next to the alternative formulation of cash-flow cost (7) that accounts for the level of equity capital. Including equity capital raises the question of whether scale economies are to be measured from cash-flow cost or from economic cost. Very few studies use economic cost, since it is difficult to

obtain a measure of the cost of capital.[12] We calculate scale economies from the economic cost function by using the shadow price as a substitute for the market price of equity capital; scale economies are calculated at each observation's observed level of equity capital.[13] Since the shadow price, $-\partial C_{CF}/\partial k$, may in fact differ from the market price, w_k, we call the price we obtain from the cost derivative a *pseudo price*, w_k*. Thus, $w_k^* = -\partial C_{CF}/\partial k$, the derivative of cash-flow cost (i.e., "short-run" cost) with respect to the conditioning level of capital, k. Note that the observed level of k minimizes economic ("long-run") cost at w_k* (see (8-9)):

(11) $$C(\mathbf{y}, n, p, \mathbf{w}_p, w_d, w_k^*) = C_{CF}(\mathbf{y}, n, p, \mathbf{w}_p, w_d, k) + w_k^* k.$$

Hence, the measure of scale economies from the economic-cost function is

(12) $$economic\text{-}cost\ economies = \frac{1}{\left[\frac{\partial C(\mathbf{y}, n, p, \mathbf{w_p}, w_d, w_k^*)}{\partial \mathbf{y}}\right] \cdot \left[\frac{\mathbf{y}}{C(\mathbf{y}, n, p, \mathbf{w_p}, w_d, w_k^*)}\right]}.$$

To measure these scale economies, we need measures of total economic cost and marginal economic cost. These can be obtained from the conditional cash-flow cost function. First, since the level of equity capital, k, minimizes economic cost, the marginal cash-flow ("short-run") cost equals the marginal economic ("long-run") cost:[14]

(13) $$\frac{\partial C(\mathbf{y}, n, p, \mathbf{w_p}, w_d, w_k^*)}{\partial \mathbf{y}} = \frac{\partial C(\mathbf{y}, n, p, \mathbf{w_p}, w_d, k)}{\partial \mathbf{y}}.$$

Second, total economic cost, which is given by (11), is obtained by adding the shadow cost of equity, $(-\partial C_{CF}/\partial k)\cdot k$, to the cash-flow cost. Thus, substituting (11) and (13) into (12), we obtain a measure of economic-cost economies in terms of the cash-flow cost function:

---

[12]McAllister and McManus (1993) arbitrarily pick a required return, which they assume is identical across all banks. Clark (1996) uses the Capital-Asset-Pricing Model to determine a market-based, required return on equity, w_k.

[13]This procedure is adapted from Braeutigam and Daughety (1983), who show how to compute long-run scale economies from a short-run (variable) cost function. The application of their technique to measure the cost of capital was proposed by Hughes (1999a).

[14]See Braeutigam and Daughety (1983) for a proof of this well-known proposition.

$$(14)\quad economic\text{-}cost\ economies = \frac{1}{\left[\frac{\partial C_{CF}(\boldsymbol{y}, n, p, \boldsymbol{w_p}, w_d, k)}{\partial \boldsymbol{y}}\right] \cdot \left[\frac{\boldsymbol{y}}{C_{CF}(\boldsymbol{y}, n, p, \boldsymbol{w_p}, w_d, k) + \left(-\frac{\partial C_{CF}}{\partial k}\right) k}\right]}$$

$$= \frac{1 - \frac{\partial \ln C_{CF}}{\partial \ln k}}{\sum_i \frac{\partial \ln C_{CF}}{\partial \ln y_i}}.$$

Table 4 reports our estimates of economic-cost economies using (14) and the shadow prices of equity capital derived from the estimated cash-flow cost function (7). For the sample average and for all but the largest BHCs, production is characterized by slightly decreasing returns to scale. For the largest BHCs, returns to scale are not significantly different from 1. In general, these results are quite similar to the cash-flow cost economies measured from the cost function that excludes equity capital and asset quality. They suggest that simply accounting for asset quality and capital structure in the cost function is not a sufficient control to identify a diversification effect and any resulting scale economies. In the next section, we ask if the risk-taking effect masks scale economies and leads to these results.

III. Risk-Taking and Diversification Effects

To isolate the effects of risk-taking and diversification on scale economies, we regress the measure of economic-cost economies reported in Table 4 on variables that control for sources of risk-taking and diversification. We gauge a bank's diversification by its exposure to macroeconomic risk. Banks that operate a geographically diverse network of branches are more likely to reduce their exposure to macroeconomic risk than banks operating in only one state or region. To construct a proxy for BHC diversification, we begin by computing a variance-covariance matrix, **V**, of state unemployment rates over the period 1985-94. The macroeconomic risk a BHC faces is proxied by the standard deviation of its weighted-average unemployment rate in the states in which it operated in 1994, where the BHC's deposit shares in each state in 1994, ***s***, serve as the weights. The inverse of this measure, $1/[\boldsymbol{s}'\mathbf{V}\boldsymbol{s}]^{1/2}$, is our measure of macroeconomic diversification. A reduction in the weighted variance of unemployment rates increases our measure of diversification.[15]

[15]This measure was used in Hughes, Lang, Mester, and Moon (1999) to study the benefits of bank consolidation. They find that it is an important variable explaining how consolidation can improve banks' efficiency and market value.

We also control for a bank's total assets and the asset growth rate from 1993 to 1994. To control for sources of risk, we use the capital-to-asset ratio, the loan-to-asset ratio, measures of *ex ante* asset quality (the average contractual interest rate on loans) and *ex post* quality (the ratio of nonperforming loans to total assets). In addition, we control for the average cost of other borrowed money (uninsured funds).

Using GMM, we regress the measure of economic-cost economies on these variables characterizing sources of risk and diversification.[16] The results, reported in Table 5, show that

(i) controlling for size and sources of risk, an increase in diversification is associated with larger scale economies;

(ii) controlling for diversification and sources of risk, an increase in asset size is associated with larger scale economies;

(iii) controlling for sources of risk, a 1 percent increase in diversification and asset size is associated with a statistically significant increase of 0.01084 in scale economies;[17] at the mean level of scale economies this represents a 1.1 percent increase in scale economies; and

(iv) the effect of an increase in diversification and asset size on scale economies appears to be economically significant, too, since controlling for sources of risk, an increase from the minimum levels of diversification and asset-size in the sample (0.43110 and $32 million) to the maximum levels in the sample (2.0957 and $249 billion) would yield a (statistically significant) increase of 0.03307 in scale economies.

Risk-taking effects on scale economies are surprisingly varied: (i) the statistically significant, positive coefficients on the average contractual return on assets and the ratio of nonperforming loans to total assets indicate that an increase in risk due to a reduction in asset quality is associated with larger scale economies (possibly reflecting BHCs devoting few resources to risk management when they choose to make riskier loans.); (ii) the statistically significant, negative coefficient on the average interest rate on uninsured funds suggests that the positive effect of a decrease in asset quality might be partially offset if the increase in risk also caused an increase in the interest rate on uninsured funds; (iii) the significant, negative coefficient on the loan-to-asset ratio indicates that an increase in risk-taking that takes the form of substituting loans for securities and liquid assets is associated with lower scale economies; and (iv) the large, significantly positive coefficient on the capital-to-asset ratio implies that an increase in risk due to a lower capital ratio is associated with lower scale economies.

In summary, this evidence demonstrates that banks' diversification and risk-taking have statistically and economically significant effects on their scale economies. Better diversification is

---

[16]Some variables are entered in logs to make it easier to compute the effect of a proportionate increase in the variable on scale economies.

[17]This is calculated by summing the coefficients on the log of diversification and log of assets.

associated with larger scale economies while substituting loans for securities and liquid assets and increasing the leverage ratio are associated with lower scale economies. This evidence points to the need to incorporate risk into the analysis of production if the elusive scale economies are to be uncovered.

IV. Is Cost Minimization Consistent with the Data?

The theory of financial intermediation focuses on banks' comparative advantage in assessing, monitoring, and taking risk. Risk, then, is an important factor in bank managers' consideration of potential production plans. Production plans' profitability must then be evaluated not just by their expected profitability, but also by higher moments of their implied conditional probability distributions of profit. If managers simply rank production plans by their first moments, they choose the plan that has the highest expected profit, which implies that cost is minimized for the resulting output vector. But if risk is also an important consideration in production decisions, managers' rankings of production plans must account for higher moments. Hence, they may trade expected profitability for lower risk to increase the *discounted value of profit* and to lower the probability of costly financial distress. When risk influences production decisions, managers may choose more costly but less risky production plans to produce any given output vector.

The risk-incentives literature in banking emphasizes two contrasting incentives that motivate risk-taking. On the one hand, safety-net subsidies, such as mispriced deposit insurance, give banks the incentive to increase risk to exploit the put-option value of the insurance. On the other hand, the potential for costly episodes of financial distress entailing liquidity crises, regulatory intervention in a bank's operations, and even forfeiture of the valuable charter gives banks the incentive to reduce risk. Since the value-maximizing production plan must account for this risk trade-off, it may not necessarily maximize current expected profit. Value maximization, then, is a more general objective than profit maximization because it ranks production plans not just by the first moment of their implied subjective conditional distributions of profit, but also by higher moments that characterize risk.

A. Testing the Assumption of Cost Minimization

Most studies of bank technology do not test their assumption of cost minimization or profit maximization. However, there are several notable exceptions. For example, English, Grosskopf, Hayes, and Yaisawarng (1993) find that an important source of bank inefficiency results from overutilization of resources. Evanoff (1998) and Evanoff, Israilevich, and Merris (1990) estimate bank technology by using a shadow cost function and find that the shadow prices at which cost is minimized are not equal to market prices. They interpret their data's failure to satisfy the conditions for cost minimization at market prices as

evidence of regulatory distortions. Mester (1989) tests for evidence of expense-preference behavior, a particular form of overutilization of resources, in savings and loan associations. Mester (1991) tests for agency problems leading to deviations from cost-minimizing behavior in savings and loans.

Using the estimated model of operating costs defined in (6), we test our data to see if BHCs' production decisions are consistent with cost minimizing behavior. In particular, we focus on the capital structure and ask if it minimizes cost. The minimum operating cost function in (6) is conditioned on the levels of insured and uninsured deposits and equity. Then, the minimum economic cost function is the sum of minimum operating cost and the cost of equity and deposits when their levels are optimal. Thus, the cost-minimizing capital structure solves

(15) $$\min_{\mathbf{x}_{d''},k} C_P(\mathbf{y}, n, p, \mathbf{w}_p, \mathbf{w}_{d'}, \mathbf{x}_{d''}, k) + \mathbf{w}_{d''}\mathbf{x}_{d''} + w_k k \text{ s.t. } T(\mathbf{y}, n, p, \mathbf{x}, k) \le 0.$$

The first-order conditions of this minimization problem are

(15a) $$\frac{\partial C_p}{\partial k} + w_k = 0 \quad \text{and} \quad \frac{\partial C_p}{\partial \mathbf{x}_{d''}} + \mathbf{w}_{d''} = \mathbf{0}.$$

When these conditions fail to hold, say for capital, they imply the following:

(15b) $$\frac{\partial C_p}{\partial k} + w_k > 0 \Rightarrow \text{overutilization of capital,}$$

(15c) $$\frac{\partial C_p}{\partial k} + w_k < 0 \Rightarrow \text{underutilization of capital.}$$

Using the estimated operating cost function, we test whether the first-order conditions (15a) hold for each BHC in our sample. In other words, do BHCs use the cost-minimizing capital and deposit structure? For the uninsured and insured deposits tests, $\mathbf{w}_{d''}$ is proxied by the BHC's average interest rate paid to each of these two types of deposits (see Appendix 1). For the financial capital test, since we do not have a price for equity in our data and since many of the BHCs in our sample are not publicly traded, it is difficult to obtain market prices for them. So we evaluate optimality for a range of prices between w_k = 0.14 and w_k = 0.18, which are chosen as a plausible range of market return for banks' equity. That is, we ask if the level of equity capital minimizes cost at a market return on capital between 0.14 and 0.18. (See footnote 19 for more about this range.)

As shown in Table 6, our data overwhelmingly reject the hypothesis of cost minimization. The nature of the violations of the first-order conditions depends on the size of the bank. In the range of asset sizes up to $10 billion, most banks **overutilize** both types of deposits and equity capital (relative to all

other types of borrowed funds) while many banks above $10 billion **underutilize** deposits and capital.[18,19]

B. Implications of the Violation of Cost Minimization

Overutilization of capital implies that the capital-to-asset ratio is too large to minimize cost—although a larger ratio reduces the risk of insolvency and protects charter value. Underutilization implies that it is too small—although a smaller ratio increases risk and the value of safety-net subsidies. If this difference between larger and smaller banks reflects different incentives to take risk, it represents a value-enhancing capital allocation even though it fails to minimize cost. If risk matters in production decisions, why model scale economies along the cost-minimizing expansion path?

Of course, any production plan that is **technically efficient** can be made to minimize cost at the appropriate shadow prices, and scale economies can be measured along the **shadow-cost-minimizing expansion path**. In section II, we adopted this strategy to measure economic-cost economies from the conditional cash-flow cost function: we assumed that each bank's observed level of capital minimizes shadow cost, which does not equal cost measured at market prices. But this approach has two fundamental problems.

First, the measure of scale economies derived from shadow prices relies on prices that are not observed to compute a cost that is not incurred. Although this measure of scale economies characterizes the cost-minimizing expansion path, it does not seem necessarily useful in explaining the behavior of banks' observed cash-flow costs as they expand.

Second, it requires production decisions to be technically efficient, which means that in the production of any given output vector, banks cannot use extra resources to reduce risk: they must necessarily use the minimum resources required to produce the given output vector without consideration of risk. For example, consider two banks that produce the same portfolio of loans and financial services. One uses extra labor to assess and monitor credit risk and, hence, to reduce risk. The less risky bank is technically inefficient, although it is also less risky.

The shadow-price technique of gauging technology is useful as long as risk does not matter. When risk does not influence production decisions, production plans are ranked by the first moment of their implied subjective, conditional probability distributions of profit. Profit and cost can be defined in terms of any set of shadow prices so that any technically efficient production plan can be made to minimize cost

---

[18]We also conduct the test for the optimality of capital with the estimated cash-flow cost function (7) and obtain the same conclusion: most smaller banks overutilize capital while many larger banks underutilize it.

[19]As shown in Table 2, the mean shadow price is rather low for BHCs with less than $10 billion in assets and rather high for those with more than $10 billion. This spread in shadow prices implies that the pattern in over- and underutilization of capital will hold for a much wider range of market prices than 0.14 to 0.18.

at some set of shadow prices. **But when risk influences the ranking of production plans, higher moments of the distributions matter in the rankings so that plans that are not technically efficient may nevertheless maximize value**—expected profit discounted by the risk-adjusted rate of interest less any expected costs of financial distress. **And there is no set of shadow prices at which a technically inefficient production plan can be made to minimize shadow cost or maximize shadow profit.** This is the most compelling reason that the assumptions of profit maximization and cost minimization are inadequate for the task of modeling risky production.

Incorporating risk into the analysis of production requires not a substitution of shadow prices for market prices when the data fail to satisfy the first-order conditions for cost minimization (and profit maximization), but a substitution of a different managerial objective function that is consistent with the data at observed market prices. **If bank production is modeled so that observed production decisions represent an equilibrium at market prices, then the expansion path and the associated measure of scale economies that are derived from this model could incorporate value-maximizing production decisions that incur extra cost for reduced risk.** Thus, when risk influences production decisions, scale economies should be measured along the value-maximizing expansion path rather than the cost-minimizing path.

In the next section, we present a more general model of production that allows higher moments of probability distributions of profit to influence managers' rankings of production plans. In the absence of agency problems between managers and outsider-owners, such rankings can be assumed to reflect the owners' objective of value maximization. In the presence of agency problems, such rankings may capture managers' private concerns for perquisites and risk—perhaps risk avoidance to protect their relatively undiversified investment of human capital. As an empirical matter, we can expect agency problems. Consequently, we include a technique that distinguishes efficient managers from inefficient ones to identify the efficient expansion path. We then link the efficiency criterion to value maximization—the distinction between efficient and inefficient firms can be used to bound the value-maximizing expansion path by the efficient, utility-maximizing path.

V. Scale Economies Along the Value-Maximizing Expansion Path

The value-maximizing expansion path differs from the cost-minimizing expansion path in that it accounts for the market-priced risk of production decisions. A bank's production plan, $(\boldsymbol{y}, n, p, \boldsymbol{x}, k)$, influences its market value of equity, *MVE*, through its effect on the expected cash flow, *E(CFE)*, and the

market-priced risk of the cash flow, which determines the discount rate, w_k, applied to it.[20]

A production plan's risk influences not just the discount rate on cash flow (the required return on equity), but also the expected cost of financial distress, which in the case of commercial banks can involve liquidity crises, regulatory intervention, and even revocation of the valuable charter. A plan that maximizes profit (and minimizes cost) does not necessarily maximize market value, since profit maximization does not take into account the plan's effect on the discount rate, the required return on debt, and the expected distress costs.[21]

A. Incorporating Risk into Managers' Rankings of Production Plans

If managers' decision-making is modeled assuming that managers maximize profits (minimize cost), their ranking of production plans depends only on the first moment of the plans' subjective conditional probability distributions of profit. If, instead, their decision-making is modeled assuming that managers maximize value, their ranking of production plans must account for higher moments that characterize the plans' riskiness. To allow higher moments to influence rankings, we represent managers' rankings with a managerial utility function defined over production plans.[22]

Hughes and Moon (1995) show that this representation of the managerial utility function is a generalization of the utility function defined over expected profit and profit risk or expected return and return risk—the first two moments of the profit distribution. It is sufficiently general to incorporate profit maximization where only the first moment influences the ranking and value maximization where higher moments also affect the ranking. To rank production plans, managers must translate plans into subjective, conditional probability distributions of profit. Their beliefs about the probability distribution of states of the world, s_t, and about how plans interact with states to yield a realization of profit, $\pi_t = g(\mathbf{y}_t, n_t, p_t, \mathbf{x}_t, k_t,$

---

[20]Hence, the market value of a bank's equity is given by the expectation conditional on information at time 1, $MVE_1 = \Sigma_{t=1}^{\infty} E\{[CFE_t(\mathbf{y_t}, n_t, p_t, \mathbf{x_t}, k_t)]/\prod_{s=1}^{\infty}[1+w_{k,s}(\mathbf{y}_s, n_s, p_s, x_s, k_s; \theta_s)]\}$. The expected cash flow takes into account solvent as well as insolvent states of the world and, consequently, includes payments to factors of production and to stakeholders and payments to third parties during episodes of financial distress. Ignoring depreciation, after-tax net cash flow is designated by $\pi_t = (1-\tau)(\boldsymbol{p}_t \boldsymbol{y}_t - \boldsymbol{w}_{p,t}\boldsymbol{x}_{p,t} - \boldsymbol{w}_{d,t}\boldsymbol{x}_{d,t})$, where τ is the tax rate on profit. Letting $p_\pi = 1/(1-\tau)$ be the price of a dollar of after-tax cash flow in terms of before-tax dollars, the before-tax cash flow is given by $\Pi_t = p_\pi \pi_t = \boldsymbol{p}\boldsymbol{y}_t - \boldsymbol{w}_{p,t}\boldsymbol{x}_{p,t} - \boldsymbol{w}_{d,t}\boldsymbol{x}_{d,t}$. The expected cash flow consists of the expected after-tax profit less the expected costs of financial distress, CFD_t: $E(CFE_t) = E(\pi_t - CFD_t)$. The analysis of how a bank's production plan influences its market value is detailed in Hughes, Lang, Moon, and Pagano (1997) and in Hughes, Lang, Mester, and Moon (1999).

[21]See Hughes (1999a) and (1999b) for a detailed discussion of this point. The most preferred production system, described in sections V.A, B, and C, is analyzed in more detail in Hughes, Lang, Mester, and Moon (1995, 1996, and 1999).

[22]This technique was first proposed by Hughes (1989) to model cost functions for hospitals and for education (1990). It was developed in its current form for commercial banks by Hughes, Lang, Mester, and Moon (1995, 1996, 1999) and by Hughes and Moon (1995).

s_t), imply a subjective distribution of profit that is conditional on the production plan: $f(\pi_t; \mathbf{y}_t, n_t, p_t, \mathbf{x}_t, k_t)$. Under certain restrictive conditions, this distribution can be represented by its first two moments, $E(\pi_t; \mathbf{y}_t, n_t, p_t, \mathbf{x}_t, k_t)$ and $S(\pi_t; \mathbf{y}_t, n_t, p_t, \mathbf{x}_t, k_t)$. Rather than define a utility function over these two moments, we define it over profit and the production plan, $U(\pi_t; \mathbf{y}_t, n_t, p_t, \mathbf{x}_t, k_t)$, which is equivalent to defining it over the conditional probability distributions $f(\cdot)$. Thus, this generalized utility function can represent value maximization as well as profit maximization. Of course, it can also represent rankings of production plans that reflect agency problems—say, entrenched managers that avoid risk because their firm-based wealth is not well diversified or less skilled managers who try to disguise their substandard performance by taking excessive risk to earn a better expected return.[23]

We assume the path utility-maximizing BHCs take when they grow will correspond to the market-value maximizing path of efficient BHCs in the sample. Efficiency is determined relative to an expected market value versus risk frontier. This is equivalent to assuming capital market discipline is effective at these firms. Said differently, for the most efficient banks in our sample, the highest ranked plans—their managers' most preferred production plans—approximate the value-maximizing expansion path.

B. Managers' Most Preferred Production Plan[24]

The managers' most preferred production plan maximizes utility, subject to the definition of profit ((16b) below) and to the technology ((16c) below). We include the risk-free rate of interest, r, in the utility function and m, noninterest income in the definition of profit. The sum of interest income, $\mathbf{py}$, and noninterest income, m, gives total revenue. We condition the utility maximization problem on the output vector to facilitate the calculation of scale economies and on the level of equity capital so that we can normalize profit by capital to obtain a rate of return on equity. Hence, the most preferred production plan is the solution to the problem

$$\max_{\pi, \mathbf{x}} U(\pi, \mathbf{x}; \mathbf{y}, n, p, r, k) \tag{16a}$$

$$\text{s.t. } p_\pi \pi = \mathbf{p}\,\mathbf{y} + m - \mathbf{w}_p \mathbf{x}_p - \mathbf{w}_d \mathbf{x}_d \tag{16b}$$

$$T(\mathbf{y}, n, p, \mathbf{x}, k) \leq 0. \tag{16c}$$

The solution gives the most preferred profit function, $\pi^* = \pi(\mathbf{y}, n, \mathbf{v}, m, k)$, and the most preferred input demand functions, $\mathbf{x}^* = \mathbf{x}(\mathbf{y}, n, \mathbf{v}, m, k)$, where $\mathbf{v} = (\mathbf{w}, p, r, p_\pi)$, π is after-tax net cash flow, $p_\pi = 1/(1-\tau)$ is the price of a dollar of after-tax cash flow in terms of before-tax dollars, and τ is the tax rate on profit.

---

[23]See Gorton and Rosen (1995).

[24]Much of the description of this model has appeared elsewhere (Hughes, Lang, Mester, and Moon, 1995, 1996, 1999); it is included here as an aid to the reader.

Note that the profit function is not necessarily the maximum profit function. The most preferred profit function and input demand functions define the highest ranked production plan or, equivalently, the highest-ranked subjective conditional probability distribution of profit. They reflect managers' assessment of how the moments of this distribution influence market value and other goals that may arise from agency problems.

Although the historical risk resulting from past production decisions is readily observable, the *ex ante* risk that motivates current production decisions is not. But it is *indirectly* observable in these demand functions because they reveal the *rankings* of production plans. We turn next to the question of how these rankings can be recovered from production data and what can be inferred from them about risk.

C. Almost Ideal Demand (Production) System

Just as the estimation of consumers' utility-maximizing demands recovers their preferences for goods and services from their budget data, the estimation of the most preferred profit and input demand functions recovers managers' preferences for production plans or, equivalently, for subjective probability distributions conditional on the production plan from their production data. We adapt the expenditure function of the Almost Ideal Demand System (Deaton and Muellbauer, 1980) to represent generalized managerial preferences and use it to derive the functional forms for the utility-maximizing demands for profit and the production plan. These profit and input demand functions are expressed as shares of total revenue, $\boldsymbol{py} + m$, and sum to one. In addition to the share equations, we estimate the first order condition for the optimal level of equity capital, k, which is a conditioning argument in the share equations. (See Appendix 4 and HLMM 1996 for details of the derivation.) Thus, the model to be estimated is:

(17a) $$\frac{p_\pi \pi}{\mathbf{p}\mathbf{y} + m} = \frac{\partial ln\, \mathbf{P}}{\partial ln\, p_\pi} + \mu\,[ln(\mathbf{p}\mathbf{y} + m) - ln\, \mathbf{P}]$$

(17b) $$\frac{w_i x_i}{\mathbf{p}\mathbf{y} + m} = \frac{\partial ln\, \mathbf{P}}{\partial ln\, w_i} + \nu_i[ln(\mathbf{p}\mathbf{y} + m) - ln\, \mathbf{P}] \quad \forall\, i$$

(17c) $$\frac{\partial V(\cdot)}{\partial k} = \frac{\partial V(\cdot)}{\partial \ln k}\frac{\partial \ln k}{\partial k} = 0,$$

where $ln\, \mathbf{P} = \alpha_0 + \sum_i \alpha_i\, ln\, z_i + (\tfrac{1}{2})\sum_i \sum_j \alpha_{ij}\, ln\, z_i\, ln\, z_j$, $\boldsymbol{z} = (\boldsymbol{y}, n, \boldsymbol{v}, k)$, and the indirect utility function, $V(\cdot)$, of the maximization problem (16a)-(16c) is

(17d) $$V(\cdot) = \frac{ln(\boldsymbol{p}\boldsymbol{y} + m) - ln\, P}{\beta_0\left(\prod_i y_i^{\beta_i}\right)\left(\prod_j w_i^{\nu_j}\right) p_\pi^{\mu} k^{\kappa}} \quad .$$

The details of the derivation of these equations are found in Appendix 4 and in Hughes, Lang Mester, and Moon (hereafter denoted HLMM) (1996).

Adding-up, homogeneity, and symmetry were imposed (see HLMM (1995, 1996) for details of these restrictions). Because preferences of managers represent their beliefs about the probabilities of future states of the world and how those states interact with production plans to generate realizations of profit, we expect managers' preferences to change over time. To deal with this problem, cross-sectional data were used in HLMM (1996) (the same data used to estimate the previous cost functions) to estimate the production system with nonlinear two-stage least squares, which is a generalized method of moments.

HLMM (1995) show that when the comparative-static restrictions implied by the assumption of profit maximization are imposed on the AID Production System, it becomes identically equal to the standard translog profit (cost) function and share equations.[25] These restrictions allow a test for the consistency of the data with profit maximization (cost minimization).

D. Measuring Scale Economies from the Most-Preferred Cost Function

Since the utility-maximizing demand for profit is conditioned on the output vector, the managers' most preferred cost function can be readily computed from it:

(18) $$C_{MP}(\boldsymbol{y}, n, \boldsymbol{v}, m, k) = \boldsymbol{p}\,\boldsymbol{y} + m - p_\pi\pi(\boldsymbol{y}, n, \boldsymbol{v}, m, k).$$

Scale economies computed from the most preferred cost function are defined by

(19) $$most\text{-}preferred\ cost\ economies = \frac{C_{MP}}{\sum_i y_i\left[\frac{\partial C_{MP}}{\partial y_i} + \frac{\partial C_{MP}}{\partial k}\frac{\partial k}{\partial y_i}\right]} = \frac{\boldsymbol{p}\cdot\boldsymbol{y} + m - p_\pi\pi}{\sum_i y_i\left[p_i - \frac{\partial p_\pi\pi}{\partial y_i} - \frac{\partial p_\pi\pi}{\partial k}\frac{\partial k}{\partial y_i}\right]},$$

where $\partial k/\partial y_i$ is computed from the first-order condition (17c).

Scale economies measured by the most-preferred cost function describe the elasticity of cost along the **utility-maximizing expansion path**. This path is a generalization of the **cost-minimizing path** and can accommodate **value-maximizing production decisions**. Hence, it accounts for managers' assessment of how their production decisions affect the bank's exposure to market-priced risk. The theoretical relationship between most-preferred cost economies and cost economies measured from the minimum cost

[25]The AID System does not require the translog form. It can nest any flexible functional form.

function is discussed in Appendix 3. In particular, we show that just adding a control for *ex post* risk into the traditional cost function cannot adequately capture the idea that managers might use resources to manage risk.

Because the AID System is a flexible functional form, the estimate of scale economies in (19) varies from BHC to BHC in the sample. Table 7 reports the mean of these estimates over the entire sample and for BHCs in five asset-size groups.[26]

E. Scale Economies, Diversification, and Risk-Taking

The mean measure of scale economies for the full sample is 1.14. For the smallest banks in the sample the mean is 1.12; for the largest, 1.25. Scale economies that are greater than one and that increase with size represent a disequilibrium in U. S. banking that is the result of historical restrictions on both intrastate and interstate branching. As states began relaxing these restrictions in the 1980s, a wave of mergers ensued. Increasing scale economies are consistent with this wave, which has witnessed mergers among the largest banks that have created banks of unprecedented size where, again, scale economies are cited as one of the benefits of consolidation. The magnitude of these scale economies and their relationship to size have been confirmed in other applications of the Most Preferred Production System (MPPS) and by a few other studies.[27]

Following our procedure in section III, we next regress most-preferred cost economies on variables characterizing risk-taking and diversification; results are shown in Table 8. We again find that scale

---

[26]The estimates differ slightly from those reported in HLMM (1996) because the samples differ slightly.

[27]DeYoung, Hughes, and Moon (1998) estimated the MPPS for U.S. national banks in 1994 and, although they did not report the measures of scale economies, found an average measure of 1.12 that increased from 1.08 for banks with less than $300 million in assets to 1.21 for banks whose assets exceeded $50 billion. HLMM (1995) use the MPPS to estimate scale economies for 286 U. S. banks that exceed $1 billion in assets. Using a cross section of 1990 data, they obtained an average of 1.15 for the full sample. Scale economies increased from 1.10 for the smallest quartile to 1.21 for the largest quartile. When they imposed the parameter restrictions implied by cost minimization, the measure of scale economies fell to the usual range: 1.02 for the smallest quartile to 1.05 for the largest. This contrast between scale economies measured along the utility-maximizing expansion path and along the cost-minimizing path suggests that the behavioral assumption used to recover technological relationships is very important.

The few other studies that have found evidence of scale economies in banking have important differences with the models that have found no such evidence. Berger and Mester (1997) include the level of equity capital and measures of output quality and historical risk in their model of profit maximization, which is estimated as a frontier. The optimal scale for banks in each of six asset-size groups is found to be two to three times larger than the average size bank in the size group—a result that suggests no size group represents a long-run equilibrium. Clark (1996) measures scale economies from economic cost and accounts for risk by using the Capital Asset Pricing Model to compute the required return on equity capital. He obtains evidence of scale economies from his model when he uses frontier estimation techniques. Hughes and Mester (1998) include a measure of historical risk and account for the level of equity capital in a model of cost minimization, but demand for capital maximizes utility and does not necessarily minimize cost. Hughes (1999a) reviews these studies in greater detail.

economies increase with size when we control for sources of risk-taking. Although better diversification increases scale economies, the effect is not statistically significant. However, controlling for sources of risk-taking, a proportional increase in both diversification and assets is associated with higher scale economies, and the effect, 0.03559, is statistically significant and three times larger than the effect of the same variation, 0.01084, measured by the minimum cost function (see Table 5). An increase from the minimum levels of diversification and assets in the sample to the maximum levels, *ceteris paribus*, implies an increase of 0.2015 in most-preferred cost economies. As expected, increased risk-taking is associated with lower scale economies when we control for size: an increase in the average contractual return on assets, an increase in the loans-to-assets ratio, and a decrease in the capital-to-assets ratio are associated with lower scale economies.

F. Defining Efficiency to Approximate the Value-Maximizing Expansion Path

As we discussed above, we assume that the utility-maximizing expansion path for the efficient BHCs in the sample will correspond to a value-maximizing expansion path. Efficient BHCs are those that achieve their highest potential market values.

To implement this concept of efficiency using the most preferred production plan, we must link the estimated most preferred production plan to the moments of its implied distribution of profit and show that these moments are relevant to observed market values. These steps are described in detail in HLMM (1999) and in Hughes (1999a, 1999b). We briefly outline them here.

To measure efficiency, we first derive the expected return on equity (ROE) from the estimated most preferred production plan model. The expected ROE is readily obtained from the estimated profit share equation (21a), since it is conditioned on equity capital:

(20) $$E(p_\pi \pi/k) = [s_\pi(\hat{\beta})][(\boldsymbol{p}\boldsymbol{y} + m)/k],$$

where $s_\pi(\hat{\beta})$ is the estimated profit share. Next we need to measure ROE risk, which is less straightforward. We use the standard error of the predicted ROE—an econometric measure of prediction risk, which is a function of the exogenous variables of the production system, ($\boldsymbol{y}$, n, $\boldsymbol{v}$, m, k). So the ROE's prediction risk depends on production decisions and the economic environment—the mix of outputs $\boldsymbol{y}$, asset quality n, the price and interest rate environment $\boldsymbol{v}$, off-balance activity m, and the level of equity capital k. The standard error of the predicted profit share is defined by

(21) $$S(p_\pi \pi/(\boldsymbol{p}\boldsymbol{y} + m)) = [\hat{\boldsymbol{X}}' \hat{var}(\hat{\beta})\, \hat{\boldsymbol{X}}]^{1/2},$$

where $[\hat{\boldsymbol{X}}' \hat{var}(\hat{\beta})\, \hat{\boldsymbol{X}}]$ is the estimated asymptotic variance of the predicted share, $s_\pi(\hat{\beta})$, and

$\hat{\boldsymbol{X}}=\partial s_{\pi}(\hat{\beta})/\partial\beta$. Then the prediction risk of ROE is given by[28]

(22) $$S(p_{\pi}\pi/k) = [\hat{\boldsymbol{X}}' \, v\hat{a}r(\hat{\beta}) \, \hat{\boldsymbol{X}}]^{1/2} \; [(\boldsymbol{p}\boldsymbol{y} + m)/k].$$

Neither the production-based measure of expected return nor the econometric measure of prediction risk is standard. Nevertheless, for our purpose of identifying BHCs that are relatively successful in achieving their highest potential market value, we need only show that these measures can adequately explain market value. Since market value is the discounted expected cash flow, the production-based expected profit must account not only for the current-period expected profit, but also must proxy future expected profit. In addition, profit risk must be a good proxy for the market-priced risk that establishes the discount rate. The expected cash flow includes expected distress costs as well as expected profit. It seems less likely that the production model captures potential distress costs. We turn to market-value data to assess the adequacy of the production-based measures in explaining value.

Using the 190 BHCs that are publicly traded, we regress the *ln*(end-of-year 1994 market value) of these banks on their *ln*(expected profit, $E_i(p_{\pi}\pi)$), and *ln*(profit risk, $S_i(p_{\pi}\pi)$).[29] The resulting estimated equation is as follows:

(23) $$ln(\text{market value of equity}_i) = \underset{(0.379)}{-0.784} + \underset{(0.088)}{1.617} \, ln\, E_i(p_{\pi}\pi) - \underset{(0.086)}{0.672} \, ln\, S_i(p_{\pi}\pi).$$

(standard errors in parentheses)

The coefficients on expected return and return risk have the theoretically correct signs and are significant at the 1 percent level. The adjusted R-squared is 0.96, so our production-based measures of expected return and return risk substantially explain market value. We can then use these to derive market return efficiency measures based on a stochastic frontier (Jondrow, Lovell, Materov, and Schmidt (1982) developed the stochastic frontier methodology for measuring efficiency).

In particular, allowing the frontier to be nonlinear, we fit return as a quadratic function of risk:

(24) $$E_i(p_{\pi}\pi/k) = \Gamma_0 + \Gamma_1 S_i(p_{\pi}\pi/k) + \Gamma_2[S_i(p_{\pi}\pi/k)]^2 + \epsilon_i,$$

where the error term, $\epsilon_i = v_i - u_i$, is composed of a two-sided component, v_i, distributed $N(0, \sigma_v^2)$, which accounts for unmeasured randomness in the data generation process, and a one-sided component, $u_i > 0$, distributed half normally, $N(0,\sigma_u^2)$, which gauges inefficiency. A BHC's **return inefficiency** is measured by the conditional expectation of u_i given ϵ_i. It represents the difference between a BHC's expected ROE and the frontier value of ROE, for given level of risk, adjusted for noise.

---

[28]Note, we use nonlinear two-stage least squares, a general method of moments, to estimate the share equations so that we do not impose homoscedasticity on the error terms. Thus, the resulting variance-covariance matrix of parameter estimates, $v\hat{a}r(\hat{\beta})$, reflects non-constant error variance across banks and captures how the production plan and economic environment of each bank influence the prediction risk attached to its expected return.

[29]This result is also reported in HLMM (1999).

To interpret this measure, note that if two BHCs have the same level of risk, S^0, one is more return efficient than the other if it has a higher expected return , i.e., its risk-return combination is closer to the efficient frontier. It also means that the more efficient BHC has a higher market-to-book value of equity.[30]

We will assume that the production decisions of the most efficient banks, those whose risk-return combinations are closest to the frontier, approximate value-maximizing decisions and compute mean scale economies for the most return-efficient banks. There is one caveat: when we measure efficiency, the peer group of any BHC is defined by risk, i.e., we compare BHCs with the same level of risk. Hence, the loss of market value due to a suboptimal choice of risk is not taken into account. For a variety of reasons such as differing charter values, there will be no unique optimal level of risk for all banks. It might seem plausible that BHCs that achieve the highest expected returns at any given level of risk also choose risk efficiently. But as a robustness check, we compute mean scale economies for efficient subsamples defined by two alternative measures of efficiency derived directly from market values. These two measures include inefficiency due to suboptimal levels of risk, and we can compare the consistency of these two market-value-based measures of efficiency with the production-based measure.

The market-value-based measures of efficiency are those computed in HLMM (1999), which followed the methods in Hughes, Lang, Moon, and Pagano (1997). Using stochastic frontier estimation techniques, we fit the following two frontiers:

$$\text{Market Value of Assets}_i = \Psi_0 + \Psi_1 \text{Adjusted Book Value of Assets}_i + \Psi_2(\text{Adjusted Book Value of Assets}_i)^2 + \xi_i^A, \tag{25}$$

$$\text{Market Value of Equity}_i = \Phi_0 + \Phi_1 \text{Adjusted Book Value of Equity}_i + \Phi_2(\text{Adjusted Book Value of Equity}_i)^2 + \xi_i^E, \tag{26}$$

where the adjusted book value of equity is the book value of equity minus goodwill and the adjusted book value of assets is the book value of assets minus goodwill; $\xi_i^E \equiv \nu_i^E - u_i^E$ and $\xi_i^A \equiv \nu_i^A - u_i^A$ are composite error terms, with ν_i^E and ν_i^A normally distributed with zero means, and u_i^E and u_i^A positive and half-normally distributed. The **market-value asset inefficiency** of a BHC is measured by the conditional mean of u_i^A

[30]To see this, consider two BHCs with the same level of risk. Then by (24), the BHC that is more return-efficient has higher expected ROE, that is, $E(p_\pi\pi/k) = E(p_\pi\pi)/k$. By estimated equation (23), the estimated numerator of expected ROE, namely, expected profit, $E_i(p_\pi\pi)$, is a function of the market value of equity and profit risk, $S_i(p_\pi\pi)$. And the denominator of expected ROE is the book value of equity, k. Thus, when we compare two banks with the same level of risk, we can conclude that the more return-efficient bank also has a higher market value-to-book value. Linking the production-based measures of expected return and return risk to market value allows us to infer that the frontier defines the highest potential market-to-book ratio for any given level of risk.

given ξ_i^A, and represents the amount by which the BHC could increase the market value of its assets if it were as well positioned in the marketplace and as efficient as the best-practice BHCs. The relevant peer group is BHCs with the same book-value of assets. This measure would include inefficiency due to suboptimal risk decisions, but it excludes inefficiency due to a suboptimal size.[31] Similarly, the **market-value equity inefficiency** of a BHC is measured by the conditional mean of u_i^E given ξ_i^E. The relevant peer group is BHCs with the same book-value of equity. So a more efficient BHC has a higher market-to-book value of equity than its less efficient peer.

We now use our three different measures of efficiency to determine the efficient set of BHCs. These efficient BHCs are assumed to maximize market value when they maximize their utility. The three efficiency measures are the production-based measure of **return (ROE) inefficiency—**the lost potential return at any *given risk level*; the **market-value measure of asset inefficiency**—the lost potential market value of assets at any given investment in assets; and the **market-value measure of equity inefficiency**—the lost potential market value of equity at any given investment in equity. As noted above, the value-maximizing expansion path controls for the investment in assets and asks, what is the highest potential market value—either of assets or of equity?[32,33]

G. Measuring Scale Economies along the Value-Maximizing Expansion Path

Using each of our three measures of market-value efficiency, we identify the most efficient quarter (and least efficient quarter) of BHCs in each of the five size groups and calculate the mean measure of scale economies for each group. As shown in Table 9, in all size categories but the largest, the mean measure of scale economies for the most efficient banks is larger than the mean for the least efficient banks

---

[31]Nevertheless, this is ideal because a value-maximizing expansion path concerns which plans maximize expected market value *at any given size*; it does not determine what the optimal size is. This is similar to the simple scale economies measure, which says how minimum cost changes as output level changes, but does not determine the optimal level of output.

[32]Agency problems such as asset substitution imply that the asset-based measure is preferable to the equity-based one.

[33]One could use the fitted relationship in (32) to convert each bank's expected profit, $E_i(p_\pi\pi)$, and profit risk, $S_i(p_\pi\pi)$, into its expected market value and, dividing by book value, into its expected ratio of market-to-book value. Then efficiency comparisons could be made *directly* from this ratio, which eliminates the need to fit a frontier. But this procedure has the important drawback of assuming that the marginal effect of expected risk on market value is constant across all banks. If this were true, an optimal risk level would be given by the tangency of linear iso-market value lines to the risk-return frontier in risk-return space. But banks' market opportunities (growth prospects, market power, etc.), their charter values, and their distress costs differ, so their optimal risk levels will also differ. The fitted relationship in (32) is not intended to capture these effects, i.e., it does not determine the tradeoff between expected return and risk from the production model.

and larger than the mean of all banks in the group.[34]

Comparing mean scale economies for efficient banks by size groups does not control for risk. To investigate the effects of differences in efficiency on the measure of scale economies while also controlling for sources of risk-taking, we repeat the regression reported in Table 8, but we add the production-based measure of return inefficiency to the list of independent variables. The results are reported in Table 10. Comparing the results in Table 8 with those in Table 10, it is clear that controlling for efficiency differences among banks sharpens the precision of the estimation.

Qualitatively, the results remain the same. Controlling for return inefficiency, sources of risk-taking, and macroeconomic diversification, an increase in size is associated with an increase in scale economies. Similarly, an improvement in macroeconomic diversification is also associated with an increase in scale economies (although it is not statistically significant at the 10% level). A proportional increase in size and diversification, *ceteris paribus*, is associated with higher scale economies, and the magnitude is large and similar to that obtained without controlling for efficiency. We also find that, controlling for the level of macroeconomic diversification, asset size, and sources of risk-taking, more efficient banks operate with higher levels of scale economies.

VI. Conclusion

Are scale economies in banking elusive or illusive? We have offered evidence obtained by incorporating capital structure and risk-taking into models of bank production that strongly suggests that the scale economies so often cited by merging banks do, indeed, exist, but are elusive. They are influenced by banks' risk-taking and can, in fact, be obscured by risk-taking. Hence, to uncover evidence of these elusive scale economies requires incorporating capital structure and risk-taking into the analysis of production.

[34]Clark (1996) found a similar pattern: estimating the economic cost function (3) as an average relationship gave significantly smaller measures of scale economies than estimating it as a frontier.

Bibliography

Berger, Allen N., and Loretta J. Mester, 1997, "Inside the Black Box: What Explains Differences in the Efficiencies of Financial Institutions?" *Journal of Banking and Finance*, 21, 895-947.

Braeutigam, Ronald R., and Andrew F. Daughety, 1983, "On the Estimation of Returns to Scale Using Variable Cost Functions," *Economic Letters*, 11, 25-31.

Bhattacharya, Sudipto, and Anjan Thakor, 1993, "Contemporary Banking Theory," *Journal of Financial Intermediation* 3, 2-50.

Calomiris, Charles W., and Charles M. Kahn, 1991, "The Role of Demandable Debt in Structuring Optimal Banking Arrangements," *American Economic Review*, 81, 497-513.

Clark, Jeffrey A., 1996, "Economic Cost, Scale Efficiency, and Competitive Viability in Banking," *Journal of Money, Credit, and Banking*, 28, 342-364.

Deaton, Angus, and John Muellbauer, 1980, "An Almost Ideal Demand System," *American Economic Review*, 70, 312-326.

Demsetz, Rebecca S., Marc R. Saidenberg, and Philip E. Strahan, 1996, "Banks with Something to Lose: The Disciplinary Role of Franchise Value," Federal Reserve Bank of New York *Economic Policy Review*, 2, 1-14.

DeYoung, Robert, Joseph P. Hughes, and Choon-Geol Moon, 1998, "Regulatory Covenant Enforcement and the Efficiency of Risk-Taking at U.S. Commercial Banks," Economics Working Paper 98-1, Office of the Comptroller of the Currency, Washington, D.C. (forthcoming in the *Journal of Economics and Business*).

English, M., S. Grosskopf, K. Hayes, and S. Yaisawarng, 1993, "Output Allocative and Technical Efficiency of Banks," *Journal of Banking and Finance*, 17, 349-366.

Evanoff, Douglas D., 1998, "Assessing the Impact of Regulation on Bank Cost Efficiency," *Economic Perspectives*, Federal Reserve Bank of Chicago, 22, 21-32.

Evanoff, Douglas D., Philip R. Israilevich, and Randall C. Merris, 1990, "Relative Price Efficiency, Technical Change, and Scale Economies for Large Commercial Banks," *Journal of Regulatory Economics*, 2, 281-298.

Flannery, Mark J., 1994, "Debt Maturity and the Deadweight Cost of Leverage: Optimally Financing Banking Firms," *American Economic Review*, 84, 320-331.

Gorton, Gary, and Richard Rosen, 1995, "Corporate Control, Portfolio Choice, and the Decline of Banking," *Journal of Finance*, 50, 1377-1420.

Grossman, Richard S., 1992, "Deposit Insurance, Regulation, and Moral Hazard in the Thrift Industry: Evidence from the 1930's," *American Economic Review*, 82, 800-821.

Hughes, Joseph P., 1989, "Hospital Cost Functions: The Case Where Revenues Affect Production," Department of Economics, Rutgers University.

Hughes, Joseph P., 1990, "The Theory and Estimation of Revenue-Driven Costs: The Case of Higher Education," Department of Economics, Rutgers University.

Hughes, Joseph P., 1999a, "Incorporating Risk into the Analysis of Production," *Atlantic Economic Journal*, 27:1, 1-23.

Hughes, Joseph P., 1999b, "Measuring Efficiency When Competitive Prices Aggregate Differences in Product Quality and Risk," Proceedings of the Conference on the Microeconomics of Financial Intermediation, University of Venice, *Research in Economics/Ricerche Economiche*, 53, 47-76.

Hughes, Joseph P., William Lang, Loretta J. Mester, and Choon-Geol Moon, 1995, "Recovering Technologies that Account for Generalized Managerial Preferences: An Application to Non-Risk-Neutral Banks," Wharton Financial Institutions Center, Working Paper 95-16 and Federal Reserve Bank of Philadelphia Working Paper No. 95-8/R.

Hughes, Joseph P., William Lang, Loretta J. Mester, and Choon-Geol Moon, 1996, "Efficient Banking Under Interstate Branching," *Journal of Money, Credit, and Banking*, 28, 1045-1071.

Hughes, Joseph P., William Lang, Loretta Mester, and Choon-Geol Moon, 1999, "The Dollars and Sense of Bank Consolidation," *Journal of Banking and Finance*, 23, 291-324.

Hughes, Joseph P., William Lang, Choon-Geol Moon, and Michael Pagano, 1997 (revised 1999), "Measuring the Efficiency of Capital Allocation in Commercial Banking," Federal Reserve Bank of Philadelphia, Working Paper 98-2.

Hughes, Joseph P., and Loretta J. Mester, 1993, "A Quality and Risk-Adjusted Cost Function for Banks: Evidence on the 'Too-Big-to-Fail' Doctrine," *Journal of Productivity Analysis*, 4, 292-315.

Hughes, Joseph P., and Loretta J. Mester, 1998, "Bank Capitalization and Cost: Evidence of Scale Economies in Risk Management and Signaling," *The Review of Economics and Statistics*, 80, 314-325.

Hughes, Joseph P. and Choon-Geol Moon, 1995 (revised 1997), "Measuring Bank Efficiency When Managers Trade Return for Reduced Risk," Working Paper, Department of Economics, Rutgers University.

Humphrey, D.B., and L.B. Pulley, 1997, "Banks' Responses to Deregulation: Profits, Technology, and Efficiency," *Journal of Money, Credit, and Banking*, 29, 73-93.

Jondrow, J., C.A.K. Lovell, I.S. Materov, and P. Schmidt, 1982, "On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model," *Journal of Econometrics*, 19, 233-238.

Keeley, Michael C., 1990, "Deposit Insurance, Risk, and Market Power in Banking," *American Economic Review*, 80, 1183-1200.

Lucas, Deborah, and Robert J. McDonald, 1992, "Bank Financing and Investment Decisions with Asymmetric Information about Loan Quality," *Rand Journal of Economics*, 23, 86-105.

McAllister, Patrick H., and Douglas McManus, 1993, "Resolving the Scale Efficiency Puzzle in Banking," *Journal of Banking and Finance*, 17, 389-405.

Marcus, Alan J., 1984, "Deregulation and Bank Financial Policy," *Journal of Banking and Finance*, 8, 557-565.

Merton, Robert C., 1977, "An Analytic Derivation of the Cost of Deposit Insurance Loan Guarantees," *Journal of Banking and Finance* 1, 3-11.

Mester, Loretta J., 1989, "Testing for Expense Preference Behavior: Mutual Versus Stock Savings and Loans," *The RAND Journal of Economics*, 20, 483-498.

Mester, Loretta J., 1991, "Agency Costs Among Savings and Loans," *Journal of Financial Intermediation*, 1,1991, 257-278.

Mester, Loretta J., Leonard I. Nakamura, and Micheline Renault, 1998, "Checking Accounts and Bank Monitoring," Federal Reserve Bank of Philadelphia, Working Paper 98-25.

Modigliani, F., and M.H. Miller, 1958, "The Cost of Capital, Corporation Finance, and the Theory of Investment," *American Economic Review* 48, 261-297.

Table 1

Derivative of Operating (Variable) Cost with Respect to Uninsured Deposits

Total Assets	No. of BHCs	Mean	Std. Dev.	T-Stat.	Prob.
≤ $300 million	109	−0.00108 *	0.00034	3.17605	0.00149
$300 million - 2 billion	215	−0.00389 *	0.00115	3.38380	0.00071
$2 billion - 10 billion	67	−0.03976 *	0.00890	4.46491	0.00001
$10 billion - 50 billion	35	−0.57449 *	0.21573	2.66293	0.00775
> $50 billion	15	−19.84262 **,†	10.82535	1.83298	0.06681
Omitting 3 outliers	12	−3.74348 *	1.30563	2.86718	0.00414

Derivative of Operating (Variable) Cost with Respect to Insured Deposits

Total Assets	Mean	Std Dev.	T-Stat.	Prob.
≤ $300 million	−0.00069 *	0.00010	6.85969	6.90092×10^{-12}
$300 million - 2 billion	−0.00302 *	0.00037	8.11544	4.44089×10^{-16}
$2 billion - 10 billion	−0.02168 *	0.00266	8.15639	4.44089×10^{-16}
$10 billion - 50 billion	−0.14229 *	0.03718	3.82757	0.00013
> $50 billion	2.71844 **,†	1.54489	1.75963	0.07847
Omitting 3 outliers	−0.45571	0.31970	1.42544	0.15403

The means are calculated as the mean of the derivatives calculated at each observation rather than the derivative evaluated at the mean of the data.

† Three outliers have distorted this mean. There are no outliers in other groups.

* Significantly different from zero at the 1 percent level.
** Significantly different from zero at the 10 percent level.

Number of observations: 441 total sample; 109 in the subsample ≤ $300 million; 215 in the subsample $300 million - 2 billion; 67 in the subsample $2 billion - 10 billion; 35 in the subsample $10 billion - 50 billion; and 15 in the subsample > $50 billion.

Table 2

Shadow Price of Equity Capital

$$-\frac{\partial C_{CF}(\boldsymbol{y}, n, p, \boldsymbol{w_p}, \boldsymbol{w_d}, k)}{\partial k}$$

Total Assets	Mean	Std. Dev.	T-Stat.	Prob.
full sample	0.14932 *	0.04055	3.68223	0.00012
≤ $300 million	0.00238 *	0.00058	4.09301	0.00002
$300 million - 2 billion	0.00788 *	0.00172	4.58840	2.23330×10^{-6}
$2 billion - 10 billion	0.04953 *	0.01358	3.64634	0.00013
$10 billion - 50 billion	0.34746 *	0.08322	4.17536	0.00001
> $50 billion	3.22794 *,†	0.96955	3.32933	0.00044
Omitting 3 outliers	1.77690 *	0.47452	3.74460	0.00018

The means are calculated as the mean of the derivatives calculated at each observation rather than the derivative evaluated at the mean of the data.

† Three outliers have distorted this mean.

* Significantly different from zero at the 1 percent level.

Number of observations: 441 total sample; 109 in the subsample ≤ $300 million; 215 in the subsample $300 million - 2 billion; 67 in the subsample $2 billion - 10 billion; 35 in the subsample $10 billion - 50 billion; and 15 in the subsample > $50 billion.

Table 3

Estimated Cash-Flow Scale Economies from the Cash-Flow Cost Function that Omits Equity Capital

$$C_{CF'}(\boldsymbol{y}, \boldsymbol{w_p}, \boldsymbol{w_d}) = \min_{\boldsymbol{x_p}, \boldsymbol{x_d}} (\boldsymbol{w_p x_p} + \boldsymbol{w_d x_d}) \text{ s.t. } T(\boldsymbol{y}, \boldsymbol{x}) \leq 0.$$

$$cash\text{-}flow\ \ scale\ \ economies\ = \frac{1}{\sum_i \frac{\partial \ln C_{CF'}}{\partial \ln y_i}}$$

Total Assets	Mean	Std. Dev.	T-Stat. (≠0)	T-Stat. (≠1)
full sample	1.01158 *,†	0.00522	193.90549	2.22006
≤ $300 million	1.01190 *	0.00791	127.92744	1.50421
$300 million - 2 billion	1.01238 *,†	0.00575	176.01769	2.15231
$2 billion - 10 billion	1.01137 *,‡	0.00583	173.54541	1.95111
$10 billion - 50 billion	1.00769 *	0.00818	123.21237	0.93982
> $50 billion	1.00789 *	0.01066	94.56706	0.73996

The means are calculated as the mean estimate of scale economies calculated at each observation rather than the scale economies evaluated at the mean of the data.

* Significantly different from zero at the 1 percent level.

† Significantly different from one at the 5 percent level.

‡ Significantly different from one at the 10 percent level.

Table 4

Estimated Economic-Cost Scale Economies from the Cash-Flow Cost Function Conditioned on the Level of Equity Capital

$$C_{CF}(\mathbf{y}, n, p, \mathbf{w}_p, \mathbf{w}_d, k) = \min_{\mathbf{x}_p, \mathbf{x}_d} (\mathbf{w}_p \mathbf{x}_p + \mathbf{w}_d \mathbf{x}_d) \text{ s.t. } T(\mathbf{y}, n, p, \mathbf{x}, k) \leq 0 \text{ and } k = k^0$$

$$economic\text{-}cost\ economies = \frac{1 - \frac{\partial \ln C_{CF}}{\partial \ln k}}{\sum_i \frac{\partial \ln C_{CF}}{\partial \ln y_i}}$$

Total Assets	Mean	Std. Dev.	T-Stat. (≠0)	T-Stat. (≠1)
full sample	0.98155 *,†	0.00755	130.08281	2.44449
≤ $300 million	0.97749 *,†	0.00871	112.19073	2.58349
$300 million - 2 billion	0.98364 *,†	0.00773	127.19727	2.11499
$2 billion - 10 billion	0.98197 *,‡	0.00946	103.77513	1.90546
$10 billion - 50 billion	0.97817 *,‡	0.01208	80.95614	1.80704
> $50 billion	0.98719 *	0.01545	63.90000	0.82905

The means are calculated as the mean estimate of scale economies calculated at each observation rather than the scale economies evaluated at the mean of the data.

* Significantly different from zero at the 1 percent level.

† Significantly different from one at the 5 percent level.

‡ Significantly different from one at the 10 percent level.

Table 5

Isolating the Effects of Risk-Taking and Diversification on Scale Economies

Dependent Variable: Economic-Cost Economies

Independent Variable	Parameter Estimate	Standard Error	T-Statistic
constant	1.20341 *	0.02754	43.69
***ln*(index of diversification)**	**0.00866** *	**0.00193**	**4.49**
***ln*(total assets)**	**0.00218** *	**0.00045**	**4.83**
ln(average contractual asset return)	0.13348 *	0.01024	13.03
ln(average uninsured funds interest rate)	−0.03037 *	0.00187	−16.24
nonperforming loans/total assets	0.27135 *	0.06788	4.00
loans/total assets	−0.05866 *	0.00779	−7.53
equity/total assets	0.38953 *	0.03637	10.71
asset growth rate	0.01065 **	0.00562	1.89

Estimated using GMM. Standard errors are computed from a heteroscedasticity-consistent covariance matrix estimate (Robust-White).
Number of observations = 441
Adjusted R-squared = 0.568432

* Significantly different from zero at the 1 percent level.
**Significantly different from zero at the 10 percent level.

• change in scale economies due to a proportional increase in diversification and assets, *ceteris paribus*

$$= \frac{\partial \text{ scale economies}}{\partial \ln(\text{diversification index})} + \frac{\partial \text{ scale economies}}{\partial \ln(\text{assets})}$$

= 0.00886 + 0.00218
= 0.01084*
with standard error = 0.001985 and t-statistic = 5.462

• change in scale economies due to a increase in diversification and assets from their minimum levels to their maximum levels in the sample (i.e., from 0.4311 to 2.06569 and from $32 million to $249 billion, respectively), *ceteris paribus*

= 0.00866×[*ln*(2.06569) − *ln*(0.4311)] + 0.00218×[*ln*(249) − *ln*(0.032)]
= 0.03307*
with standard error = 0.005043 and t-statistic = 6.557

Table 6
Tests of First-Order Conditions for the Cost-Minimizing Level of Equity Capital

optimal: $\frac{\partial C_p}{\partial k} + w_k = 0 \Rightarrow$ cost-minimizing level of capital

over: $\frac{\partial C_p}{\partial k} + w_k > 0 \Rightarrow$ overutilization of capital

under: $\frac{\partial C_p}{\partial k} + w_k < 0 \Rightarrow$ underutilization of capital

The values in the table below give the percentage of observations where the capital level is optimal, over, or underutilized at $w_k \in [0.14, 0.18]$ at significance levels 1% and 10%. The "over" column reports the proportion of observations where the hypothesis that $\partial C_p/\partial k + 0.14 \leq 0$ is rejected at the 1% (10%) level of significance. The "under" column reports the proportion where the hypothesis $\partial C_p/\partial k + 0.18 \geq 0$ is rejected at the 1% (10%) level of significance. The "optimal" column reports the proportion of observations where the hypothesis that $\partial C_p/\partial k + 0.14 \leq 0$ and $\partial C_p/\partial k + 0.18 \geq 0$ cannot be rejected at the 1% (10%) level of significance.

	1%			10%		
Total Assets	**optimal**	**over**	**under**	**optimal**	**over**	**under**
full sample	12.7	84.1	2.9	8.2	85.9	5.9
≤ $300 million	0.0	100.0	0.0	0.0	100.0	0.0
$300 million - 2 billion	1.8	98.2	0.0	1.4	98.6	0.0
$2 billion - 10 billion	26.9	73.1	0.0	26.9	82.1	0.0
$10 billion - 50 billion	80.0	0.0	20.0	60.0	2.9	37.1
> $50 billion	60.0	0.0	40.0	13.3	0.0	86.7

Tests of First-Order Conditions for the Cost-Minimizing Level of Uninsured Deposits

	1%			10%		
Total Assets	**optimal**	**over**	**under**	**optimal**	**over**	**under**
full sample	18.1	78.0	3.9	9.0	81.2	9.8
≤ $300 million	0.0	100.0	0.0	0.0	100.0	0.0
$300 million - 2 billion	3.2	96.8	0.0	2.3	97.7	0.0
$2 billion - 10 billion	53.7	37.3	9.0	29.9	55.2	14.9
$10 billion - 50 billion	82.9	0.0	17.1	37.1	0.0	62.9
> $50 billion	66.7	0.0	33.3	26.7	0.0	73.3

Tests of First-Order Conditions for the Cost-Minimizing Level of Insured Deposits

	1%			10%		
Total Assets	**optimal**	**over**	**under**	**optimal**	**over**	**under**
full sample	10.0	80.7	9.3	4.5	83.4	12.0
≤ $300 million	0.0	100.0	0.0	0.0	100.0	0.0
$300 million - 2 billion	4.1	95.9	0.0	2.7	96.8	0.5
$2 billion - 10 billion	29.9	58.2	11.9	10.4	70.2	19.4
$10 billion - 50 billion	28.6	0.0	71.4	17.1	0.0	82.9
> $50 billion	46.7	0.0	53.3	20.0	13.3	66.7

Table 7

Estimated Cash-Flow Scale Economies from the Most-Preferred Production System

Total Assets	Mean	Std. Dev.	T-Stat. (≠0)	T-Stat. (≠1)
full sample	1.144532 *,†	0.010111	113.194	14.294
≤ $300 million	1.117420 *,†	0.008658	129.069	13.562
$300 million - 2 billion	1.125790 *,†	0.009038	124.562	13.918
$2 billion - 10 billion	1.171363 *,†	0.011877	98.628	14.428
$10 billion - 50 billion	1.247416 *,†	0.018151	68.725	13.631
> $50 billion	1.250270 *,†	0.017810	70.199	14.052

The means are calculated as the mean estimate of scale economies rather than the scale economies evaluated at the mean of the data.

* Significantly different from zero at the 1 percent level.
† Significantly different from one at the 1 percent level.

Table 8

Isolating the Effects of Risk-Taking and Diversification on Scale Economies Measured by the Most-Preferred Production System

Dependent Variable: Cash-Flow Scale Economies

Independent Variable	Parameter Estimate	Standard Error	T-Statistic
constant	0.01468	0.17956	0.08176
***ln*(index of diversification)**	**0.01584**	**0.01510**	**1.04929**
***ln*(total assets)**	**0.01975** *	**0.00321**	**6.16064**
ln(average contractual asset return)	−0.36906 *	0.06558	−5.62741
ln(average uninsured funds interest rate)	0.02148	0.01388	1.54761
nonperforming loans/total assets	−0.41434	0.55267	−0.74970
loans/total assets	−0.08970 **	0.05131	−1.74827
equity/total assets	−0.00033	0.24533	−0.00136
asset growth rate	−0.03877	0.03845	−1.00853

Number of observations = 441
Estimated using GMM. Standard errors are computed from a heteroscedasticity-consistent covariance matrix estimate (Robust-White).
Adjusted R-squared = 0.237

* Significantly different from zero at the 1 percent level.
**Significantly different from zero at the 10 percent level.

• change in scale economies due to a proportional increase in diversification and assets, *ceteris paribus*

$$= \frac{\partial \text{ scale economies}}{\partial \, ln(\text{diversification index})} + \frac{\partial \text{ scale economies}}{\partial \, ln(\text{assets})}$$

= 0.01584 + 0.01975
= 0.03559**
with standard error = 0.015614 and t-statistic = 2.279

• change in scale economies due to a increase in diversification and assets from their minimum levels to their maximum levels in the sample (i.e., from 0.4311 to 2.06569 and from $32 million to $249 billion, respectively), *ceteris paribus*

= 0.01584×[*ln*(2.06569)−*ln*(0.4311)] + 0.01976×[*ln*(249)−*ln*(0.032)]
= 0.20148*
with standard error = 0.037728 and t-statistic = 5.301

Table 9

Estimated Cash-Flow Scale Economies from the Most-Preferred Production System for the Value-Maximizing Expansion Path

Total Assets	**Utility-Max. Expansion Path**	**Value-Max. ROE Efficiency**	**Value-Max. MV-Asset Efficiency**	**Value-Max. MV-Equity Efficiency**
	Mean Scale Economies for All Banks in the Group	Mean Scale Economies for the 25% Most Efficient (Mean Scale Economies for the 25% Least Efficient)		
full sample	1.145	1.188 (1.130)	1.240 (1.139)	1.236 (1.110)
≤ $300 million	1.117	1.164 (1.097)	1.166 (1.079)	1.177 (1.079)
$300 million - 2 billion	1.126	1.170 (1.104)	1.106 (1.079)	1.108 (1.112)
$2 billion - 10 billion	1.171	1.224 (1.161)	1.161 (1.130)	1.160 (1.149)
$10 billion - 50 billion	1.247	1.280 (1.291)	1.242 (1.192)	1.285 (1.196)
> $50 billion	1.250	1.220 (1.344)	1.213 (1.343)	1.244 (1.343)

Table 10

Isolating the Effects of Risk-Taking and Diversification on Scale Economies Measured by the Most-Preferred Production System

Dependent Variable: Cash-Flow Scale Economies

Independent Variable	Parameter Estimate	Standard Error	T-Statistic
constant	−0.48564 *	0.16404	−2.96048
***ln*(index of diversification)**	**0.01275**	**0.01328**	**0.95996**
***ln*(total assets)**	**0.01447 ***	**0.00286**	**5.06042**
ln(average contractual asset return)	−0.59481 *	0.06106	−9.74206
ln(average uninsured funds interest rate)	0.03660 *	0.01228	2.97975
nonperforming loans/total assets	1.14266 *	0.50532	2.26125
loans/total assets	−0.21009 *	0.04638	−4.53020
equity/total assets	1.99587 *	0.27902	7.15319
asset growth rate	−0.02473	0.03384	−0.73076
ROE inefficiency	−0.78125 *	0.06923	−11.2856

Number of observations = 441
Estimated using GMM. Standard errors are computed from a heteroscedasticity-consistent covariance matrix estimate (Robust-White).
Adjusted R-squared = 0.422

* Significantly different from zero at the 1 percent level.
**Significantly different from zero at the 5 percent level.

• change in scale economies due to a proportional increase in diversification and assets, *ceteris paribus*

$$= \frac{\partial \text{ scale economies}}{\partial \, ln(\text{diversification index})} + \frac{\partial \text{ scale economies}}{\partial \, ln(\text{assets})}$$

= 0.01275 + 0.01447
= 0.02722**
with standard error = 0.013766 and t-statistic = 1.979

• change in scale economies due to a increase in diversification and assets from their minimum levels to their maximum levels in the sample (i.e., from 0.4311 to 2.06569 and from $32 million to $249 billion, respectively), *ceteris paribus*

= 0.01275×[*ln*(2.06569)−*ln*(0.4311)] + 0.01447×[*ln*(249)−*ln*(0.032)]
= 0.16137*
with standard error = 0.03398 and t-statistic = 4.403

Appendix 1

The Data

The various specifications of cost are estimated using 1994 data obtained from the Y-9C Call Reports filed quarterly by bank holding companies operating in the United States. Rather than focus on individual banks, we examine the highest level bank holding companies. These are holding companies that are not owned by other companies. They may own only one bank, but in many cases, they own multiple banks that operate in different states. We focus on the highest level holding companies, since the investment strategies of individual banks reflect the composite strategy of the top holding company and so are not necessarily independent. There are 441 companies, which range in size from $33 million to $250 billion in consolidated assets.

Of these, 190 are publicly traded. We obtain the number of shares outstanding from the Standard & Poor's Compustat database and end-of-year stock prices from the Center for Research in Securities Prices (CRSP).

The vector $\boldsymbol{y}$ consists of five outputs: liquid assets, short-term securities, long-term securities, loans and leases net of unearned income, and other assets. Labor and physical capital constitute the physical inputs, $\boldsymbol{x}_p$. Debt, $\boldsymbol{x}_d$, includes insured deposits, uninsured deposits, and other borrowed money. Insured deposits are deposits in domestic offices excluding time deposits over $100,000; uninsured deposits are domestic time deposits over $100,000; and other borrowed money comprises foreign deposits, federal funds purchased, securities sold under agreement to repurchase, other borrowed funds, subordinated debt, and mandatory convertible debt. Equity capital is measured by the sum of shareholders' equity, loan-loss reserves, and subordinated debt (Tier 1 and Tier 2 capital). With the exception of equity capital, input prices are computed by dividing the input expenditure by its quantity. *Ex post* asset quality is measured by the amount of nonperforming assets, *n*—accruing and nonaccruing loans, leases, and other assets that are past due over 90 days. *Ex ante* asset quality is gauged by the average contractual return on assets, which is the ratio of income accruing to assets divided by the quantity of accruing assets. The variable *m* is measured by noninterest income. All these variables are computed as the average of their values at the end of the four quarters of 1994.

The state tax rates are obtained from *The Book of the States*, published by the Council of State Governments, and from *Significant Aspects of Fiscal Federalism*, published by the U.S. Advisory Commission on Intergovernmental Relations.

Appendix 2

Can Demand Deposits Be Modeled Both as Inputs and Outputs?

Reasoning that deposits have the characteristics of both inputs and outputs, some studies have included both the price and the quantity of deposits in the cost function. We avoid this formulation since it adds nothing theoretically to the specification of the operating cost function (1) in the text and can yield a misleading measure of scale economies. If deposits are in fact inputs but are treated as outputs in computing scale economies, then their "marginal costs" are used to compute their cost elasticities and these elasticities are added to the other output elasticities to calculate scale economies. But the derivative of cost function (5) in the text, which includes both the prices and quantities of deposits, is not really a marginal cost of deposits: it is a **test** of whether the level of deposits minimizes cost. Taking the derivative of (5) with respect to deposits gives

$$\text{(A2.1)} \qquad \frac{\partial C_{CF}(\boldsymbol{y}, \boldsymbol{w_p}, w_d, x_d)}{\partial x_d} = \frac{\partial [\min(\boldsymbol{w_p x_p})]}{\partial x_d} + w_d$$

$$= 0 \text{ in cost-minimizing equilibrium.}$$

That is, the partial derivative must equal zero when deposits are inputs ($\partial[\min(\boldsymbol{w}_p \boldsymbol{x}_p)]/\partial x_d < 0 \Rightarrow$ input) *and* their levels minimize cash-flow cost.[35] If these studies find that the "marginal cost" of deposits is positive, $\partial[\min(\boldsymbol{w}_p \boldsymbol{x}_p)]/\partial x_d + w_d > 0$, they have really found that the first-order condition of cost minimization is violated and that deposits are overutilized. Hence, including the "marginal cost" of deposits in the scale economies calculation is not justifiable when deposits function as inputs and do not minimize cost. If scale economies are measured by the *inverse* sum of output elasticities (see (9) in the text), the measure will be biased downward when deposits are overutilized and upward when deposits are underutilized.

[35] See section III of the text for an explanation of this point.

Appendix 3

Cost-Minimizing versus Utility-Maximizing Expansion Paths

Scale economies measured by the most-preferred cost function describe the elasticity of cost along the utility-maximizing expansion path. This path is a generalization of the cost-minimizing path and can accommodate value-maximizing production decisions. Hence, it accounts for managers' assessment of how their production decisions affect the bank's exposure to market-priced risk.

The contrast between scale economies measured along the utility-maximizing expansion path and those measured along the cost-minimizing path is apparent in the utility-maximizing first-order conditions derived from (17a-c) in the text:

$$\frac{\lambda \partial T(\cdot)/\partial x_i}{\lambda \partial T(\cdot)/\partial x_j} = \frac{\mu w_i - \partial U/\partial x_i}{\mu w_j - \partial U/\partial x_j}, \tag{A3.1}$$

where λ and μ are Lagrange multipliers. When managers rank production plans by the first-moment of their implied probability distributions of profit, production plans influence utility only through their effect on profit. Thus, while $\partial U/\partial \pi > 0$, components of the production plan do not inherently affect utility so that $\partial U/\partial x_j = 0$. Moreover, in this case production must be technically efficient so that $T(\cdot) = 0$ and $\lambda > 0$. Thus, profit maximization (cost minimization) implies the familiar equality between the marginal rates of technical substitution and the input price ratios:

$$\frac{\partial T(\cdot)/\partial x_i}{\partial T(\cdot)/\partial x_j} = \frac{w_i}{w_j}. \tag{A3.2}$$

When production is technically efficient, but managers have preferences for inputs apart from their influence on profit, inputs affect utility so that $\partial U/\partial x_j > 0$. Such preferences could result from regulatory incentives, as Evanoff (1998) and Evanoff, Israilevich, and Merris (1990) have emphasized. In this case the first-order conditions take the form

$$\frac{\partial T(\cdot)/\partial x_i}{\partial T(\cdot)/\partial x_j} = \frac{\mu w_i - \partial U/\partial x_i}{\mu w_j - \partial U/\partial x_j}. \tag{A3.3}$$

Note that this is the shadow-price formulation: the shadow prices are given by $\mu w_i - \partial U/\partial x_i = w_i^*$. Thus, marginal rates of technical substitution equal ratios of shadow prices.

In contrast, when managers consider how production decisions influence risk, they may use additional resources to reduce the risk of producing any output vector. Hence, technical efficiency is no longer a meaningful requirement since risk matters. Thus, the transformation function becomes an inequality $T(\cdot) < 0$ so

that the technology constraint is no longer binding, which implies that the associated Lagrange multiplier equals zero: $\lambda = 0$. In this case the utility-maximizing first-order conditions become

$$\frac{\partial U/\partial x_i}{\partial U/\partial x_j} = \frac{w_i}{w_j}, \tag{A3.4}$$

which require the marginal rates of substitution in *"consumption"* to equal the input price ratios. This case highlights the important difference between the assumptions of profit maximization and utility maximization and their respective expansion paths. Not even the shadow price technique, whose first-order conditions are given by (A3.3), can capture the essential feature of risky production, given by (A3.4), namely, that any given output vector may require extra resources to manage risk.

Note that the idea that managers might use resources to manage risk cannot be implemented by merely adding a control for *ex post* risk into the traditional cost function. The *ex ante* level of risk that managers associate with any given production plan is not directly observable, so the cost function would have to include a measure of historical risk. But *ex ante* risk depends on expectation of *future* states of the world and how those future states interact with production plans to generate realizations of profit. And since expectations change, so do *ex ante* risk assessments. Although managers' assessments of *ex ante* risk cannot be directly measured and included in the cost function, their *rankings* of production plans, which reflect their *subjective* risk assessments, can be inferred from their choices which, by assumption, maximize utility. Only this assumption is sufficiently general to account for profit-maximizing rankings as well as rankings that are influenced by risk.

Appendix 4

Derivation of the Almost Ideal Demand (Production) System

The estimation of the most preferred profit and input demand functions recovers managers' preferences for production plans or, equivalently, for subjective probability distributions conditional on the production plan from their production data. We use the Almost Ideal Demand System (Deaton and Muellbauer, 1980) to estimate these equations.

Just as the cost function can be used to represent technology, the Almost Ideal Demand (AID) System represents preferences with the expenditure function, which gives the minimum expenditure required to achieve a given level of utility. Hence, it solves the problem,

(A4.1) $$\min_{\pi,\, \boldsymbol{x}} \; p_{\pi}\pi + \boldsymbol{w}\boldsymbol{x}$$

(A4.2) $$\text{s.t. } U^{0} - U(\pi, \boldsymbol{x}; \boldsymbol{y}, n, p, r, k) = 0$$

(A4.3) $$T(\boldsymbol{y}, n, p, \boldsymbol{x}, k) \leq 0.$$

Its solution yields the constant-utility (expenditure-minimizing) demand functions for inputs, $\boldsymbol{x}^{u} = \boldsymbol{x}^{u}(\boldsymbol{y}, n, \boldsymbol{v}, k, U^{0})$, and for profit, $\pi^{u} = \pi^{u}(\boldsymbol{y}, n, \boldsymbol{v}, k, U^{0})$. The minimum expenditure function, $E = E(\boldsymbol{y}, n, \boldsymbol{v}, k, U^{0})$ is obtained by substituting these demand functions into (A4.1). The indirect utility function, $V = V(\boldsymbol{y}, n, \boldsymbol{v}, m, k)$, follows from inverting the expenditure function. Duality between utility maximization and expenditure minimization implies that the expenditure, $\boldsymbol{py}+m$, that results in a maximum value of utility, U^{*}, is also the minimum expenditure required to achieve a utility level of $U^{0} = U^{*}$. Hence, $E(\boldsymbol{y}, n, \boldsymbol{v}, k, U^{0}) = \boldsymbol{py}+m$.

HLMM (1995, 1996, 1999) adapted the expenditure function of the AID System to represent generalized managerial preferences:

(A4.4) $$ln\, \mathrm{E}(\cdot) \;=\; ln\, \mathrm{P} \;+\; \mathrm{U}\cdot\beta_{0}\left(\prod_{i} y_{i}^{\beta_{i}}\right)\left(\prod_{j} w_{j}^{v_{j}}\right) p_{\pi}^{\mu} k^{\kappa},$$

where $\ln \mathbf{P} = \alpha_{0} + \sum_{i} \alpha_{i} \ln z_{i} + (\tfrac{1}{2})\sum_{i}\sum_{j} \alpha_{ij} \ln z_{i} \ln z_{j}$ and $\boldsymbol{z} = (\boldsymbol{y}, n, \boldsymbol{v}, k)$. Inverting the expenditure function (A4.4) gives the indirect utility function:

(A4.5) $$V(\cdot) \;=\; \frac{ln(\boldsymbol{py} + m) - ln\, P}{\beta_{0}\left(\prod_{i} y_{i}^{\beta_{i}}\right)\left(\prod_{j} w_{i}^{v_{j}}\right) p_{\pi}^{\mu} k^{\kappa}} \quad .$$

Applying Shephard's lemma to (A4.4) yields the functional forms for the profit demand constant-utility input demand equations, and substituting the indirect utility function (A4.5) into these constant-utility functions

transforms them into the utility-maximizing demand equations that are to be estimated:

(A4.6) $$\frac{\partial ln\, E}{\partial ln\, p_\pi} = \frac{p_\pi \pi}{\mathbf{p}\mathbf{y} + m} = \frac{\partial ln\, P}{\partial ln\, p_\pi} + \mu\,[ln(\mathbf{p}\mathbf{y} + m) - ln\, P]$$

(A4.7) $$\frac{\partial ln\, E}{\partial ln\, w_i} = \frac{w_i x_i}{\mathbf{p}\mathbf{y} + m} = \frac{\partial ln\, P}{\partial ln\, w_i} + \nu_i[ln(\mathbf{p}\mathbf{y} + m) - ln\, P] \quad \forall\, i.$$

In their logarithmic form, these demand equations take the form of shares of the expenditures, $w_i x_i$, and before-tax net cash flow, $p_\pi \pi$, in revenue, $\boldsymbol{p}\boldsymbol{y} + m$. They sum to one.

As in HLMM, to derive the optimal level of equity capital, which is a conditioning argument in (A4.6) and (A4.7), we use a first-order condition obtained by maximizing the conditional Lagrangean function for the utility maximization problem (16a) - (16c) in the text, evaluated at the conditional optimum,

(A4.8) $$\begin{aligned} V(\boldsymbol{y},\boldsymbol{q},n,\boldsymbol{v},m,k) \equiv\ & U(\pi(\cdot),\boldsymbol{x}(\cdot);\boldsymbol{y},\boldsymbol{q},n,\boldsymbol{v},k) \\ & + \lambda(\cdot)[\boldsymbol{p}\boldsymbol{y} + m - \boldsymbol{w}\boldsymbol{x}(\cdot) - p_\pi \pi(\cdot)] \\ & + \gamma(\cdot)[T(x(\cdot);\boldsymbol{y},\boldsymbol{q},n,k)], \end{aligned}$$

with respect to the level of equity capital, k:

(A4.9) $$\frac{\partial V(\cdot)}{\partial k} = \frac{\partial V(\cdot)}{\partial \ln k}\frac{\partial \ln k}{\partial k} = 0.$$

The system of equations estimated comprises the profit and input share equations, (A4.6) and (A4.7), and the first-order condition for the optimal level of equity capital, (A4.9). Adding-up, homogeneity, and symmetry were imposed (see HLMM (1995, 1996) for details of these restrictions).