WHO SAID LARGE BANKS DON'T EXPERIENCE SCALE ECONOMIES? EVIDENCE FROM A RISK-RETURN-DRIVEN COST FUNCTION

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

April 2013

Abstract

The Great Recession focused attention on large financial institutions and systemic risk. We investigate whether large size provides any cost advantages to the economy and, if so, whether these cost advantages are due to technological scale economies or too-big-to-fail subsidies. Estimating scale economies is made more complex by risk-taking. Better diversification resulting from larger scale generates scale economies but also incentives to take more risk. When this additional risk-taking adds to cost, it can obscure the underlying scale economies and engender misleading econometric estimates of them. Using data pre- and post-crisis, we estimate scale economies using two production models. The standard model ignores endogenous risk-taking and finds little evidence of scale economies. The model accounting for managerial risk preferences and endogenous risk-taking finds large scale economies, which are not driven by too-big-to-fail considerations. We evaluate the costs and competitive implications of breaking up the largest banks into smaller banks.

*The views expressed in this paper do not necessarily reflect those of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. This paper is available free of charge at www.philadelphiafed.org/research-and-data/publications/working-papers/.

This paper supersedes Federal Reserve Bank of Philadelphia Working Paper No. 11-27.

We thank Marc Rodriguez and Arland Kane of the Board of Governors of the Federal Reserve System for their help with the data, Sally Burke for editorial assistance, and Christian Laux, conference participants at the International Banking, Economics, and Finance Association session at the Allied Social Sciences Association meetings in January 2011, and conference participants at the Financial Intermediation Research Society Annual Meeting in June 2011 for very helpful comments. We also thank the Whitcomb Center for Research in Financial Services at the Rutgers Business School for its support of data services used in this research.

Correspondence to Hughes at Department of Economics, Rutgers University, New Brunswick, NJ 08901-1248; phone: (917) 721-0910; 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; email: Loretta.Mester@phil.frb.org.

JEL Codes: D20, D21, G21, L23.
Key Words: banking, production, risk, scale economies, too big to fail.

For years the Federal Reserve was concerned about the ever-growing size of our largest financial institutions. Federal Reserve research had been unable to find economies of scale in banking beyond a modest size.

Alan Greenspan
"The Crisis" (*Brookings Papers on Economic Activity*, Spring 2010, p. 231)

I. Introduction

The financial crisis of 2007 focused attention on large financial institutions and the role the too-big-to-fail doctrine played in driving their size. Financial reform has focused on limiting the costs that systemically important financial institutions (SIFIs) impose on the economy. However, the potential efficiency benefits of operating at a large scale have been largely neglected in policy discussions and recent research. Textbooks explain that banks should enjoy scale economies as they grow larger because the credit risk of their loans and financial services, as well as the liquidity risk of their deposits, becomes better diversified. This reduces the relative cost of managing these risks and allows banks to conserve equity capital, as well as reserves and liquid assets. In addition, textbooks point to the spreading of overhead costs, especially those associated with information technology, as another source of scale economies. Network economies, such as those found in payments systems, have been cited as another source of financial scale economies. But the financial crisis has led many to question whether such efficiencies exist or whether scale has been driven primarily by institutions seeking to exploit the cost advantages of being too big to fail.

Older empirical studies that used data from the 1980s did not find scale economies in banking except at very small banks. But more recent studies that used data from the 1990s and 2000s and more modern methods for modeling bank technology that incorporate managerial preferences for risk and endogenize bank risk-taking find significant scale economies at banks of all sizes included in the sample.[1] These studies include Hughes, Lang, Mester, and Moon (1996, 2000), Berger and Mester (1997), Hughes and Mester (1998), Hughes, Mester, and Moon (2001), Bossone and Lee (2004), Wheelock and Wilson

[1] See Mester (2010) and Hughes (forthcoming) for further discussion.

(2012), and Feng and Serletis (2010). Hughes and Mester (2010) discuss some of these modern methods of modeling bank technology and the evidence of scale economies obtained from them. Part of the difference in results between the older studies and more recent ones appears to reflect improvements in the methods researchers use for measuring scale economies and part reflects a change in banking technology, such as the use of information technologies, and environmental factors, such as geographic deregulation, which have led to a larger efficient scale of banking production.

This investigation uses the modeling techniques developed by Hughes, Lang, Mester, and Moon (1996, 2000), and Hughes, Mester, and Moon (2001). These earlier papers used 1989-90 data on individual commercial banks and 1994 data on top-tier bank holding companies in the U.S., while here we use data from 2003, 2007, and 2010.[2] During the years that separate the earlier and later data sets, advances in information technology and further implementation of this technology in banking, as well as greater diversification from geographic consolidation, might be expected to increase scale economies in banking. And indeed, consistent with the textbook prediction and with consolidation in the banking industry, we find large scale economies at small banks and even larger scale economies at large banks. In addition, we find that controlling for size, more efficient banks enjoy higher scale economies than less efficient banks. The finding of significant scale economies even at banks that are not at a size usually considered too big to fail suggests that government policy is not the only source of size-related cost economies. It also suggests that a size limit on banks would not eliminate the market incentives to grow larger and therefore may result in unintended consequences of encouraging banking activities to move outside of the regulated financial services industry.

We present evidence below that too-big-to-fail considerations are not the source of the scale economies we find. In addition, we provide estimates of the cost impact of breaking up banks into smaller institutions, as some have proposed. In performing this exercise we take into account not only the size of the banks but also the potential longer-run impact that accounts for the fact that smaller banks

[2] The BHCs in our data set range in size from $64 million to $2.27 trillion in total consolidated assets. We performed additional tests that show that our results are robust to estimating the model excluding the largest banks in the sample, estimating the model excluding smaller banks in the sample, and estimating the model excluding extreme values for the output shares.

focus on product offerings that are different from those of larger banks. Our results suggest that reducing the size of the largest financial institutions by scaling back their chosen mix of financial products and services proportionately would significantly raise the costs of production. The higher total cost of the increased number of smaller banks required to replace the output of the larger banks would likely undermine the global competitiveness of U.S. banks. On the other hand, if the broken-up banks produce the mix of financial products and services of smaller institutions, their total costs would be slightly lower. Whether this is socially beneficial, however, depends on whether the product mix offered by the largest banks was beneficial – a question that is beyond the scope of the current paper.

The paper proceeds as follows. Sections II-IV discuss the theoretical model that incorporates bank managers' risk preferences and endogenous choice of risk. Those mainly interested in the empirical results can skip to Sections V-VI, which discuss the empirical model specifications. Section VII discusses our data set. Sections VIII-X give our empirical results, and Section XI concludes.

II. Modeling Banking Risk and Why It Matters for Uncovering Scale Economies

According to the standard textbook, a cost function uses input prices to translate the production function into the minimum cost of producing output. The textbook usually illustrates the cost function in terms of an expansion path graphed on an isoquant map. The expansion path is the locus of points where the marginal rate of substitution equals the ratio of input prices. The older literature on modeling bank cost functions often applied these concepts in a very straightforward way to bank production. It considered how to specify outputs and inputs in terms of bank assets, financial services, and liabilities. After calculating input prices, it derived a cost function for econometric estimation, applied it to bank data, and computed scale economies from the fitted function. As noted above, the results usually offered no evidence of scale economies at large banks.

Hughes, Lang, Mester, and Moon (1996, 2000), and Hughes, Mester, and Moon (2001) argue that the standard specification of the cost function fails to capture an essential ingredient in bank production – *risk*. Systematic differences in risk among banks can significantly alter how their cost varies with output

and consequently engender misleading econometric estimates of their scale economies when endogenous risk-taking is not taken into account in modeling and estimating bank cost. Bank managers' risk preferences are typically not modeled in standard cost function analysis, yet managers face a risk-expected return trade-off determined by the investment strategy they choose and the economic environment in which they operate. Thus, a bank's cost depends on its risk exposure, which contains an exogenous component reflecting the economic environment and an endogenous component reflecting the managers' choice of risk exposure.

The standard textbook explains that banks might enjoy scale economies derived from the diversification of risk obtained from a larger portfolio of loans and a larger base of deposits. These diversification benefits allow larger banks to manage risk with relatively fewer resources. In other words, a larger scale of operations improves a bank's risk-return trade-off. Figure 1 shows a smaller bank's investment strategies on the risk-return frontier labeled *I* and a larger bank's strategies on frontier *II*. Suppose that in Figure 1, point *A* represents production of a smaller, less diversified output, say, some quantity of loans with a particular probability distribution of default that reflects the contractual interest rate charged and the resources allocated to risk assessment and monitoring. Point *B* represents a larger quantity of loans with the same contractual interest rate but better diversification and, hence, an improved probability distribution of default and lower overall risk. The better diversification allows the costs of risk management to increase less than proportionately with the loan volume while maintaining an improved probability distribution of default. Thus, the response of cost to the increase in output from point A to point B reflects scale economies and the expected return at *B* exceeds that at *A*.

Suppose, instead, that the larger, better diversified portfolio of loans is produced with the investment strategy at point *C*. The strategy at *C* preserves the risk exposure of *A*, and the better diversification improves the expected return. The bank at *C* may charge a higher contractual interest rate, which would tend to increase risk by attracting riskier borrowers, but the better diversification offsets the additional risk. The cost of managing the larger loan portfolio at the same risk as *A* may still increase less

than proportionately, but the increase will be greater than that occasioned by *B*. Thus, the change in cost from *A* to *C* may still show scale economies, though smaller than from *A* to *B*.

On the other hand, suppose that the bank responds to the better diversification of the larger output by adopting a more risky investment strategy for an enhanced expected return. It charges an even higher contractual interest rate on loans than at point *C*. Better diversification does not offset the increased cost occasioned by the additional default risk. Point *D* in Figure 1 designates this strategy. The increased inherent default risk due to the higher contractual interest rate results in costs of risk management that increase more than proportionately with the loan volume (from *A* to *D*), and production appears to exhibit the counter-intuitive *scale diseconomies* found by empirical studies of banking cost that fail to account for endogenous risk-taking.[3]

While the investment strategies at *B*, *C*, and *D* entail producing the *same quantity of loans*, the expected return and its associated cost and risk of producing the loans differ across the three strategies. Figure 2 illustrates this point. It characterizes the production technology for a quantity of loans represented by the isoquant shown in the figure. The mix of debt and equity used to fund the loans is ignored. Instead, the diagram shows the quantity of physical capital and labor used in the process of credit evaluation and loan monitoring. (As the argument that follows illustrates, this isoquant is not well defined in traditional terms.) Point *C* shows the least costly way to produce the particular quantity of loans with the risk exposure associated with the investment strategy *C* in Figure 1. If a bank adopted the less risky strategy, *B*, it might use less labor in credit evaluation and monitoring: point *B* in Figure 2, a less costly method of producing the *same quantity of loans*. Thus, the isoquant for this quantity of loans that passes through point *C* captures one investment strategy only. If the isoquant included a characterization of the risk exposure, there would be another isoquant passing through point *B* for the same quantity of loans produced with the lower risk strategy. On the other hand, if a bank adopted the

[3] Demsetz and Strahan (1997) contend that larger banks are better diversified but take on more risk than smaller banks. Like us, they argue that to find evidence of this better diversification, researchers must control for the sources of endogenous risk. They estimate asset pricing models to obtain firm-specific risk, which they in turn regress on sources of risk taking and on asset size. With no controls, risk and asset size are weakly negatively related, but controlling for risk, the negative relationship is strong and large in magnitude – evidence of better diversification.

more risky strategy, *D*, it would use more labor, the corresponding point *D* in Figure 2, a more costly method than *C*. Thus, the cost of producing this particular quantity of loans depends on a bank's choice of risk exposure and its expected return. As in Hughes, Lang, Mester, and Moon (1996, 2000) and Hughes, Mester, and Moon (2001), we refer to this risk-return-driven cost as the managerial most preferred cost function, since it reflects managers' preferences over investment strategies that reflect the risk-expected return trade-off.

As explained in Hughes, Mester, and Moon (2001), failing to account for endogenous risk-taking when estimating a production model can produce misleading estimates of scale economies and cost elasticities. If production is observed at points *A* and *B*, a naïve calculation of the cost elasticity from the difference in cost measured at these two points would appear to yield evidence of scale economies. If production is observed at points *A* and *D*, a naïve calculation from their difference in cost would appear to give evidence of scale diseconomies. Thus, the specification of the cost function to be estimated must account for endogenous risk-taking to detect the scale economies associated with better diversification.

III. Modeling Managers' Preferences for Expected Return and Risk

Figures 1 and 2 illustrate that the cost of producing the larger, better diversified output depends on managers' choice of investment strategy in response to the better risk-expected return trade-off. Thus, cost is not independent of managers' risk preferences. Why might risk influence banks' production choices?

Modern banking theory emphasizes that bank managers face dichotomous investment strategies for maximizing value: one, higher risk; the other, lower risk (Marcus, 1984). The higher risk strategy, characterized in part by a lower capital ratio and lower asset quality, exploits mispriced deposit insurance, too-big-to-fail policies, and other benefits of the governmental safety net. Of course, this strategy also increases the risk of financial distress – possibly involving regulatory intervention in the operations of the bank, liquidity crises, and even insolvency and loss of the bank's charter. Such a risky strategy enhances a bank's value when its investment opportunities are not particularly valuable: the expected gains from

exploiting safety-net subsidies outweigh the potential losses entailed in episodes of financial distress. On the other hand, if a bank enjoys valuable investment opportunities, these market advantages increase its expected costs of financial distress. When the expected losses involved in financial distress exceed the expected gains from exploiting the safety net, banks enhance their value by pursuing a lower-risk strategy involving a higher capital ratio and higher asset quality.[4] Both of these investment strategies maximize firm value. Hence, risk-neutral managers would pursue them. They manage risk when doing so maximizes value (Tufano, 1996).

These value-maximizing, dichotomous investment strategies highlight the importance of accounting for endogenous risk-taking in estimating production costs in banking. Modeling managers' risk preferences forms the foundation for building a model of bank production and cost.

We turn first to some notational matters. We represent bank technology by the transformation function, $T(\mathbf{y}, n, p, \mathbf{x}, k) \leq 0$, where $\mathbf{y}$ denotes information-intensive loans and financial services; k, equity capital; $\mathbf{x}_d$, demandable debt and other types of debt; $\mathbf{x}_b$, labor and physical capital; and $\mathbf{x} = (\mathbf{x}_b, \mathbf{x}_d)$. The price of the i-th type of input is designated by w_i so that the economic cost of producing the output vector $\mathbf{y}$ is given by $\mathbf{w}_b\mathbf{x}_b + \mathbf{w}_d\mathbf{x}_d + w_k k$. If the cost of equity capital is omitted, $\mathbf{w}_b\mathbf{x}_b + \mathbf{w}_d\mathbf{x}_d$ gives the *cash-flow cost* (C_{CF}). We characterize asset quality by two types of proxies: *ex ante* measures are given by the vector of average contractual interest rates on assets such as securities and loans, $\mathbf{p}$, which, given the risk-free interest rate, r, captures an average risk premium, and an *ex post* measure, the dollar amount of nonperforming loans, n.

Rather than express managers' preferences in terms of how they rank expected return and return risk, the first two moments of the subjective distribution of returns, we ask how managers rank production plans. Production plans are more basic: to rank production plans, managers must translate plans into subjective, conditional probability distributions of profit. Managers' beliefs about the probability distribution of states of the world, s_t, and about how the interaction of production plans with states yields

[4] For empirical evidence of these dichotomous strategies, see Keeley (1990) and Hughes, Lang, Moon, and Pagano (1997 and 2004).

a realization of after-tax profit, $\pi = g(\boldsymbol{y}, n, \boldsymbol{p}, r, \boldsymbol{x}, k, s)$, imply a subjective distribution of profit that is conditional on the production plan: $f(\pi; \boldsymbol{y}, n, \boldsymbol{p}, r, \boldsymbol{x}, k)$. Under certain restrictive conditions, this distribution can be represented by its first two moments, $E(\pi, \boldsymbol{y}, n, \boldsymbol{p}, r, \boldsymbol{x}, k)$ and $S(\pi, \boldsymbol{y}, n, \boldsymbol{p}, r, \boldsymbol{x}, k)$. Rather than define a utility function over these two moments, we define it over profit and the production plan, $U(\pi, \boldsymbol{y}, n, \boldsymbol{p}, r, \boldsymbol{x}, k)$, which is equivalent to defining it over the conditional probability distributions $f(\cdot)$. This generalized managerial utility function subsumes the case of profit maximization where only the first moment of the conditional distribution of profit influences utility; however, it also explains cases where higher moments influence utility so that managers can trade profit to achieve other objectives involving risk.

IV. Modeling Cost When Risk Is Endogenous

The cost of producing a particular output vector $\boldsymbol{y}$ – financial assets and services – depends on the employment of inputs $\boldsymbol{x}$ and k – labor, physical capital, debt, and equity. How managers choose to produce any particular output vector can be modeled as a utility-maximization problem. Hence, the choice from the production strategies highlighted in Figures 1 and 2, points *B*, *C*, and *D*, solves the utility-maximization problem.

Since the utility function ranks production plans – output and input vectors and the resulting profit – banks maximize utility conditional on the output vector by solving for the utility-maximizing profit and the constituent vector of inputs required to produce it. Let m designate noninterest income and let $\boldsymbol{p} \cdot \boldsymbol{y}$ represent interest income. Total revenue is given by $\boldsymbol{p} \cdot \boldsymbol{y} + m$. Letting π designate after-tax profit and t, the tax rate on profit, and $p_\pi = 1/(1 - t)$, the price of a dollar of after-tax profit in terms of before-tax dollars, the before-tax accounting or cash-flow profit is defined as $p_\pi \pi = \boldsymbol{p} \cdot \boldsymbol{y} + m - \boldsymbol{w_b} \cdot \boldsymbol{x_b} - \boldsymbol{w_d} \cdot \boldsymbol{x_d}$.

The utility-maximization problem is given by:

$$\text{(1a)} \qquad \max_{\pi, \boldsymbol{x}} U(\pi, \boldsymbol{x}; \boldsymbol{y}, n, \boldsymbol{p}, r, k)$$

$$\text{(1b)} \qquad \text{s.t. } p_\pi \pi = \boldsymbol{p} \cdot \boldsymbol{y} + m - \boldsymbol{w_b} \cdot \boldsymbol{x_b} - \boldsymbol{w_d} \cdot \boldsymbol{x_d}$$

(1c) $$T(\mathbf{y}, n, \mathbf{p}, r, \mathbf{x}, k) \leq 0.$$

The solution gives the *managers' most preferred profit function*, $\pi^* = \pi_{MP}(\mathbf{y}, n, \mathbf{v}, k)$, and the *managers' most preferred input demand functions*, $\mathbf{x}^* = \mathbf{x}_{MP}(\mathbf{y}, n, \mathbf{v}, k)$, where $\mathbf{v} = (\mathbf{w}, \mathbf{p}, r, m, p_\pi)$. The *managers' most preferred cost function* follows trivially from the profit function:

(2) $$C_{MP}(\mathbf{y}, n, \mathbf{v}, k) = \mathbf{p} \cdot \mathbf{y} + m - p_\pi \pi_{MP}(\mathbf{y}, n, \mathbf{v}, k).$$

We claimed above that this utility-maximization problem has sufficient structure to identify and control for the choice of production plan from points *B*, *C*, and *D* of Figures 1 and 2 – plans that produce the same output, $\mathbf{y}$, but differ in their risk exposure and resources allocated to managing risk. How then does the solution, the most preferred profit and cost functions and the most preferred input demand functions, depend on the risk exposure?

First, note that revenue, $\mathbf{p} \cdot \mathbf{y} + m$, drives the solution. In addition, the output prices, $\mathbf{p}$, which are contractual returns on assets such as loans and securities, control for the *ex ante* risk premium of each of those assets when they are compared to the risk-free interest rate *r*. The quantity of nonperforming assets, *n*, captures *ex post* or realized default risk. The quantity of equity capital, *k*, controls for a key component of capital structure that underlies expected return and return risk. Moreover, since the cost of equity and loan losses are excluded from the calculation of cash-flow cost and profit, the quantities of equity and nonperforming loans control for these omitted expenses. These controls, as well as the tax rate on earnings embodied in the price of a before-tax dollar, p_π, in terms of after-tax dollars, constitute a rich characterization of investment strategies that shape cost.

These variables that characterize and control for the investment strategy permit the calculation of risk-adjusted scale economies from the estimated cost function – a calculation that accounts for the bank's choice of risk exposure. In Figures 1 and 2 the problem of identifying the points *B*, *C*, and *D* for the purpose of computing scale economies is resolved by these control variables in the theoretical and empirical framework of managerial utility maximization. Note that to the extent that larger banks and

smaller banks choose a different product mix with different risk characteristics – e.g., larger banks produce more off-balance-sheet activities than smaller banks – by controlling for risk preferences, this cost model allows us to include banks of all sizes in our estimation.

V. Using the Almost Ideal Demand System to Estimate the Most Preferred Cost Function and Scale Economies

To estimate the utility-maximizing profit and input demand functions that solve the problem (1a), (1b), and (1c), we follow Hughes, Lang, Mester, and Moon (1996, 2000), and Hughes, Mester, and Moon (2001) and adapt the Almost Ideal Demand System of consumer theory, which was proposed by Deaton and Muellbauer (1980), to represent managerial preferences. As Deaton and Muellbauer note, the AI demand system is a flexible functional form that has many advantages over the translog functional form. Just as the estimation of this system using budget data recovers consumers' preferences for goods and services, its application to banks' data on production and cost recovers managers' rankings of production plans or, equivalently, their ranking of subjective probability distributions of profit conditional on the production plan. The AI production system we estimate allows for the possibility that managers trade off profit for reduced risk and, hence, incur higher costs for reduced risk. This is not possible in the translog production system and the Fourier flexible cost function often used in the literature. Indeed, the AI production system embeds the typical translog production system as a special case. In particular, the AI system allows one to test whether the firms are minimizing cost and maximizing profits. As shown in Hughes, Lang, Mester, and Moon (1996, 2000), if the data are consistent with firms minimizing cost and maximizing profits, then the AI production system reduces to the translog production system. In the literature, applications of the AI production system to banking have rejected the assumptions of cost minimization and profit maximization.

The profit equation and input demands are expressed as expenditure shares of total revenue:

(3a) $$\frac{p_\pi \pi}{\mathbf{p} \cdot \mathbf{y} + m} = \frac{\partial \ln \boldsymbol{P}}{\partial \ln p_\pi} + \mu[ln(\mathbf{p} \cdot \mathbf{y} + m) - \ln \boldsymbol{P}]$$

(3b) $$\frac{w_i x_i}{\mathbf{p} \cdot \mathbf{y} + m} = \frac{\partial \ln \boldsymbol{P}}{\partial \ln w_i} + \nu_i[ln(\mathbf{p} \cdot \mathbf{y} + m) - \ln \boldsymbol{P}] \; \forall i$$

where $ln\ \mathbf{P} = A_0 + \Sigma_i A_i\ ln\ y_i + (½)\Sigma_i \Sigma_j S_{ij}\ ln\ y_i\ ln\ y_j + \Sigma_i B_i\ ln\ w_i + (½)\Sigma_i \Sigma_j G_{ij}\ ln\ w_i\ ln\ w_j$

$+ \Sigma_i \Sigma_j D_i\ ln\ y_i\ ln\ w_j + (½)\Sigma_i \Sigma_j R_{ij}\ ln\ z_i\ ln\ z_j + \Sigma_i \Sigma_j H_{ij}\ ln\ z_i\ ln\ y_j + \Sigma_i \Sigma_j T_{ij}\ ln\ z_i\ ln\ w_j$;

and $\boldsymbol{z} = (k, n, \mathbf{p}, p_\pi)$. The input shares and profit share sum to one.

Equity capital enters the specification of the profit and input demand equations as a conditional argument. Hence, we include in the estimation a first-order condition defining the utility-maximizing value of equity capital:

(3c) $$\frac{\partial V(\cdot)}{\partial k} = \frac{\partial V(\cdot)}{\partial \ln k} \frac{\partial \ln k}{\partial k} = 0,$$

where the indirect utility function, $V(\cdot)$, of the maximization problem (1a)-(1c) is

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

Equation 2 shows how the managers' most preferred cost function is derived from the profit function. We compute the measure of scale economies, the inverse of the cost elasticity with respect to output, from this expression after substituting the optimal demand for equity capital into it:

$$\textit{most preferred cost economies} = \frac{C_{MP}}{\sum_i y_i \left[\frac{\partial C_{MP}}{\partial y_i} + \frac{\partial C_{MP}}{\partial k} \frac{\partial k}{\partial y_i} \right]}$$

$$= \frac{\mathbf{p} \cdot \mathbf{y} + m - p_\pi \pi}{\sum_i y_i \left[p_i - \frac{\partial p_\pi \pi}{\partial y_i} - \frac{\partial p_\pi \pi}{\partial k} \frac{\partial k}{\partial y_i} \right]} . \tag{5}$$

A value of the expression in equation (5) greater than one implies scale economies, and a value less than one implies scale diseconomies.

Managerial preferences 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, so managers' preferences change over time. Consequently, we use cross-sectional data and estimate the production system each year with nonlinear two-stage least squares, which is a generalized method of moments technique. The appendix gives the details of empirical specification and estimation.

In addition, Hughes (1999) and Hughes, Mester and Moon (2001) report that the estimated predicted profit from the AI production system and the standard error of that predicted profit, which is a measure of profit risk, explain 96 percent of the variation in the market value of banks' equity for a sample of 190 publicly traded bank holding companies in 1994. This indicates that the predicted profit and profit risk captured by the AI production system characterize bank performance that is priced by capital markets, which lends credibility to the model.[5] Normalizing these measures of predicted profit and profit risk by equity produces measures of expected return and return risk, which can be used to estimate an efficient risk-return frontier. The frontier can then be used to measure a financial institution's return efficiency for a given level of risk exposure. We apply these methods below when investigating scale economies at a set of efficient banks.

[5] If the arguments of the profit function are considered factors in explaining expected profit or return, the fitted coefficients in the regression of bank equity value on predicted profit and profit risk can be interpreted as marginal returns to the factors. Since the standard error of the predicted profit is a function of the variance-covariance matrix of coefficients, it resembles the variance-covariance matrix of security returns in a portfolio of traded securities. In this case, the financial institution holds a levered portfolio of traded and produced securities and services.

VI. Minimum Cost Functions and Scale Economies

The standard minimum cost function is quite different from the most preferred cost function just discussed. The standard cost function can control for some aspects of risk, including the amount of nonperforming loans, *n*, which accounts for the influence of asset quality on cost. In addition, the important role of equity capital in banking production suggests that the minimum cost function should include either the required return (price) or quantity of equity capital. When the required return is not readily available (and it is not, since most banks are not publicly traded), the minimum cost function can be conditioned on equity capital. In this case, the cost function excludes the cost of equity capital and, thus, is *cash-flow* cost. Note that this function fails to account for the revenue side of expected return and return risk that are found in the specification of the most preferred profit and cost functions. Thus, the standard cash-flow cost function is:

$$(6) \qquad C_{CF}(\boldsymbol{y}, n, \boldsymbol{w_b}, \boldsymbol{w_d}, k) = \min_{\boldsymbol{x_b}, \boldsymbol{x_d}} (\boldsymbol{w_b} \cdot \boldsymbol{x_b} + \boldsymbol{w_d} \cdot \boldsymbol{x_d}) \text{ s.t. } T(\boldsymbol{y}, n, \boldsymbol{x}, k) \leq 0 \text{ and } k = k^0.$$

We estimate this cost function and its associated share equations with a translog specification:

$$ln\, C_{CF} = \alpha_0 + \Sigma_i \alpha_i\, ln\, g_i + (\tfrac{1}{2})\Sigma_{ij} \alpha_{ij}\, ln\, g_i\, ln\, g_j \text{ and } \boldsymbol{g} = (\boldsymbol{y}, n, \boldsymbol{w}, k).$$

Scale economies based on this cash-flow cost function are:

$$(7) \qquad cash\text{-}flow\ scale\ economies\ from\ the\ C_{CF}\ cost\ function = \frac{1}{\sum_i \frac{\partial \ln C_{CF}}{\partial \ln y_i}}\ .$$

Some studies of banking technology neglect the critical role of equity capital by defining a minimum cash-flow cost function without conditioning it on the amount of equity capital:

$$(8) \qquad C_{MS}(\boldsymbol{y}, n, \boldsymbol{w_b}, \boldsymbol{w_d}) = \min_{\boldsymbol{x_b}, \boldsymbol{x_d}} (\boldsymbol{w_b} \cdot \boldsymbol{x_b} + \boldsymbol{w_d} \cdot \boldsymbol{x_d}) \text{ s.t. } T(\boldsymbol{y}, n, \boldsymbol{x}) \leq 0.$$

To illustrate the bias introduced by such a cash-flow cost function, consider two banks identical in every respect except their capital structures. One bank uses less equity and more debt to finance the same quantity of assets. Thus, its cash-flow cost of producing the same output will be greater because it incurs the interest cost of the additional debt. Since cash-flow cost does not account for the cost savings of less equity, it appears to be a more costly method of producing the same output. Had the cash-flow cost function been properly conditioned on the amount of equity capital employed, the appearance of a less efficient production method would have been dispelled. Thus, the specification of cost in (8) is theoretically mis-specified, so we label it with the *MS* subscript. For illustrative purposes, we estimate this cost function and its associated share equations with a translog specification:

$$ln\ C_{\mathbf{MS}} = \alpha_0 + \Sigma_i\, \alpha_i\ ln\ h_i + (\tfrac{1}{2})\Sigma_i\, \Sigma_j\, \alpha_{ij}\ ln\ h_i\, ln\ h_j \text{ and } \boldsymbol{h} = (\boldsymbol{y},\ n,\ \boldsymbol{w}).$$

Scale economies based on this cost function are given by:

$$(9) \qquad cash\text{-}flow\ scale\ economies\ \ from\ the\ C_{MS}\ \ cost\ \ function = \frac{1}{\sum_i \frac{\partial \ln C_{MS}}{\partial \ln y_i}}\ .$$

In contrast to these cash-flow cost functions, consider an economic cost function that includes the cost of equity capital:

$$(10) \qquad C_{\mathbf{EC}}(\boldsymbol{y},\ n,\ \boldsymbol{w_b},\ \boldsymbol{w_d},\ w_k) = \min_{\boldsymbol{x_b},\ \boldsymbol{x_d},\ k} (\boldsymbol{w_b} \cdot \boldsymbol{x_b} + \boldsymbol{w_d} \cdot \boldsymbol{x_d} + w_k k) \text{ s.t. } T(\boldsymbol{y},\ n,\ \boldsymbol{x},\ k) \le 0.$$

Since the economic cost function includes the cost of equity capital, it is conditioned on the required return (price) rather than the quantity of equity capital. When a bank is publicly traded, the required return, w_k, can be computed from an asset pricing model; however, most banks are not publicly traded. Instead, the cash-flow cost function in (6) is used to obtain a shadow price of equity capital from which

the economic cost function and its associated scale economies can be computed.[6] The first-order condition for optimal equity capital gives its shadow price:

$$(11) \qquad w_k = -\frac{\partial C_{CF}}{\partial k}.$$

Then the economic cost function is:

$$(12) \quad C_{EC}(y, n, \boldsymbol{w_b}, \boldsymbol{w_d}, w_k) = \min_k C_{CF}(y, n, \boldsymbol{w_b}, \boldsymbol{w_d}, k) + w_k k = \min_k C_{CF}(y, n, \boldsymbol{w_b}, \boldsymbol{w_d}, k) - \frac{\partial C_{CF}}{\partial k} k.$$

If we assume that the observed level of equity capital is cost-minimizing, then marginal cost computed from the cash-flow cost function equals marginal cost computed from the economic cost function:[7]

$$(13) \qquad \frac{\partial C_{EC}}{\partial y_i} = \frac{\partial C_{CF}}{\partial y_i} \; \forall i.$$

Then, using (12) and (13), the degree of scale economies based on the economic cost function is given by:

$$(14) \quad economic\ cost\ scale\ economies\ from\ the\ C_{EC}\ cost\ function = \frac{1}{\sum_i \frac{\partial \ln C_{EC}}{\partial \ln y_i}}$$

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

[6] Braeutigam and Daughety (1983) first suggested this technique, and Hughes, Mester, and Moon (2001) applied it to banking production and cost.

[7] Interpreting this proposition in terms of long-run (economic) cost and short-run variable (cash-flow) cost, it illustrates the familiar result that long-run and short-run marginal costs are equal when the value of the "fixed" input that gives rise to short-run variable cost minimizes long-run cost at the given output vector. Berger, DeYoung, Flannery, Lee, and Öztekin (2008) find that banks hold more equity capital than required by their regulators, which need not be the cost-minimizing level. Using 1994 data, Hughes, Mester, and Moon (2001) find that smaller banks appear to overutilize equity capital, while large banks appear to underutilize equity capital relative to the cost-minimizing level. The most preferred production model includes a demand for capital equation and so allows for the possibility that bank managers choose a level of capital that is not cost-minimizing.

VII. The Data

Our data set includes 842 top-tier bank holding companies (BHCs) in the United States in 2007, and for robustness, we also estimate our model for the 1,855 top-tier BHCs in 2003 and 856 top-tier BHCs in 2010. A top-tier company is not owned by another company. The data are obtained from the Y-9 C Call Reports filed quarterly with bank regulators. We model the consolidated bank rather than its constituent banks and subsidiaries because investment decisions are generally made at the consolidated level; this also allows us to avoid the problems associated with transfer pricing within the organization. The summary statistics describing these banks are found in Tables 1-5.

Estimating a flexible functional form like the AI production system requires a degree of parsimony in specifying outputs, since each output adds dozens of parameters for estimation. On the other hand, disaggregating outputs enhances the characterization of the differences in investment strategies of banks of all sizes. To balance these conflicting goals of disaggregation and parsimony, we specify five outputs and then check robustness using a variation of the output definitions.

The first output, y_1, includes the liquid assets, cash, repos, federal funds sold, and interest-bearing deposits due from banks. The second output, y_2, is securities, including U.S. Treasury and U.S. government agency securities, as well as nongovernmental securities. We distinguish securities from other liquid assets because securities, especially mortgage-backed securities, have played an important role in bank production as a source of income and also as a troubled asset whose liquidity sometimes became compromised during the period covered by the data.

The third output, y_3, captures lending activity and comprises loans on the balance sheet, assets sold during the year without securitization, and assets securitized with servicing retained or with recourse or other credit enhancements. On-balance-sheet loans entail both funding costs and costs of credit evaluation and monitoring. Loans originated and eventually sold or securitized, while ultimately not on the balance sheet, nevertheless incur costs of origination and costs of monitoring and funding during their time on the balance sheet. Loans sold with servicing retained continue to generate monitoring and servicing costs; those sold with recourse generate capital costs because of the potential that such loans

will be brought back on the balance sheet should they become troubled, as well as additional risk management expenses. Only a few studies have taken into account sold loans in specifying banking outputs (see, e.g., Mester, 1992). However, during the period studied here, such asset sales were often an important activity of banks, even small banks (see, e.g., Erel, Nadauld, and Stulz, 2011). Also, since revenue and risk drive costs in our model, it is important to include these sold assets, which generate bank revenue. While this is our preferred specification, we checked robustness with an alternative specification of y_3, which includes only on-balance-sheet loans. The results are qualitatively very similar to the ones reported below.

The fourth output, y_4, comprises trading assets, investments in unconsolidated subsidiaries, intangibles, and other assets. The fifth output, y_5, captures off-balance-sheet activities, measured by their credit equivalent amount.[8]

The six inputs are: x_1, labor; x_2, physical capital; x_3, time deposits exceeding $100,000 (uninsured);[9] x_4, all other deposits (including insured deposits); x_5, all other borrowed funds, including foreign deposits, federal funds purchased, reverse repos, trading account liabilities, mandatory convertible securities, mortgage indebtedness, commercial paper, and all other borrowed funds; and k, equity capital consisting of equity, subordinated debt, and loan loss reserves. Except for equity capital, the other five input prices are computed as the expenditure on the input divided by the quantity of the input, and cost is defined as $\boldsymbol{w} \cdot \boldsymbol{x}$. The price of a dollar of after-tax profit in terms of before-tax dollars is $p_\pi = 1 / (1 - t)$, where the tax rate, t, is the highest marginal corporate tax rate in the state in which the bank holding company is headquartered plus the highest federal marginal tax rate (which is 35 percent). Revenue, $\boldsymbol{p} \cdot \boldsymbol{y} + m$, is the sum of interest and noninterest income.

[8] Some studies proxy the amount of off-balance-sheet activities by the net income they generate. However, this measure is biased downward by losses. The credit-equivalent amount is calculated by converting the various measures of off-balance-sheet activities into the equivalent amount of on-balance-sheet assets, adjusted by the latter's risk weight. Loans are weighted at 100 percent. A stand-by letter of credit is weighted at 100 percent, too, on the grounds that it generates the same amount of exposure to default risk as an on-balance-sheet loan.

[9] The limit was temporarily increased to $250,000 in October 2008 and permanently increased by the Dodd-Frank Act of 2010. In 2003 and 2007 the limit was $100,000.

We proxy *ex post* asset quality by the amount of nonperforming loans, which is the sum of past due loans, leases, and other assets, and assets in nonaccrual status, plus gross charge-offs of on-balance-sheet assets, plus other real estate owned in satisfaction of debts (i.e., real estate owned due to foreclosures), plus charge-offs on securitized assets. Because banks differ in their aggressiveness in charging off past due assets, we include gross charge-offs in the measure of nonperforming assets to eliminate any bias that might be caused by differences in charge-off practices.[10]

We proxy *ex ante* asset quality by the average contractual interest rate, p_i, on the ith output. The difference between this yield and the risk-free rate captures the risk premium incurred by the asset. Thus, the contractual interest rate captures both a component of revenue and a dimension of asset quality. Since interest income is not reported for all the outputs we specify, we use the weighted average of output prices, $\tilde{p}$, which is measured as the sum of interest income from accruing assets, trading income, income from securitization, income from servicing, and net income from assets sold, divided by the sum of all the outputs.

Table 1 describes the full sample we used in the estimations. Banks in 2007 range in assets from $72 million to $2.19 trillion. (The data for 2003 and 2010 are given in 2007 dollars. Real assets range from $73 million to $1.4 trillion for the 1,855 banks in the 2003 sample and from $97 million to $2.2 trillion for the 856 banks in the 2010 sample. The following discussion will focus on the 2007 data, but the results for 2003 and 2010 are similar.) Because of the flexible nature of the production model and the fact that we are controlling for risk preferences and asset quality by including a measure of nonperforming loans and the average implied interest rate on output, the model permits including a wide range of bank sizes. Tables 2-5 partition the data by asset size in order to show how the variables in Table 1 differ from small to very large banks. There is no official definition of too big to fail, but asset size of $100 billion or more has been considered a threshold for too big to fail in some studies.[11]

[10] If charge-offs were not included in the definition of asset quality, then a bank that was otherwise identical to another bank but was more aggressive in charging off nonperforming loans would appear to have better *ex post* loan quality.

[11] Brewer and Jagtiani (2009) give three too-big-to-fail size thresholds: (1) banks with total book value of assets of at least $100 billion, (2) banks that are one of the 11 largest organizations in each year (currently the 11th largest

As shown in Table 2, the mean level of loans as a proportion of total assets falls somewhat as banks get larger. The liquid assets ratio is also higher for banks in the larger two size groups, with assets over $50 billion, compared to banks with assets less than $50 billion. Trading and other assets as a proportion of total assets and the ratio of the credit-equivalent amount of off-balance-sheet assets to total assets both rise with bank size.

Table 3 details differences in input utilization. While labor as a proportion of total assets does not vary much across the size groups, the physical capital ratio declines somewhat. We also find that compared to smaller banks, larger banks fund a smaller proportion of their assets with insured deposits and a larger proportion with other borrowed funds (which include foreign deposits, commercial paper, federal funds purchased, securities sold under agreement to repurchase, trading account liabilities, and other borrowed money). Compared with insured deposits and other borrowed funds, uninsured deposits are a less important source of funds for all size groups.

Table 4 provides details of differences across size groups in risk exposure and financial performance. As banks increase in size across the six groups, their mean ratio of capital to assets also increases, while the mean ratio of nonperforming assets shows no monotonic pattern related to asset size, but it has risen over time. The rate of return on assets (ROA) (measured as profits/assets) is slightly higher for larger banks, but the differences in mean ROA are negligible.[12] The average contractual return on accruing assets is higher for smaller banks than for larger banks.

VIII. Evidence of Scale Economies

Before presenting our scale economies results, we present evidence in support of the managers' most preferred model. One benefit of the managers' most preferred model is that it allows for the

---

BHC has $290 billion in assets), and (3) banks with market value of equity ≥ $20 billion. The Dodd-Frank Act requires the Federal Reserve to conduct annual supervisory stress tests for BHCs with at least $50 billion in assets. Also, the Federal Reserve has designated BHCs and nonbank financial companies with at least $50 billion in assets as "significant" for the purposes of Dodd-Frank.

[12] This measure of profits is based on the theoretical definition used in the model and is total revenue minus the expense of variable inputs, $\boldsymbol{p} \cdot \boldsymbol{y} + m - \boldsymbol{w_b} \cdot \boldsymbol{x_b} - \boldsymbol{w_d} \cdot \boldsymbol{x_d}$. Note that the expense of variable inputs excludes depreciation, taxes, and the cost of the quasi-fixed factors equity capital and loan losses.

possibility that bank managers are not necessarily maximizing profits but are pursuing additional objectives, e.g., trading off expected profit and risk. As shown in Hughes, Lang, Mester, and Moon (1996), if banks are maximizing profits alone, this implies some restrictions on the most preferred model, which can be statistically tested. In particular, profit maximization implies that a variation in tax rate will not affect the bank's choice of before-tax profit, a variation in the revenue and risk characteristics of production represented by the output price vector will not influence the bank's cost-minimizing production plan, and a variation in *m* will not influence optimal input demands. Essentially, with these restrictions imposed, the most preferred model becomes a translog model in which the input profit-share equations are cost share equations identical to those derived from the translog cost function and the profit share equation is equivalent to the translog cost function.

We tested the linear restrictions for profit maximization using a Wald test and found that managers are not acting to maximize profit alone – i.e., the data overwhelmingly support estimation of the most preferred model. The value of the test statistic is 305.8 for 2007, 1041.1 for 2003, and 447.6 for 2010, with the p-value very close to 0 in all three cases.

The plausibility of the most preferred model and its estimates of scale economies depend in part on how well the estimated model gauges expected profit in the long run as well as in the short run. The accounting data used in estimating the model necessarily focus on the short run, but market value captures discounted expected future profits. Thus, as discussed in Hughes, Mester, and Moon (2001), we can gauge how well the production model captures from current period data managerial preferences for a longer horizon by examining how closely related the model's estimated expected profit, $E(p_\pi\pi)$, and profit risk, which is measured as the standard error of expected profit, $S(p_\pi\pi)$, are to market value.[13] Regressing the BHC's year-end market value of equity (*MVE*) on $E(p_\pi\pi)$ and $S(p_\pi\pi)$ for the publicly traded BHCs yields:

For 2007: $MVE = 526504 + 5.9325\, E(p_\pi\pi) - 75.4126\, S(p_\pi\pi)$ Adj. $R^2 = 0.9416$

[13] The standard error of predicted profit, which is a function of the variance-covariance matrix of the estimated parameters of the model, resembles the variance of a portfolio return when the parameter estimates are viewed as marginal expected returns. Thus, we used this measure as a proxy for the profit risk of banks' produced portfolios of financial products and services.

(331584) (0.3871) (10.0766) No. obs. = 226

For 2003: MVE = 98282 + 7.6219 $E(p_\pi\pi)$ – 108.3289 $S(p_\pi\pi)$ Adj. R^2 = 0.9924
(7469) (0.5958) (24.6706) No. obs. = 349

For 2010: MVE = 210593 + 6.2117 $E(p_\pi\pi)$ – 116.9897 $S(p_\pi\pi)$ Adj. R^2 = 0.9671
(258878) (0.4512) (13.2605) No. obs. = 220

As can be seen, the coefficients on expected profit and profit risk have the theoretically correct sign and are highly significant at the 1 percent level. The adjusted R^2s are very high, indicating that our production-based measures of expected profit and profit risk substantially explain market value.

Next, we investigate scale economies. We estimate the cost function and input share equations for the theoretically mis-specified cash-flow cost function (omitting the amount of equity capital), the theoretically proper cash-flow cost function (conditioned on the amount of equity capital), and the most preferred profit function and input demand functions.

Table 6 presents the estimated scale economies for these models for 2007. We also present the results for 2003, an earlier year that was prior to the crisis, and for 2010, a later year that was after the crisis. For each year, the first column of results shows that for the mis-specified cost function that omits any role for equity capital, all six size groups show evidence of scale economies that are statistically significantly greater than one. The differences across size groups in these measured scale economies are slight. And the means differ little from the medians. To obtain some intuition for the magnitude of the measures, consider two values, 1.03 and 1.06, at each end of the range for the six groups for 2007. If all outputs increase by 10 percent, at a scale measure of 1.03, cost increases by 9.7 percent; and, at 1.06, cost increases by 9.4 percent.

For each year, the second column of results in Table 6 shows estimates of scale economies for the theoretically correct specification that includes the quantity of equity capital as a conditioning argument but omits the cost of equity in the calculation of cost. Hence, we term the result "correct cash-flow cost." For the most part, these estimates show essentially constant returns to scale or, in some cases,

diseconomies of scale – e.g., the larger banks in 2010 and the smaller banks in 2003.[14] The estimates in column 2 tend to be lower than the estimates in column 1. This result reflects the relative costs of debt vs. equity financing.[15]

For each year, column 3 of Table 6 reports these scale economies based on economic cost that includes the cost of equity. In nearly all cases, adding the cost of equity increases the scale economies compared to the estimates of the cash-flow cost function that is conditioned on the level of equity but excludes its cost (column 2). In 2003 and 2007, and for banks with assets under $10 billion in 2010, the estimates are significantly greater than one and range from 1.03 and 1.06.

None of the cost models used to derive the scale estimates in columns 1-3 distinguish the differences in risk-expected return trade-offs that are inherent in the investment strategies of large and small banks. The most preferred cost function controls for these differences. For each year, we report the estimates of scale economies obtained from this cost function in column 4. The mean value of scale economies for the full sample is a significant 1.14 in 2007, 1.18 in 2003, and 1.25 in 2010. The estimated scale economies increase with bank size. For example, in 2007, estimated scale economies range from 1.12 for banks with less than $800 million in assets, to 1.34 for banks with over $100 billion in assets. For a 10 percent increase in all outputs, a scale measure of 1.12 implies an 8.8 percent increase in cost, while a scale measure of 1.34 implies a 7.5 percent increase in cost. Similar results hold for 2003 and 2010.

Robustness. We perform several robustness tests.

[14] The only exception is the largest size category of banks with assets over $100 billion in 2007, which shows scale economies.

[15] As discussed in Berger and Mester (1997), capital provides an alternative to deposits as a funding source for loans, so estimates of scale economies will depend on a bank's relative reliance on debt vs. equity financing, the relative costs of raising debt and equity, and whether the level of capital is controlled for in the cost function specification. On the one hand, interest paid on debt counts as a cost in the cash-flow cost functions but dividends paid on capital do not. On the other hand, the cost of raising equity is typically higher than the cost of raising deposits. If the first effect dominates, then measured cost will be higher for banks that use proportionately more debt to fund their assets, and scale economies would tend to be lower when the level of equity is controlled for (column 2) than when it is not controlled for (column 1). If the second effect dominates, then the opposite would obtain. Our results indicating that controlling for equity in the cost function tends to produce lower estimates of scale economies than when capital is not controlled for suggest that the first effect dominates.

(1) Even though our model is very flexible and we control for risk preferences and output quality, there may be some concern that we are including banks with very different production technologies in the estimation and that this is driving our results. However, this does not appear to be the case. First, we re-estimate our model excluding banks with assets of $2 billion or less. This leaves a sample of 215 bank holding companies. As seen in column 2 of Table 7, our scale results are very similar to those obtained with the full sample. For example, for 2007, scale economies are significantly different from one at the 1 percent level and increase with bank size, from 1.145 for banks in the $2 billion to $10 billion size category up to 1.365 for banks with assets greater than $100 billion.

(2) Our results are also robust to re-estimating the model for the sample of banks that omits those with extreme values of output shares. This leaves a sample of 830 bank holding companies. As seen in column 3 of Table 7, for 2007, scale economies are significantly different from one at the 1 percent level and increase with bank size, from 1.133 for banks with assets under $0.8 billion and 1.267 for banks with assets greater than $100 billion.

(3) We also investigated an alternative specification of outputs, namely, instead of measuring y_3 as loans on the balance sheet plus assets sold without securitization plus securitized loans with servicing retained or with recourse or other credit enhancements, we measured it as loans on the balance sheet. Again, as shown in column 3 of Table 7, our results are robust, with scale economies being significantly different from one at the 1 percent level and increasing with bank size, from 1.136 for banks with assets under $0.8 billion and 1.348 for banks with assets greater than $100 billion.

(4) Many studies of bank cost that impose the assumption of cost minimization on the data estimate the cost function as a frontier in order to characterize efficient production and to measure the degree to which banks depart from efficiency. However, this estimated cost frontier cannot explain the inefficient production decisions of banks whose profit and cost are not close to the frontier. Because the most preferred production system we estimate here is derived from a model of managerial utility maximization that does not impose cost minimization and profit maximization on the data, it can capture production decisions in which managers trade profit and higher cost for reduced risk, which may be a

value-maximizing production decision but not a cost-minimizing decision. Thus, the most preferred profit and cost functions cannot be estimated as a frontier.

Instead, to investigate how production inefficiency might affect our results, we follow Hughes, Lang, Moon, and Pagano (1997), Hughes, Mester, and Moon (2001), and Habib and Ljungqvist (2005) and estimate a *market-value frontier* and identify efficient firms as the quartile of firms that produce market value closest to their highest potential value based on this frontier. To obtain the market value frontier, for those banks that are publicly traded, we estimate a stochastic frontier of the market value of assets as a quadratic function of the book value of assets (adjusted to remove goodwill), which allows the frontier to be nonlinear. The stochastic frontier eliminates the influence of random error (luck) and identifies market value inefficiency as the shortfall of a bank's achieved market value from its highest potential (frontier) market value adjusted for random error. Letting MVA_i denote the market value of the i-th bank's assets and BVA_i, their book value less goodwill, we use maximum likelihood estimation to fit the frontier relationship,

$$MVA_i = \alpha + \beta\,(BVA_i) + \gamma\,(BVA_i)^2 + \varepsilon_i, \tag{15}$$

where $\varepsilon_i \equiv v_i - \mu_i$ is a composite error term comprising v_i, which is normally distributed with zero mean, and μ_i, which is positive and half-normally distributed. The i^{th} bank's market-value inefficiency is measured by the mean of the conditional distribution of μ_i given ε_i, $E(\mu_i|\varepsilon_i)$.

Having estimated market-value efficiency, we then compare the scale economies of the most efficient quartile of banks with those of the full sample (and with those in the least efficient quartile) to determine whether the results differ for market-value-maximizing banks versus utility-maximizing banks.

Table 8 shows the results of these estimations.[16] As shown, scale measures for the market-value efficient banks in the sample are increasing with bank size, confirming the results we obtained for the full sample.

[16] Note that there is a reduction in sample size because not all of the banks are publicly traded; this is especially true for the smaller banks in the sample; hence, we aggregate our two smallest size categories and display results for banks with less than $2 billion in assets. Also note that for 2010, the skewness of the residual is positive and we did not obtain convergence; thus, we could not reject the hypothesis that the OLS estimates represent the frontier or, equivalently, that value is maximized along the utility-maximizing expansion path.

IX. Evidence on Whether Scale Economies Are Driven by Too-Big-To-Fail Considerations

One question is whether the scale economies we find at very large banks are driven by their being too big to fail (TBTF), which might give them a cost advantage over other banks. There is no simple categorization of banks as TBTF. For the purposes of our analysis, consider banks with assets greater than $100 billion as being TBTF, which is consistent with the definitions suggested in Brewer and Jagtiani (2009). Here we present evidence that our scale results are not driven solely by TBTF considerations.

First, as presented above, we find scale economies not only at banks with assets > $100 billion but also at smaller banks, which are too small to be considered TBTF under any reasonable definition.

Second, we re-estimated our cost model for our sample of banks dropping the TBTF banks, i.e, banks with assets > $100 billion, and then calculated what scale economies would be for the TBTF banks, and for banks of other sizes, using this parameterization. The results for 2007 are shown in Table 7, column (5). Here, we once again found significant scale economies that increase with bank size. For banks with assets > $100 billion, the mean scale economies were 1.35 (compared with 1.34 in the baseline model estimated with the full sample of banks discussed above and presented in Table 7, column 1).

Third, to the extent that TBTF enables banks to enjoy lower funding costs because of lower risk premiums on the borrowed funds, it could be that our finding of scale economies at the largest banks in the sample is driven by these lower funding costs and that if these banks faced the same cost of funds as smaller banks, they would not enjoy scale economies. To investigate this possibility, we calculated what the scale economies for the TBTF banks would have been had the cost of the three inputs representing funding costs, namely, w_3 = uninsured deposit rate, w_4 = insured deposit rate, and w_5 = other borrowed funds rate, been the median values for the banks with assets ≤ $100 billion. The results for 2007 are shown in Table 7, column 6. Again, we find significant scale economies that also increase with size. For

banks with assets > $100 billion, the mean scale economies were 1.35 (compared with 1.34 in the baseline model).[17]

Thus, while there may be a funding cost advantage among the largest banks (perhaps because they are considered TBTF), our production model controls for this funding advantage in its computation of scale economies, and there is no evidence that a funding cost advantage influences scale economies.

X. Policy Implications

A current policy question is how regulators should handle TBTF banks. One suggestion has been to impose a size limit on banks to try to prevent them from growing to be too big to fail in the first place (see, e.g., Boyd and Jagannathan). Boyd and Heitz (2012) estimate the benefits and costs of breaking up

[17] Using data from 2001 to 2010 on a sample of 152 highest-level bank holding companies located in 37 countries, Davies and Tracey (forthcoming) estimate a cash-flow cost function (similar to equation 6 above) and as a robustness test, an economic cost function (similar to equation 10 above). They use our strategy of replacing the observed price of borrowed funds with a price that seeks to eliminate the TBTF subsidy. This pseudo price is derived from two Moody's ratings for each bank – one that assumes government assistance in financial distress and one that assumes no assistance. Interestingly, by this measure, 49 percent of banks in their smallest size category (under $100 billion in assets) have a TBTF funding subsidy, while 13 percent of banks in their largest size category (over $2 trillion in asset) do not. The authors find that using the actual observed price of borrowed funds yields evidence of scale economies that disappear when the pseudo price is used. The authors assume that this difference in measured scale economies is due to a TBTF subsidy. However, there is a critical difference between their methodology and ours. We measure scale economies by substituting the pseudo prices into the fitted measure of scale economies, which we derived from our estimated cost function estimated using the actual input prices, outputs, and control variables that the banks faced. Davies and Tracey do not use the original estimated cost function. Instead, they re-estimate the cost function and share equations using the pseudo price along with the actual observed data on the other variables, including total cost, cost shares, other input prices, outputs, and control variables. Thus, the total costs and cost shares used in the re-estimation do not match the prices used. The model assumes that banks minimize cost with respect to the pseudo prices, but since these pseudo prices do not give rise to the observed cost and cost shares, the resulting re-estimated technology is difficult to interpret.

Another difficulty is that Davies and Tracey measure the prices of labor and physical capital as the input expenditure divided by total assets, rather than the standard in the literature, which is input expenditure divided by the amount of the input. Thus, the Davies and Tracey measures make these two prices a function of how the input expense varies with assets, which confounds measures of scale economies. For example, if there are scale economies, it is likely that the input expenditure will increase less than proportionately with assets so that the input price will decrease as total assets increase. The bias created in estimated scale economies by an input price that is proxied by a measure that is a function of these economies is not clear. Indeed, when the authors define the price of physical capital as its expense divided by fixed assets rather than total assets, the magnitude of measured of scale economies is significantly higher over all bank sizes.

A third reason it is difficult to interpret the Davies and Tracey results is that their cost functions do not account for size-related differences in endogenous risk-taking: if larger, better diversified institutions seek higher expected return by taking more risk, then the extra cost associated with larger scale and higher risk increases the estimated cost elasticity and obscures some or all of the underlying technological scale economies that are driven by better diversification, larger networks, and so on. Thus, it is not possible from their specification to distinguish TBTF effects from risk-taking effects.

systemically important banks into smaller institutions and find that the benefits far outweigh the costs. They compute costs from the estimates of scale economies in Hughes, Mester, and Moon (2001) and in Wheelock and Wilson (2012). As discussed in Mester (2010), there would be several consequences of such a size limit, some of which might be unintended. Indeed, should scale economies be as strong as suggested in our results, banks would be motivated to try to circumvent such a limit. On the face of it, our estimates of scale economies suggest that such a size limit, by limiting the attainment of scale economies, would be quite costly. However, this is actually a more difficult question than it might seem. Typically, when researchers perform such calculations, they vary the scale of operations alone. And the estimates of scale economies essentially do that as well, by keeping product mix locally constant as scale is expanded. However, not only is the scale of operations different for large and small banks, but the output mix also differs considerably, e.g., large banks have a considerably higher share of off-balance-sheet output. This variation in output mix turns out to be important when evaluating the potential cost impact of a size limit on banks.

In particular, we ask, what would be the change in cost if we broke up the 17 banks with assets greater than $100 billion in 2007 into banks with assets of $100 billion? We will decompose the change in costs into two parts: the *scale effect*, which calculates the change in costs ignoring any change in output mix, and the *mix effect*, which calculates the change in costs from the change in the output mix that occurs when scale changes. Let YL = total assets of a bank with assets > $100 billion (a large bank), YS = the size limit we are imposing (here, $100 billion), HL represent the output mix (output shares) of the large bank, and HS represent the output mix of a $100 billion bank. Then based on our estimated cost function, we can compute the ratio of the estimated cost of a set of n $100 billion banks, $nC(YS,HS)$, to the estimated cost of a large bank, $C(YL,HL)$, where $n = YL/YS$:

$$(16)\quad \frac{nC(YS,HS)}{C(YL,HL)} = \frac{\frac{YL}{YS}C(YS,HS)}{C(YL,HL)} = \frac{\frac{YL}{YS}C(YS,HL)}{C(YL,HL)} \times \frac{\frac{YL}{YS}C(YS,HS)}{\frac{YL}{YS}C(YS,HL)}$$

$$= \quad \text{scale effect} \times \text{mix effect.}$$

Our estimated scale economies can be used to calculate the *scale effect*:

$$(17)\quad \text{scale effect} = \frac{\frac{YL}{YS}C(YS,HL)}{C(YL,HL)} = \frac{YS}{YL}\left\{\left[\left(\frac{YL}{YS}-1\right)\frac{1}{scale}\right]+1\right\}.$$

To calculate the *mix effect*, we need to know what product mix a bank with $100 billion in assets would produce in order to calculate *C(YS,HS)*. Because there is no bank in the sample that has $100 billion in assets (and even if there were, it would not necessarily be representative), we calculate the *mix effect* in two different ways. First, we approximate *C(YS,HS)* by the mean *C(·)* of the 10 firms in the size category $60 billion to $140 billion in assets, which spans $100 billion. Second, we approximate *C(YS,HS)* by evaluating the estimated cost function *C(·)* at the median output shares and other non-output variables of banks in the asset size category of $50 billion to $100 billion, where we adjust y_3 so that the output levels sum to $100 billion.

Based on our estimates for 2007, the sum of estimated costs for the 17 banks in the largest size category (with assets over $100 billion) is $410 billion. These banks hold a total of $9.1 trillion in assets. Thus, expressed as a percentage, the average cost per dollar of assets is 4.5 percent. If these 17 banks were broken into smaller banks with $100 billion in assets but with no change in their output mix, costs would increase from $410 billion to $1.48 trillion. This scale effect means the average cost per dollar of assets would increase from 4.5 percent to 16.3 percent, an increase of 11.8 percentage points. This increase in average cost per dollar of assets suggests that restricting the size of financial institutions in a manner consistent with this exercise could seriously affect their competitiveness in global financial markets if institutions in other countries were not similarly constrained in size.

On the other hand, banks that are forced to downsize might also change their product mix over the longer run to one more consistent with a smaller scale of operations. We can compute the mix effect in equation (16), which compares the total cost of the 17 largest institutions scaled back in size and producing their original product mix with the total cost of the 17 banks reduced in size but producing the product mix of smaller institutions. Our estimated mix effect suggests that adjusting their output shares to those appropriate to their smaller size would lower costs from the projected $1.48 trillion for the large-bank product mix to $260 billion to $360 billion, which is 2.9 percent to 3.9 percent of assets. Adding the scale and mix effects, our estimates suggest that the total impact of breaking up the banks into smaller institutions producing the financial products and services of smaller institutions, all else equal, would be a cost savings of $54-$151 billion.

These calculations are only intended to be suggestive of one issue that must be considered in calculating the cost impact of imposing a size limit, namely, the effect on costs not only of a change in the scale of operations but also in the mix of outputs banks would choose to produce. In particular, the cost-savings estimates do not include a value for the output mix produced by larger banks. Moreover, these calculations ignore other consequences of such a policy, and they are only rough estimates and are dependent on the method of calculation. To see this, consider next a simpler method of comparing the total cost of the largest financial institutions with that of smaller institutions. The 17 largest institutions in our 2007 sample have $9.1 trillion in assets and an estimated total cost of $410 billion. The sum of the estimated costs for the 12 banks in the second largest size category with assets between $50 billion and $100 billion is $33 billion, and these banks hold a total of $778 billion in assets. A simple back-of-the-envelope calculation indicates that redistributing the $9.1 trillion of assets in the largest size category to the next largest size category would result in costs of $379 billion (= (9.1 trillion / 778 billion) × $33 billion), which, compared to the $410 billion cost of banks in the largest category, suggests a cost savings of about $30 billion. Again, such a calculation assumes a change in output mix to that of banks in the second largest size category (and it also assumes that the values of the other variables in the cost function, in particular, input prices, would be consistent with those at banks in the second largest size category).

These various cost comparisons suggest that the mix of financial products and services being offered by different size institutions is an important consideration when evaluating a policy of breaking up the largest financial institutions into ones of smaller size. Our results suggest that the product mix offered by the largest institutions cannot be produced economically by smaller institutions that must compete with larger banks in global markets. Nevertheless, should a break-up policy be enacted, the scale economies that exist will give the smaller banks an economic incentive to seek ways around the policy, perhaps using alternative organizational structures. The general equilibrium consequences in terms of cost, competitiveness, and financial stability of a policy to simply break up the large banks is not at all straightforward.

XI. Conclusions

We find evidence of large scale economies at smaller banks and even larger economies at large banks – economies consistent with the standard textbook arguments – when the production model endogenizes managers' choice of risk vs. expected return. The standard minimum cost function, even one that controls for equity capital, is not able to capture these scale economies.

Our results indicate that these measured scale economies do not result from the cost advantages large banks may derive from too-big-to-fail considerations. Instead, they follow from technological advantages, such as diversification and the spreading of information costs and other costs that do not increase proportionately with size. Significant scale economies in banking suggest that technological factors, as well as TBTF cost advantages, appear to be an important driver of banks' increasing size. While we do not know if the benefits of large size outweigh the potential costs in terms of systemic risk that large scale may impose on the financial system, our results suggest that strict size limits to control such costs will not likely be effective, since they work against market forces. Our results also indicate that one should consider both scale and product mix when evaluating such a policy.

Appendix

Empirical model and estimation: The Almost Ideal Production System[18]

The managers' most preferred (MP) production model comprises the profit share equation (3a), the input share equations (3b), and the first-order condition for the optimal level of equity capital, *k*, which is a conditioning argument in the share equations. The profit and input demand functions are shares expressed as shares of total revenue, $\boldsymbol{p} \cdot \boldsymbol{y} + m$, and sum to one. They are derived by applying Shephard's Lemma to the managerial expenditure function, which is dual to the utility maximization problem (1a-1c). Thus, the model to be estimated is:

$$\frac{p_\pi \pi}{\mathbf{p}\cdot\mathbf{y}+m} = \frac{\partial \ln \boldsymbol{P}}{\partial \ln p_\pi} + \mu[ln(\mathbf{p}\cdot\mathbf{y}+m) - \ln \boldsymbol{P}] \tag{A1.1}$$

$$\frac{w_i x_i}{\mathbf{p}\cdot\mathbf{y}+m} = \frac{\partial \ln \boldsymbol{P}}{\partial \ln w_i} + \nu_i[ln(\mathbf{p}\cdot\mathbf{y}+m) - \ln \boldsymbol{P}] \;\forall i \tag{A1.2}$$

$$\frac{\partial V(\cdot)}{\partial k} = \frac{\partial V(\cdot)}{\partial \ln k}\frac{\partial \ln k}{\partial k} = 0, \tag{A1.3}$$

where $ln\ \mathbf{P} = A_0 + \Sigma_i A_i\ ln\ y_i + (½)\Sigma_i \Sigma_j S_{ij}\ ln\ y_i\ ln\ y_j + \Sigma_i B_i\ ln\ w_i + (½)\Sigma_i \Sigma_j G_{ij}\ ln\ w_i\ ln\ w_j$

$+ \Sigma_i \Sigma_j D_i\ ln\ y_i\ ln\ w_j + (½)\Sigma_i \Sigma_j R_{ij}\ ln\ z_i\ ln\ z_j + \Sigma_i \Sigma_j H_{ij}\ ln\ z_i\ ln\ y_j + \Sigma_i \Sigma_j T_{ij}\ ln\ z_i\ ln\ w_j$;

$\boldsymbol{z} = (k, n, \boldsymbol{p}, p_\pi)$, and

$$V(\cdot) = \frac{\ln(\mathbf{p}\cdot\mathbf{y}+m) - \ln \boldsymbol{P}}{\beta_0\left(\prod_i y_i^{\beta_i}\right)\left(\prod_j w_i^{\nu_i}\right)p_\pi^{\mu}k^{\kappa}}. \tag{A1.4}$$

To save on degrees of freedom, in our estimation we replace the vector of output prices, $\boldsymbol{p}$, with the weighted-average output price, $\tilde{p} = \Sigma_i p_i (y_i/\Sigma_j y_j)$. The risk-free rate, *r*, is the same for all bank

[18] The exposition in this appendix is adapted from Hughes, Lang, Mester, and Moon (2000).

holding companies, so the coefficients on terms involving *r* are not estimated. Written out, the equations to be estimated are:

$$\begin{aligned}\frac{p_\pi \pi}{\mathbf{p}\cdot\mathbf{y}+m} &= F_4 + R_{44}\ln p_\pi + R_{34}\ln \tilde{p} + \sum_j H_{4j}\ln y_j + \sum_s T_{4s}\ln w_s \\ &\quad + R_{24}\ln n + R_{14}\ln k + \mu[ln(\mathbf{p}\cdot\mathbf{y}+m) - \ln \boldsymbol{P}] \\ &= F_4 + \sum_j H_{4j}\ln y_j + \sum_s T_{4s}\ln w_s \\ &\quad + \sum_i R_{i4}\ln z_i + \mu[ln(\mathbf{p}\cdot\mathbf{y}+m) - \ln \boldsymbol{P}]\end{aligned} \tag{A1.1′}$$

$$\begin{aligned}\frac{w_i x_i}{\mathbf{p}\cdot\mathbf{y}+m} &= B_i + \sum_s G_{ij}\ln w_s + T_{3i}\ln \tilde{p} + \sum_j D_{ji}\ln y_j + T_{4i}\ln p_\pi \\ &\quad + T_{3i}\ln n + T_{1i}\ln k + \nu_i[ln(\mathbf{p}\cdot\mathbf{y}+m) - \ln \boldsymbol{P}] \\ &= B_i + \sum_j D_{ji}\ln y_j + \sum_s G_{ij}\ln w_s \\ &\quad + \sum_s T_{si}\ln z_s + \nu_i[ln(\mathbf{p}\cdot\mathbf{y}+m) - \ln \boldsymbol{P}]\end{aligned} \tag{A1.2′}$$

$$\begin{aligned}&F_1 + R_{11}\ln k + R_{13}\ln \tilde{p} + \sum_j H_{1j}\ln y_j + \sum_s T_{1s}\ln w_s + R_{14}\ln p_\pi \\ &\quad + R_{12}\ln k + \kappa[ln(\mathbf{p}\cdot\mathbf{y}+m) - \ln \boldsymbol{P}] = 0 \\ &\Rightarrow F_1 + \sum_j H_{1j}\ln y_j + \sum_s T_{1s}\ln w_s + \sum_j R_{1j}\ln z_j + \kappa[ln(\mathbf{p}\cdot\mathbf{y}+m) - \ln \boldsymbol{P}] = 0\end{aligned} \tag{A1.3′}$$

where

$$
\begin{aligned}
\ln \boldsymbol{P} = {} & A_0 + F_3 \ln \tilde{p} + \sum_i A_i \ln y_i + \sum_j B_j \ln w_j \\
& + F_4 \ln p_\pi + F_2 \ln n + F_1 \ln k + \frac{1}{2} R_{33} (\ln \tilde{p})^2 + \frac{1}{2} \sum_i \sum_j S_{ij} \ln y_i \ln y_j \\
& + \frac{1}{2} \sum_s \sum_t G_{st} \ln w_s \ln w_t + \frac{1}{2} R_{44} (\ln p_\pi)^2 \\
& + \frac{1}{2} R_{22} (\ln n)^2 + \frac{1}{2} R_{11} (\ln k)^2 \\
& + \sum_j H_{3j} \ln \tilde{p} \ln y_j + \sum_s T_{3s} \ln \tilde{p} \ln w_s + R_{34} \ln \tilde{p} \ln p_\pi \\
& + R_{23} \ln \tilde{p} \ln n + R_{13} \ln \tilde{p} \ln k \\
& + \sum_i \sum_s D_{is} \ln y_i \ln w_s + \sum_j H_{4j} \ln y_j \ln p_\pi \\
& + \sum_j H_{2j} \ln y_j \ln n + \sum_j H_{1j} \ln y_j \ln k \\
& + \sum_s T_{4s} \ln w_s \ln p_\pi \\
& + \sum_s T_{2s} \ln w_s \ln n + \sum_s T_{1s} \ln w_s \ln k \\
& + R_{24} \ln p_\pi \ln n + R_{14} \ln p_\pi \ln k \\
& + R_{12} \ln n \ln k
\end{aligned}
$$

$$
\begin{aligned}
= {} & A_0 + \sum_i A_i \ln y_i + \sum_s B_s \ln w_s + \sum_j F_j \ln z_j \\
& + \frac{1}{2} \sum_i \sum_j S_{ij} \ln y_i \ln y_j + \frac{1}{2} \sum_s \sum_t G_{st} \ln w_s \ln w_t + \frac{1}{2} \sum_i \sum_j R_{ij} \ln z_i \ln z_j \\
& + \sum_i \sum_s D_{is} \ln y_i \ln w_s + \sum_i \sum_j H_{ij} \ln z_i \ln y_j + \sum_i \sum_s T_{is} \ln z_i \ln w_s
\end{aligned}
$$

and $p_\pi = 1/(1-t)$.

We impose several conditions on the parameters of the model. *Symmetry* requires that[19]

(S1) $S_{ij} = S_{ji} \ \forall$ i,j,

(S2) $T_{4s} = T_{s4} \ \forall$ s, and

(S3) $G_{si} = G_{is} \ \forall$ s,i.

[19] (S1) must be imposed in the estimation of the share equations, since the constituent coefficients cannot be separately identified. However, (S2) and (S3) involve coefficients of prices that are used by Shephard's Lemma to obtain the share equations. Consequently, they appear in separate share equations and are, thus, identifiable. It is a judgment call as to whether one imposes these symmetry conditions. We impose them in our estimation.

The input and profit revenue share equations sum to one, which implies the following *adding-up* conditions:

(A1) $\Sigma_i B_i + F_4 = 1$,

(A2) $\Sigma_i G_{si} + T_{4s} = 0, \forall s$,

(A3) $\Sigma_i T_{3i} + R_{34} = 0$,

(A4) $\Sigma_i D_{ji} + H_{4j} = 0 \forall j$,

(A5) $\Sigma_i T_{4i} + R_{44} = 0$,

(A6) $\Sigma_i T_{1i} + R_{14} = 0$,

(A7) $\Sigma_i T_{2i} + R_{24} = 0$, and

(A8) $\Sigma v_j + \mu = 0$.

The input and profit share equations are *homogeneous of degree zero* in ($\boldsymbol{w}$, $\tilde{p}$, r, p_π). The only homogeneity condition imposed in the estimation is:

(H1) $\Sigma v_j + \mu = 0$,

which is equivalent to the adding-up condition (A8). The other homogeneity conditions contain coefficients on variables involving the risk-free rate, *r*. These coefficients are not estimated, since *r* does not vary across banks, but the homogeneity conditions can be used to recover these coefficients.

To summarize: in estimating the model, we imposed (S1), (A1)-(A7), and (A8) ≡ (H1).

We estimated the model using nonlinear two-stage least squares, a generalized method of moments. Starting values were obtained by setting the constant terms, B_i, in the input share equations at the average value of the input share across banks in the sample, the constant term, *F4*, in the profit share equation at the average value of the profit share across banks in the sample, and all other parameters in the input share, profit share, and equity capital demand equation equal to 0.

Bibliography

Berger, A.N., DeYoung, R., Flannery, M.J., Lee, D. and Öztekin, Ö. (2008). How do large banking organizations manage their capital ratios? *Journal of Financial Services Research* **34**, 123-149.

Berger, A.N. and Mester, L.J. (1997). Inside the black box: what explains differences in the efficiencies of financial institutions, *Journal of Banking and Finance* **21**, 895-947.

Bossone, B. and Lee, J.-K. (2004). In finance, size matters: the 'systemic scale economies' hypothesis, IMF Staff Papers, 51:1.

Boyd, J.H. and Heitz, A. (2012). The social costs and benefits of too-big-to-fail banks: a "bounding" exercise." Working Paper. University of Minnesota.

Boyd, J.H. and Jagannathan, R. (2009). Avoiding the next crisis. *The Economists' Voice*, **6**.

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

Brewer, E. and Jagtiani, J. (2009). How much did banks pay to become too-big-to-fail and to become systemically important? Federal Reserve Bank of Philadelphia Working Paper No. 09-34.

Davies, R. and Tracey, B. (forthcoming). Too big to be efficient? The impact of too-big-to-fail factors on scale economies for banks. *Journal of Money, Credit, and Banking*.

Deaton, A. and Muellbauer, J. (1980). An almost ideal demand system, *American Economic Review* **70**, 312-326.

Demsetz, R.S. and Strahan, P.E. (1997). Diversification, size, and risk at bank holding companies, *Journal of Money, Credit, and Banking* **29**, 300-313.

Erel, I., Nadauld, T.D., and Stulz, R.M. (2011). Why did U.S. banks invest in highly-rated securitization tranches? NBER Working Paper No. 17269.

Feng, G. and A. Serletis. (2010). Efficiency, technical change, and returns to scale in large US banks: Panel data evidence from an output distance function satisfying theoretical regularity, *Journal of Banking and Finance* **34**, 127-138.

Greenspan, Alan. (2010). The crisis, Brookings Papers on Economic Activity Spring 2010, 201-246.

Habib, M. and A. Ljungqvist. (2005). Firm value and managerial incentives: A stochastic frontier approach, *Journal of Business* **78**, 2053-2094.

Hughes, J.P. (1999). Incorporating risk into the analysis of production, Presidential Address at the Atlantic Economic Society, *Atlantic Economic Journal* **27**, 1–23.

Hughes, J.P. (forthcoming). The elusive scale economies of the largest banks and their implications for global competitiveness, in *The Role of Central Banks in Financial Stability: How Has It Changed?* Proceedings of the Fourteenth Annual International Banking Conference, Federal Reserve Bank of Chicago and the European Central Bank.

Hughes, J.P., Lang, W., Mester, L.J. and Moon C.-G. (1996). Efficient banking under interstate branching, *Journal of Money, Credit, and Banking* **28**, 1045-1071.

Hughes, J.P., Lang, W., Mester, L.J. and Moon C.-G. (2000). Recovering risky technologies using the almost ideal demand system: an application to U.S. banking, *Journal of Financial Services Research* **18**, 5-27.

Hughes, J.P., Lang, W., Moon C.-G. and Pagano, M. (1997 and 2004). Measuring the efficiency of capital allocation in commercial banking, Federal Reserve Bank of Philadelphia Working Paper 98-2, (revised as Working Paper 2004-1, Rutgers University Economics Department).

Hughes, J.P. and Mester, L.J. (1998). Bank capitalization and cost: evidence of scale economies in risk management and signaling, *Review of Economics and Statistics* **80**, 314-325.

Hughes, J.P. and Mester, L.J. (2010). Efficiency in banking: theory, practice, and evidence, Chapter 19 in *The Oxford Handbook of Banking*, edited by A.N. Berger, P. Molyneux, and J. Wilson, Oxford University Press.

Hughes, J.P., Mester, L.J. and Moon C.-G. (2001). Are scale economies in banking elusive or illusive? Evidence obtained by incorporating capital structure and risk-taking into models of bank production, *Journal of Banking and Finance* **25**, 2169-2208.

Keeley, Michael C. (1990). Deposit insurance, risk, and market power in banking, *American Economic Review* **80**, 1183-1200.

Marcus, A.J. (1984). Deregulation and bank financial policy, *Journal of Banking and Finance* **8**, 557-565.

Mester, L.J. (1992). Traditional and nontraditional banking: An information-theoretic approach, *Journal of Banking and Finance* **16**, 545-566.

Mester, L.J. (2010). Scale economies in banking and financial regulatory reform, *The Region*, Federal Reserve Bank of Minneapolis, September 2010, 10-13.

Tufano, Peter (1996). Who manages risk? An empirical examination of risk management practices in the gold mining industry, *Journal of Finance* **50**, 1097-1137.

Wheelock, D. and Wilson, P. (2012). Do large banks have lower costs? New estimates of returns to scale for US banks, *Journal of Money, Credit, and Banking* **44**, 171-99.

Figure 1

A smaller bank's investment strategies are depicted on the risk-return frontier labeled *I* and the larger bank's strategies on frontier *II*. The improved trade-off along frontier *II* results from the better diversification of the larger bank. Point *A* represents production of a smaller, less diversified output, say, some quantity of loans with a particular probability distribution of default that reflects the contractual interest rate charged and the resources allocated to risk assessment and monitoring. Point *B* represents a larger quantity of loans with the same contractual interest rate but better diversification and, hence, an improved probability distribution of default and lower overall risk. The better diversification allows the costs of risk management to increase less than proportionately with the loan volume while maintaining an improved probability distribution of default. Thus, the response of cost to the increase in output from point *A* to point *B* reflects scale economies. And the expected return at *B* exceeds that at *A*.

On the other hand, suppose the bank responds to the better diversification of the larger output by adopting a more risky investment strategy for an enhanced expected return. Better diversification does not offset the increased cost occasioned by the additional default risk. Point *D* designates this strategy. The increased inherent default risk due to the higher contractual interest rate results in costs of risk management that increase more than proportionately with the loan volume (from *A* to *D*), and production appears to exhibit the counter-intuitive *scale diseconomies* found by empirical studies of banking cost that fail to account for endogenous risk-taking.

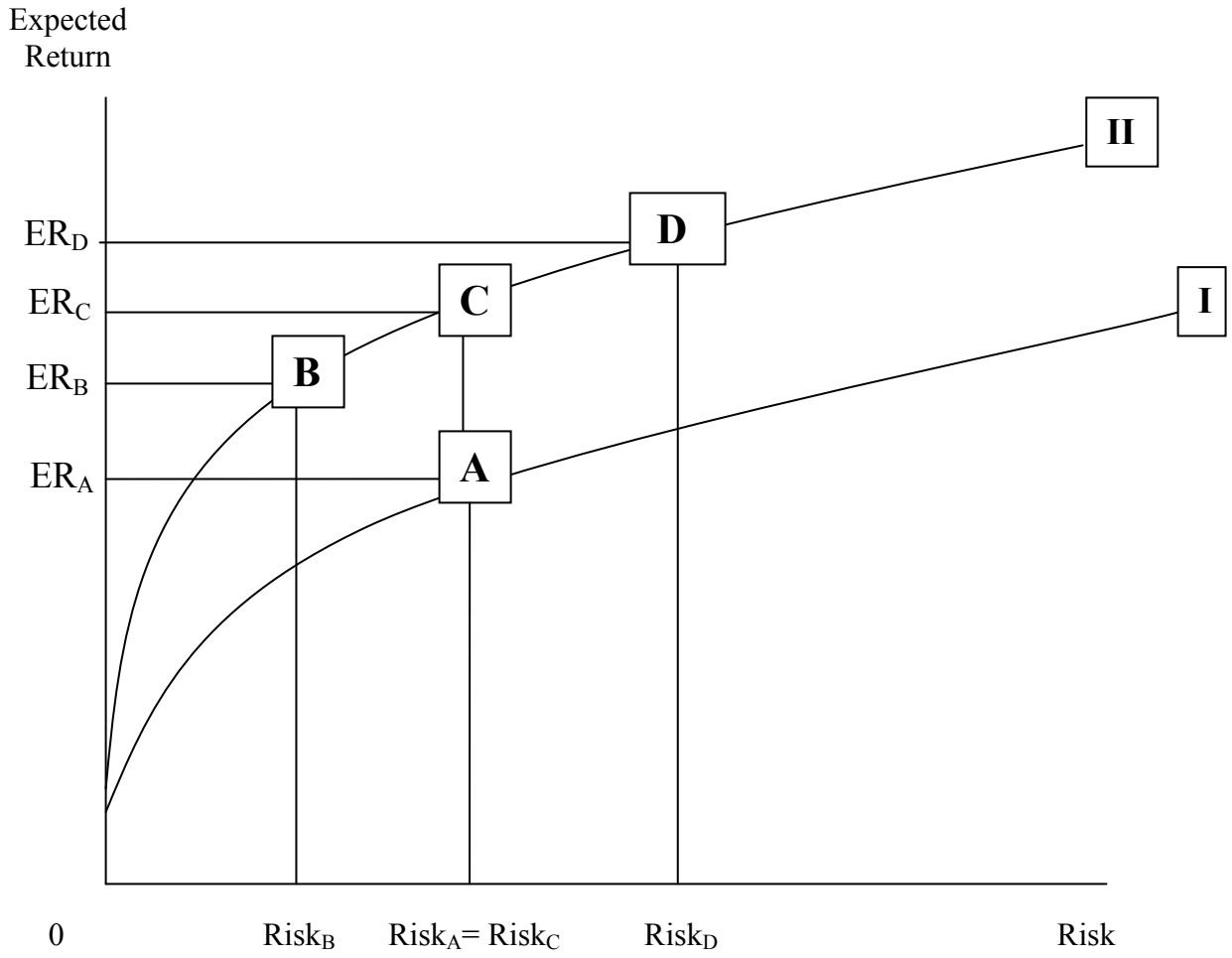

Figure 2

The investment strategies in Figure 1 are illustrated for the case of the larger output along frontier *II*. The production technology for a given quantity of loans is represented by the isoquant. The mix of debt and equity used to fund the loans is ignored. The diagram shows the quantity of physical capital and labor used in the process of credit evaluation and loan monitoring. Point *C* shows the least costly way to produce the particular quantity of loans with the risk exposure associated with the investment strategy *C* in Figure 1. If a bank adopted the less risky strategy, *B*, it might use less labor in credit evaluation and monitoring: point *B*. This is a less costly method of producing the *same quantity of loans*. Thus, the isoquant for this quantity of loans that passes through point *C* captures one investment strategy only. There would be another isoquant passing through point *B* for the same quantity of loans produced with a lower risk strategy. On the other hand, if a bank adopted the more risky strategy, *D*, it would use more labor. The corresponding point *D* is a more costly method than point *C*. Thus, the cost of producing this particular quantity of loans depends on a bank's choice of risk exposure and its expected return. We shall refer to this characterization of cost as *risk-return-driven cost*.

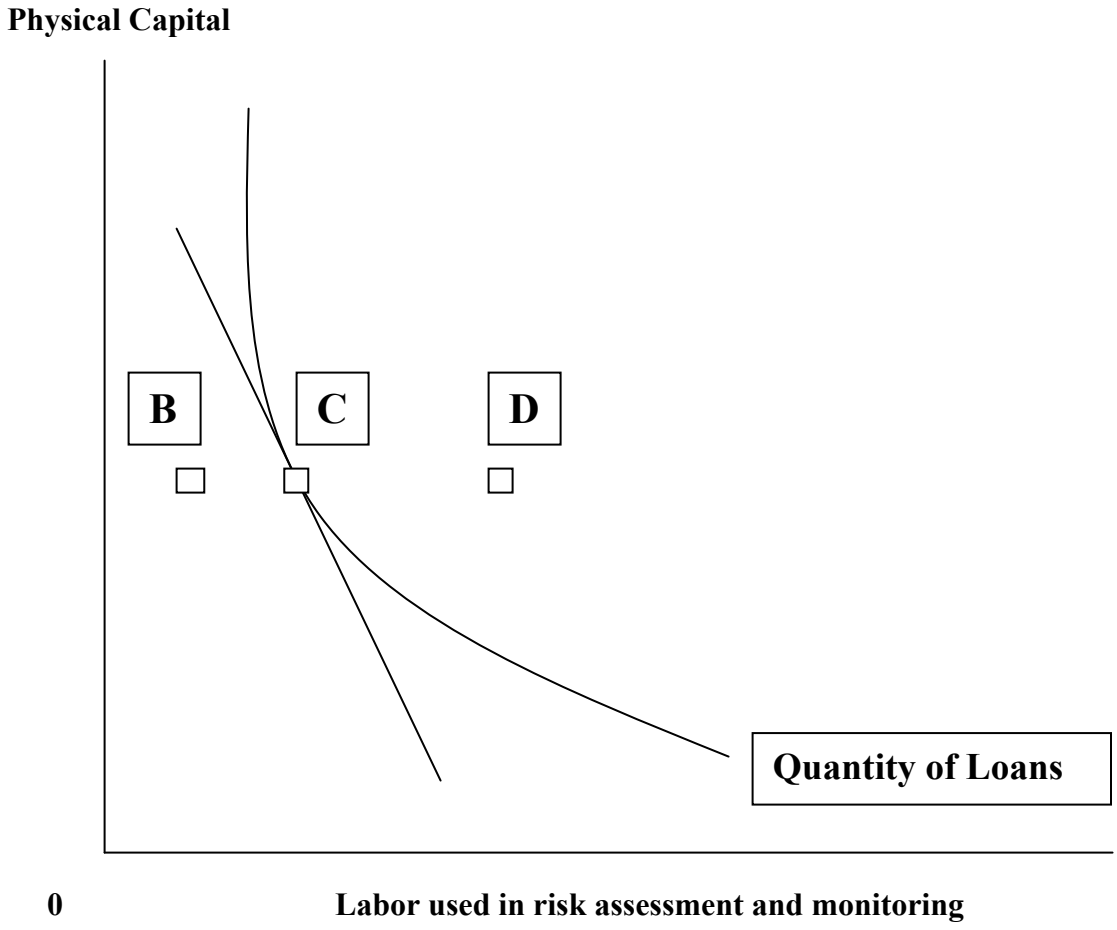

Table 1. Summary Statistics: Full Sample

The data, obtained from the Y9-C Call Reports filed quarterly with regulators, include 842 top-tier U.S. bank holding companies in 2007. A top-tier company is not owned by another company.

	2007 (no. obs. = 842)		2003 (no. obs. = 1855)		2010 (no. obs. = 856)	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Total Assets in $1000s	13,692,833	941,224	4,773,835	385,638	13,654,789	905,371
Total Revenue in $1000s	1,024,870	70,026	324,661	24,144	804,550	51,150
Financial Performance						
Equity Capital/Assets	0.102	0.097	0.100	0.096	0.107	0.102
Nonperforming Assets/Assets	0.022	0.016	0.016	0.013	0.058	0.044
Profit/Revenue	0.317	0.315	0.409	0.407	0.421	0.431
Profit/Assets	0.024	0.023	0.026	0.025	0.024	0.024
Asset Allocation						
Liquid Assets: y_1/Assets	0.044	0.033	0.061	0.048	0.075	0.059
Securities: y_2/Assets	0.174	0.159	0.237	0.217	0.200	0.184
Loans, Securitiz., Serv.: y3/Assets	0.730	0.741	0.656	0.670	0.657	0.667
Trading, Other Assets: y_4/Assets	0.051	0.041	0.035	0.030	0.052	0.044
Off-Balance-Sheet Items: y_5/Assets	0.060	0.034	0.035	0.018	0.037	0.021
Input Utilization						
Labor (FTEs): x_1/assets	0.00027	0.00027	0.00034	0.00033	0.00025	0.00024
Physical Capital: x_2/Assets	0.019	0.018	0.019	0.018	0.019	0.018
Uninsured Deposits: x_3/Assets	0.149	0.134	0.126	0.113	0.150	0.135
Insured Deposits: x_4/Assets	0.610	0.625	0.668	0.685	0.650	0.664
Other Borrowed Funds: x_5/assets	0.122	0.104	0.097	0.078	0.085	0.069
Prices						
Average Interest Rate on Assets	0.062	0.062	0.053	0.052	0.047	0.047
Wage Rate: w_1 in $1000s	63.322	59.391	57.983	54.581	63.850	59.816
Price of Physical Capital: w_2	0.288	0.215	0.287	0.226	0.289	0.210
Uninsured Deposit Rate: w_3	0.048	0.047	0.027	0.027	0.020	0.019
Insured Deposit Rate: w_4	0.028	0.027	0.014	0.014	0.010	0.010
Other Borrowed Funds rate: w_5	0.054	0.048	0.037	0.033	0.042	0.036
Tax Rate	0.421	0.420	0.414	0.415	0.413	0.419
1/(1–Tax Rate)	1.730	1.724	1.711	1.709	1.708	1.721

Table 2. Summary Statistics: Asset Allocation by Size Groups

The data on top-tier holding companies were obtained from the Y9-C Call Reports filed quarterly with regulators. A top-tier company is not owned by another company. Banks in the largest size category, with assets exceeding $100 billion, are often perceived as being too big to fail (Brewer and Jagtiani, 2009).

Total Assets < $0.8 billion	**2007 (no. obs. = 328)**		**2003 (no. obs. = 1403)**		**2010 (no. obs. = 364)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Total Assets in $1000s	571582.29	589526.00	354397.55	305631.82	575945.700	578292.34
Liquid Assets: y1/Assets	0.045	0.036	0.063	0.052	0.077	0.064
Securities: y2/Assets	0.182	0.163	0.234	0.212	0.187	0.171
Loans, Securitiz., Serv.: y3/Assets	0.727	0.742	0.660	0.674	0.674	0.682
Trading, Other Assets: y4/Assets	0.040	0.036	0.030	0.028	0.043	0.040
Off-Balance-Sheet Items: y5/Assets	0.032	0.025	0.022	0.014	0.025	0.016

Total Assets $0.8 bill − $2 bill	**2007 (no. obs. = 299)**		**2003 (no. obs. = 252)**		**2010 (no. obs. = 286)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Total Assets in $1000s	1195048.46	1119251.00	1203192.26	1132334.32	1201903.01	1112394.36
Liquid Assets: y1/Assets	0.043	0.033	0.054	0.044	0.071	0.057
Securities: y2/Assets	0.166	0.153	0.236	0.217	0.209	0.200
Loans, Securitiz., Serv.: y3/Assets	0.741	0.753	0.662	0.670	0.659	0.660
Trading, Other Assets: y4/Assets	0.045	0.040	0.038	0.034	0.045	0.041
Off-Balance-Sheet Items: y5/Assets	0.041	0.032	0.035	0.028	0.028	0.021

Total Assets $2 bill − $10 bill	**2007 (no. obs. = 155)**		**2003 (no. obs. = 131)**		**2010 (no. obs. = 148)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Total Assets in $1000s	4091767.77	3350126.00	4150822.68	3330456.24	4080141.44	3247250.81
Liquid Assets: y1/Assets	0.038	0.029	0.044	0.035	0.075	0.054
Loans, Securitiz., Serv.: y3/Assets	0.177	0.164	0.269	0.249	0.207	0.191
Loans: y3/Assets	0.721	0.726	0.631	0.652	0.639	0.657
Trading, Other Assets: y4/Assets	0.060	0.056	0.049	0.046	0.067	0.059
Off-Balance-Sheet Items: y5/Assets	0.055	0.046	0.052	0.039	0.035	0.028

Total Assets $10 bill − $50 bill	**2007 (no. obs. = 31)**		**2003 (no. obs. = 43)**		**2010 (no. obs. = 33)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Total Assets in $1000s	16562010.32	13871556.00	23548740.60	20915197.21	19176056.99	16671150.70
Liquid Assets: y1/Assets	0.041	0.029	0.053	0.034	0.067	0.052
Securities: y2/Assets	0.187	0.170	0.257	0.253	0.241	0.216
Loans, Securitiz., Serv.: y3/Assets	0.717	0.711	0.607	0.632	0.610	0.635
Trading, Other Assets: y4/Assets	0.078	0.069	0.076	0.063	0.077	0.075
Off-Balance-Sheet Items: y5/Assets	0.089	0.073	0.193	0.066	0.065	0.051

Total Assets $50 bill − $100 bill	**2007 (no. obs. = 12)**		**2003 (no. obs. = 10)**		**2010 (no. obs. = 10)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Total Assets in $1000s	64794778.92	61460925.50	70559690.46	63144003.68	66617190.34	65700576.60
Liquid Assets: y1/Assets	0.079	0.044	0.085	0.039	0.088	0.036
Securities: y2/Assets	0.139	0.128	0.204	0.200	0.161	0.138
Loans, Securitiz., Serv.: y3/Assets	0.681	0.741	0.597	0.665	0.637	0.672
Trading, Other Assets: y4/Assets	0.121	0.112	0.109	0.097	0.120	0.110
Off-Balance-Sheet Items: y5/Assets	0.441	0.221	0.519	0.234	0.256	0.146

Total Assets > $100 bill	**2007 (no. obs. = 17)**		**2003 (no. obs. = 16)**		**2010 (no. obs. = 15)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Total Assets in $1000s	532904914.0	179573933.0	362068066.0	178293125.0	615484563.0	187776084.00
Liquid Assets: y1/Assets	0.087	0.039	0.093	0.086	0.095	0.052
Securities: y2/Assets	0.153	0.131	0.176	0.165	0.211	0.159
Loans, Securitiz., Serv.: y3/Assets	0.726	0.689	0.557	0.537	0.529	0.638
Trading, Other Assets: y4/Assets	0.181	0.154	0.168	0.136	0.168	0.142
Off-Balance-Sheet Items: y5/Assets	0.657	0.250	0.350	0.248	0.309	0.176

Table 3. Summary Statistics: Input Utilization by Size Groups

The data on top-tier holding companies were obtained from the Y9-C Call Reports filed quarterly with regulators. A top-tier company is not owned by another company. Banks in the largest size category, with assets exceeding $100 billion, are often perceived as being too big to fail (Brewer and Jagtiani, 2009).

Total Assets < $0.8 billion	**2007 (no. obs. = 328)**		**2003 (no. obs. = 1403)**		**2010 (no. obs. = 364)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Labor (FTEs): x_1/assets	0.00030	0.00029	0.00035	0.00035	0.00027	0.00026
Physical Capital: x_2/Assets	0.0213	0.0198	0.0199	0.0188	0.0211	0.0199
Uninsured Deposits: x_3/Assets	0.1558	0.1395	0.1289	0.1178	0.1612	0.1442
Insured Deposits: x_4/Assets	0.6344	0.6465	0.6828	0.6941	0.6580	0.6680
Other Borrowed Funds: x_5/assets	0.0969	0.0885	0.0822	0.0687	0.0737	0.0637

Total Assets $0.8 bill − $2 bill	**2007 (no. obs. = 299)**		**2003 (no. obs. = 252)**		**2010 (no. obs. = 286)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Labor (FTEs): x_1/assets	0.00027	0.00026	0.00032	0.00031	0.00024	0.00024
Physical Capital: x_2/Assets	0.0198	0.0195	0.0180	0.0174	0.0201	0.0183
Uninsured Deposits: x_3/Assets	0.1508	0.1380	0.1252	0.1090	0.1580	0.1411
Insured Deposits: x_4/Assets	0.6185	0.6240	0.6549	0.6770	0.6547	0.6708
Other Borrowed Funds: x_5/assets	0.1143	0.0984	0.1113	0.0966	0.0795	0.0723

Total Assets $2 bill − $10 bill	**2007 (no. obs. = 155)**		**2003 (no. obs. = 131)**		**2010 (no. obs. = 148)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Labor (FTEs): x_1/assets	0.00024	0.00024	0.00029	0.00029	0.00024	0.00022
Physical Capital: x_2/Assets	0.0173	0.0157	0.0151	0.0147	0.0165	0.0145
Uninsured Deposits: x_3/Assets	0.1471	0.1221	0.1199	0.1002	0.1366	0.1274
Insured Deposits: x_4/Assets	0.5909	0.6050	0.6184	0.6239	0.6468	0.6596
Other Borrowed Funds: x_5/assets	0.1384	0.1276	0.1479	0.1308	0.0855	0.0663

Total Assets $10 bill − $50 bill	**2007 (no. obs. = 31)**		**2003 (no. obs. = 43)**		**2010 (no. obs. = 33)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Labor (FTEs): x_1/assets	0.00022	0.00022	0.00028	0.00027	0.00021	0.00020
Physical Capital: x_2/Assets	0.0157	0.0123	0.0145	0.0120	0.0147	0.0121
Uninsured Deposits: x_3/Assets	0.1209	0.0944	0.0831	0.0654	0.0892	0.0710
Insured Deposits: x_4/Assets	0.5532	0.5642	0.5612	0.5742	0.6344	0.6521
Other Borrowed Funds: x_5/assets	0.1931	0.1745	0.2139	0.1977	0.1359	0.0914

Total Assets $50 bill − $100 bill	**2007 (no. obs. = 12)**		**2003 (no. obs. = 10)**		**2010 (no. obs. = 10)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Labor (FTEs): x_1/assets	0.00017	0.00018	0.00028	0.00025	0.00016	0.00016
Physical Capital: x_2/Assets	0.0092	0.0083	0.0156	0.0120	0.0094	0.0091
Uninsured Deposits: x_3/Assets	0.1061	0.1013	0.0819	0.0668	0.0649	0.0560
Insured Deposits: x_4/Assets	0.4749	0.4922	0.5066	0.5509	0.5627	0.5965
Other Borrowed Funds: x_5/assets	0.2563	0.2216	0.2433	0.2154	0.2128	0.1286

Total Assets > $100 bill	**2007 (no. obs. = 17)**		**2003 (no. obs. = 16)**		**2010 (no. obs. = 15)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Labor (FTEs): x_1/assets	0.00018	0.00019	0.00023	0.00022	0.00016	0.00017
Physical Capital: x_2/Assets	0.0097	0.0087	0.0098	0.0096	0.0103	0.0094
Uninsured Deposits: x_3/Assets	0.0769	0.0692	0.0441	0.0385	0.0378	0.0389
Insured Deposits: x_4/Assets	0.3799	0.4604	0.4077	0.4827	0.5049	0.5863
Other Borrowed Funds: x_5/assets	0.3670	0.2881	0.3774	0.3348	0.2760	0.1798

Table 4. Summary Statistics: Risk and Financial Performance by Size Groups

The data on top-tier holding companies were obtained from the Y9-C Call Reports filed quarterly with regulators. A top-tier company is not owned by another company. Banks in the largest size category, with assets exceeding $100 billion, are often perceived as being too big to fail (Brewer and Jagtiani, 2009).

Total Assets < $0.8 billion	2007 (no. obs. = 328)		2003 (no. obs. = 1403)		2010 (no. obs. = 364)	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Equity Capital/Assets	0.099	0.095	0.099	0.095	0.101	0.099
Average Interest Rate on Assets	0.064	0.063	0.054	0.053	0.050	0.049
Nonperforming Assets/Assets	0.025	0.018	0.017	0.013	0.061	0.044
Total Revenue ($1000s)	42232.860	42302.000	22443.550	19035.080	33154.270	32122.470
Profit/Revenue	0.302	0.304	0.400	0.401	0.396	0.411
Profit/Assets	0.023	0.022	0.025	0.025	0.023	0.023

Total Assets $0.8 bill − $2 bill	2007 (no. obs. = 299)		2003 (no. obs. = 252)		2010 (no. obs. = 286)	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Equity Capital/Assets	0.100	0.095	0.097	0.094	0.102	0.099
Average Interest Rate on Assets	0.062	0.062	0.051	0.050	0.047	0.047
Nonperforming Assets/Assets	0.020	0.015	0.014	0.012	0.055	0.041
Total Revenue ($1000s)	88222.160	81848.000	75112.510	67723.170	65780.780	61923.680
Profit/Revenue	0.311	0.309	0.420	0.416	0.412	0.425
Profit/Assets	0.023	0.022	0.026	0.025	0.023	0.023

Total Assets $2 bill − $10 bill	2007 (no. obs. = 155)		2003 (no. obs. = 131)		2010 (no. obs. = 148)	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Equity Capital/Assets	0.105	0.103	0.100	0.096	0.119	0.115
Average Interest Rate on Assets	0.061	0.060	0.049	0.049	0.044	0.044
Nonperforming Assets/Assets	0.020	0.016	0.014	0.013	0.060	0.047
Total Revenue ($1000s)	300481.430	246612.000	255253.190	207907.430	230735.420	180073.680
Profit/Revenue	0.336	0.329	0.441	0.437	0.466	0.469
Profit/Assets	0.025	0.024	0.027	0.027	0.026	0.026

Total Assets $10 bill − $50 bill	2007 (no. obs. = 31)		2003 (no. obs. = 43)		2010 (no. obs. = 33)	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Equity Capital/Assets	0.114	0.112	0.113	0.111	0.127	0.127
Average Interest Rate on Assets	0.055	0.056	0.044	0.045	0.040	0.039
Nonperforming Assets/Assets	0.017	0.015	0.013	0.012	0.047	0.045
Total Revenue ($1000s)	1209522.320	1032973.000	1622270.070	1400739.440	1028775.920	848829.270
Profit/Revenue	0.353	0.355	0.483	0.490	0.483	0.493
Profit/Assets	0.026	0.023	0.033	0.032	0.026	0.027

Total Assets $50 bill − $100 bill	2007 (no. obs. = 12)		2003 (no. obs. = 10)		2010 (no. obs. = 10)	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Equity Capital/Assets	0.142	0.135	0.138	0.136	0.154	0.160
Average Interest Rate on Assets	0.041	0.044	0.046	0.040	0.037	0.033
Nonperforming Assets/Assets	0.016	0.014	0.018	0.016	0.045	0.042
Total Revenue ($1000s)	4359240.750	4349244.500	5432698.800	4164318.310	3846224.880	3407826.540
Profit/Revenue	0.385	0.389	0.505	0.493	0.528	0.510
Profit/Assets	0.026	0.028	0.041	0.032	0.033	0.027

Total Assets > $100 bill	2007 (no. obs. = 17)		2003 (no. obs. = 16)		2010 (no. obs. = 15)	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Equity Capital/Assets	0.131	0.130	0.114	0.109	0.149	0.144
Average Interest Rate on Assets	0.040	0.042	0.041	0.040	0.036	0.035
Nonperforming Assets/Assets	0.026	0.020	0.020	0.018	0.056	0.060
Total Revenue ($1000s)	40372301.180	15015000.000	24644113.330	13658238.340	36750160.480	16167413.680
Profit/Revenue	0.404	0.398	0.517	0.517	0.544	0.545
Profit/Assets	0.033	0.031	0.036	0.035	0.034	0.032

Table 5. Summary Statistics: Prices by Size Groups

The data on top-tier holding companies were obtained from the Y9-C Call Reports filed quarterly with regulators. A top-tier company is not owned by another company. Banks in the largest size category, with assets exceeding $100 billion, are often perceived as being too big to fail (Brewer and Jagtiani, 2009).

Total Assets < $0.8 billion	**2007 (no. obs. = 328)**		**2003 (no. obs. = 1403)**		**2010 (no. obs. = 364)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Wage Rate: w_1	58.495	56.801	55.978	53.250	60.591	57.922
Price of Physical Capital: w_2	0.261	0.200	0.275	0.213	0.273	0.189
Uninsured Deposit Rate: w_3	0.049	0.047	0.027	0.027	0.021	0.020
Insured Deposit Rate: w_4	0.028	0.028	0.015	0.015	0.011	0.011
Other Borrowed Funds rate: w_5	0.057	0.050	0.038	0.035	0.046	0.039
Tax Rate	0.420	0.420	0.414	0.413	0.414	0.415
Price of After-Tax Profit (1/(1–t))	1.728	1.724	1.709	1.702	1.712	1.709
Total Assets $0.8 bill − $2 bill	**2007 (no. obs. = 299)**		**2003 (no. obs. = 252)**		**2010 (no. obs. = 286)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Wage Rate: w_1	63.261	59.208	60.726	57.694	62.751	59.377
Price of Physical Capital: w_2	0.276	0.206	0.296	0.236	0.263	0.197
Uninsured Deposit Rate: w_3	0.049	0.048	0.026	0.026	0.020	0.020
Insured Deposit Rate: w_4	0.029	0.028	0.013	0.013	0.010	0.010
Other Borrowed Funds rate: w_5	0.053	0.049	0.032	0.029	0.042	0.037
Tax Rate	0.421	0.420	0.414	0.415	0.412	0.419
Price of After-Tax Profit (1/(1–t))	1.730	1.724	1.711	1.709	1.704	1.721
Total Assets $2 bill − $10 bill	**2007 (no. obs. = 155)**		**2003 (no. obs. = 131)**		**2010 (no. obs. = 148)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Wage Rate: w_1	68.491	62.710	67.708	60..640	66.464	62.618
Price of Physical Capital: w_2	0.345	0.251	0.346	0.272	0.346	0.265
Uninsured Deposit Rate: w_3	0.047	0.047	0.024	0.025	0.018	0.018
Insured Deposit Rate: w_4	0.026	0.026	0.012	0.012	0.008	0.008
Other Borrowed Funds rate: w_5	0.052	0.046	0.032	0.029	0.037	0.031
Tax Rate	0.421	0.423	0.416	0.423	0.412	0.420
Price of After-Tax Profit (1/(1–t))	1.729	1.733	1.718	1.733	1.705	1.725
Total Assets $10 bill − $50 bill	**2007 (no. obs. = 31)**		**2003 (no. obs. = 43)**		**2010 (no. obs. = 33)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Wage Rate: w_1	68.748	66.348	71.799	63.985	77.693	71.852
Price of Physical Capital: w_2	0.302	0.273	0.363	0.313	0.337	0.289
Uninsured Deposit Rate: w_3	0.054	0.050	0.022	0.023	0.015	0.014
Insured Deposit Rate: w_4	0.024	0.025	0.010	0.009	0.006	0.006
Other Borrowed Funds rate: w_5	0.047	0.045	0.027	0.025	0.031	0.030
Tax Rate	0.422	0.423	0.423	0.425	0.418	0.421
Price of After-Tax Profit (1/(1–t))	1.735	1.733	1.734	1.739	1.722	1.727
Total Assets $50 bill − $100 bill	**2007 (no. obs. = 12)**		**2003 (no. obs. = 10)**		**2010 (no. obs. = 10)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Wage Rate: w_1	79.908	73.445	78.445	82.428	89.282	83.805
Price of Physical Capital: w_2	0.347	0.342	0.406	0.386	0.369	0.357
Uninsured Deposit Rate: w_3	0.046	0.047	0.025	0.023	0.019	0.017
Insured Deposit Rate: w_4	0.020	0.020	0.012	0.011	0.007	0.004
Other Borrowed Funds rate: w_5	0.039	0.044	0.019	0.018	0.015	0.011
Tax Rate	0.424	0.427	0.425	0.430	0.396	0.418
Price of After-Tax Profit (1/(1–t))	1.737	1.745	1.740	1.755	1.661	1.717
Total Assets > $100 bill	**2007 (no. obs. = 17)**		**2003 (no. obs. = 16)**		**2010 (no. obs. = 15)**	
Variable	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
Wage Rate: w_1	88.800	84.380	85.654	82.831	90.675	80.429
Price of Physical Capital: w_2	0.443	0.394	0.436	0.404	0.433	0.370
Uninsured Deposit Rate: w_3	0.043	0.048	0.026	0.025	0.020	0.022
Insured Deposit Rate: w_4	0.023	0.023	0.008	0.008	0.005	0.005
Other Borrowed Funds rate: w_5	0.043	0.043	0.021	0.019	0.024	0.021
Tax Rate	0.430	0.425	0.428	0.425	0.422	0.421
Price of After-Tax Profit (1/(1–t))	1.756	1.739	1.749	1.739	1.734	1.727

Table 6
Estimated Mean Scale Economies

Scale economies are calculated as the mean of the estimated scale economies at each point in the sample or size category (rather than scale economies evaluated at the mean of the data). Size categories are in 2007 dollars. The data on top-tier holding companies were obtained from the Y9-C Call Reports filed quarterly with regulators. A top-tier company is not owned by another company. Banks in the largest size category, with assets exceeding $100 billion, are often perceived as being too big to fail (Brewer and Jagtiani, 2009).

The estimations include the cost function and input share equations for the theoretically mis-specified cash-flow cost function (omitting the amount of equity capital) and the theoretically proper cash-flow cost function (conditioned on the amount of equity capital). The economic-cost scale economies are inferred from the theoretically proper cash-flow cost function. In addition, we estimate the managers' most preferred profit function and input demand functions, which reflect the bank's risk-expected-return trade-off, and compute the managers' most preferred cost function from the profit function.

	2007 (no. of obs. = 842)				**2003 (no. of obs. = 1855)**				**2010 (no. of obs. = 856)**			
Asset Category	**(1) Mis-specified Cash-Flow Cost Function Omits Level of Equity Mean (Std. Err.) Median**	**(2) Correct Cash-Flow Cost Function Conditioned on Level of Equity Mean (Std. Err.) Median**	**(3) Economic Cost Function Includes Shadow Cost of Equity Mean (Std. Err.) Median**	**(4) Managers' Most Preferred Cost Function Conditioned on Optimal Equity Mean (Std. Err.) Median**	**(1) Mis-specified Cash-Flow Cost Function Omits Level of Equity Mean (Std. Err.) Median**	**(2) Correct Cash-Flow Cost Function Conditioned on Level of Equity Mean (Std. Err.) Median**	**(3) Economic Cost Function Includes Shadow Cost of Equity Mean (Std. Err.) Median**	**(4) Managers' Most Preferred Cost Function Conditioned on Optimal Equity Mean (Std. Err.) Median**	**(1) Mis-specified Cash-Flow Cost Function Omits Level of Equity Mean (Std. Err.) Median**	**(2) Correct Cash-Flow Cost Function Conditioned on Level of Equity Mean (Std. Err.) Median**	**(3) Economic Cost Function Includes Shadow Cost of Equity Mean (Std. Err.) Median**	**(4) Managers' Most Preferred Cost Function Conditioned on Optimal Equity Mean (Std. Err.) Median**
Full sample	**1.0340** (0.00575) 1.0330	0.9744 (0.0199) 0.9697	**1.0333** (0.00843) 1.0317	**1.1394** (0.0503) 1.1232	**1.0292** (0.00532) 1.0282	**0.9350** (0.0146) 0.9325	**1.0474** (0.00676) 1.0457	**1.1829** (0.00640) 1.1573	**1.0664** (0.005976) 1.0633	0.9841 (0.0294) 0.9814	**1.0476** (0.0142) 1.0493	**1.2539** (0.0154) 1.2139
< $0.8 bill	**1.0274** (0.00666) 1.0260	0.9502** (0.0198) 0.9444	**1.0298** (0.00917) 1.0287	**1.1227** (0.0436) 1.1149	**1.0271** (0.00575) 1.0263	**0.9241** (0.0144) 0.9228	**1.0492** (0.00706) 1.0475	**1.1690** (0.00605) 1.1500	**1.0735** (0.0115) 1.0708	1.0076 (0.0324) 1.0006	**1.0636** (0.0165) 1.0639	**1.2269** (0.0140) 1.1925
$0.8 bill – $2 bill	**1.0343** (0.00596) 1.0328	0.9748 (0.0198) 0.9715	**1.0331** (0.00844) 1.0315	**1.1334** (0.0497) 1.1200	**1.0321** (0.00505) 1.0306	**0.9553** (0.0169) 0.9555	**1.0455** (0.00729) 1.0421	**1.2001** (0.00698) 1.1717	**1.0608** (0.00979) 1.0592	0.9892 (0.0297) 0.9836	**1.0495** (0.0142) 1.0489	**1.2291** (0.0137) 1.2112
$2 bill – $10 bill	**1.0402** (0.00547) 1.0389	0.9932 (0.0232) 0.9924	**1.0361** (0.00990) 1.0356	**1.1489** (0.0522) 1.1398	**1.0366** (0.00548) 1.0356	0.9795 (0.0225) 0.9791	**1.0395** (0.00930) 1.0346	**1.1990** (0.00775) 1.1806	**1.0643** (0.00936) 1.0597	0.9505 (0.0307) 0.9398	1.0294** (0.01497) 1.0318	**1.3258** (0.0213) 1.2531
$10 bill – $50 bill	**1.0483** (0.00672) 1.0481	1.0348 (0.0324) 1.0237	**1.0417** (0.0146) 1.0383	**1.1821** (0.0658) 1.1584	**1.0422** (0.00756) 1.0385	0.9976 (0.0335) 0.9856	1.0305** (0.0135) 1.0203	**1.3676** (0.0159) 1.2686	**1.0528** (0.0119) 1.0441	0.9272* (0.0411) 0.9084	1.0018 (0.0214) 1.0010	**1.3110** (0.0256) 1.2796
$50 bill – $100 bill	**1.0492** (0.00808) 1.0553	1.0169 (0.0393) 1.0185	1.0441** (0.0180) 1.0456	**1.2309** (0.0785) 1.2087	**1.0544** (0.00848) 1.0446	0.9863 (0.0416) 0.9773	1.0353** (0.0169) 1.0228	**1.4065** (0.0189) 1.2877	**1.0641** (0.0146) 1.0555	**0.8765** (0.0457) 0.8664	0.9789 (0.02639) 0.9837	**1.4270** (0.0392) 1.3739
> $100 bill	**1.0597** (0.0112) 1.0693	1.1224** (0.0624) 1.0624	1.0584** (0.0269) 1.0580	**1.3353** (0.1295) 1.2765	**1.0537** (0.0112) 1.0526	1.0081 (0.0521) 1.0062	1.0373* (0.0205) 1.0303	**1.3567** (0.0181) 1.3317	**1.0540** (0.0182) 1.0350	**0.8419** (0.0568) 0.8369	0.9506 (0.0310) 0.9697	**1.4315** (0.0489) 1.4172

Standard errors are given in parentheses.
All estimates of scale economies are significantly different from 0 at the 1 percent level.
Estimates of scale economies in **bold** are significantly different from 1 at the 1 percent level.
* Significantly different from 1 at the 10 percent level
** Significantly different from 1 at the 5 percent level

Table 7
Robustness Results for Estimated Mean Scale Economies

Scale economies are calculated as the mean of the estimated scale economies at each point in the sample or size category (rather than scale economies evaluated at the mean of the data). Size categories are in 2007 dollars. The data on top-tier holding companies were obtained from the Y9-C Call Reports filed quarterly with regulators. A top-tier company is not owned by another company. Banks in the largest size category, with assets exceeding $100 billion, are often perceived as being too big to fail (Brewer and Jagtiani, 2009).

For each specification, we estimate the managers' most preferred profit function and input demand functions, which reflect the bank's risk-expected return trade-off, and compute the managers' most preferred cost function from the profit function. The results pertain to 2007.

	Managers' Most Preferred Cost Function Conditioned on Optimal Equity 2007 data					
Asset Category	**(1) Baseline model** **Mean** **(Std. Err.)** **Median** **No. of obs.**	**(2) Re-estimated excluding BHCs with total assets ≤ $2 billion** **Mean** **(Std. Err.)** **Median** **No. of obs.**	**(3) Re-estimated excluding observations with most extreme output shares** **Mean** **(Std. Err.)** **Median** **No. of obs.**	**(4) Re-estimated with alternative definition of y_3** **Mean** **(Std. Err.)** **Median** **No. of obs.**	**(5) Re-estimated excluding BHCs with total assets > $100 billion, then calculates scale economies for all banks** **Mean** **(Std. Err.)** **Median** **No. of obs. in calculation of scale economies**	**(6) Baseline estimation but calculating scale economies for BHCs with total assets > $100 billion using the median input prices for w_3, w_4, w_5 for BHCs with total assets < $100 billion in assets** **Mean** **(Std. Err.)** **Median** **No. of obs.**
Full sample	**1.1394** (0.0503) 1.1232 n = 842	**1.1753** (0.0202) 1.1753 n = 215	**1.1452** (0.00856) 1.1323 n = 830	**1.1490** (0.00952) 1.1341 n = 842	**1.1419** (0.00849) 1.1239 n = 842	**1.1396** (0.0502) 1.1232 n = 842
< $0.8 bill	**1.1227** (0.0436) 1.1149 n = 328		**1.1328** (0.00789) 1.1231 n = 326	**1.1364** (0.00874) 1.1272 n = 328	**1.1264** (0.00801) 1.1177 n = 328	**Same as (1) Baseline column**
$0.8 bill – $2 bill	**1.1334** (0.0497) 1.1200 n = 299		**1.1410** (0.00846) 1.1294 n = 296	**1.1421** (0.00930) 1.1297 n = 299	**1.1362** (0.00844) 1.1237 n = 299	**Same as (1) Baseline column**
$2 bill – $10 bill	**1.1489** (0.0522) 1.1398 n = 155	**1.1452** (0.0185) 1.1378 n = 155	**1.1567** (0.00928) 1.1482 n = 155	**1.1549** (0.0103) 1.1470 n = 155	**1.1484** (0.00931) 1.1389 n = 155	**Same as (1) Baseline column**
$10 bill – $50 bill	**1.1821** (0.0658) 1.1584 n = 31	**1.1881** (0.0229) 1.1980 n = 31	**1.1889** (0.0123) 1.1636 n = 30	**1.1782** (0.0135) 1.1518 n = 31	**1.1793** (0.0128) 1.1567 n = 31	**Same as (1) Baseline column**
$50 bill – $100 bill	**1.2309** (0.0785) 1.2087 n = 12	**1.2635** (0.0317) 1.2423 n = 12	**1.2061** (0.0147) 1.2153 n = 11	**1.2330** **(0.0177)** 1.1976 n = 12	**1.2374** (0.0186) 1.2131 n = 12	**Same as (1) Baseline column**
> $100 bill	**1.3353** (0.1295) 1.2765 n = 17	**1.3646** (0.0500) 1.3184 n = 17	**1.2670** (0.0206) 1.2762 n = 12	**1.3478** **(0.0295)** 1.2508 n = 17	**1.3481** (0.0276) 1.2679 n = 17	**1.3466** (0.1228) 1.2776 n = 17

Standard errors are given in parentheses.
All estimates of scale economies are significantly different from 0 at the 1 percent level.
Estimates of scale economies in **bold** are significantly different from 1 at the 1 percent level.
* Significantly different from 1 at the 10 percent level
** Significantly different from 1 at the 5 percent level

Table 8

Estimated Mean Scale Economies and Cost Elasticities along the Value-Maximizing Expansion Path

Mean scale economies and cost elasticities (1/scale economies) are calculated as the mean of the estimated scale economies and cost elasticities at each point in the sample or subsample. The data, obtained from the Y9-C Call Reports filed quarterly with regulators, pertain to top-tier U.S. bank holding companies. A top-tier company is not owned by another company. For the purposes of this table, the sample is restricted to publicly traded companies with market value data from Compustat.

The table reports the mean scale economies and cost elasticities for BHCs that maximize managers' utility when expanding their scale of operations, as given by the managers' most preferred cost function model and for BHCs that maximize market value when expanding their scale of operations. In the absence of agency problems between managers and outside owners, these expansion paths can be assumed to be the same. However, in the presence of agency problems, the utility-maximizing path of expansion may reflect managers' private concerns for perquisites and risk. To identify those firms where the paths coincide, we assume that an efficient BHC's utility-maximizing expansion path will be the path that maximizes market value.

To determine the efficient BHCs in the sample we estimate a stochastic frontier of the market value of assets as a function of the book value of assets adjusted to remove goodwill (equation 6 in the text). A bank's efficiency is measured by the ratio of its achieved market value to its highest potential value, which is the value on the frontier. Value-maximizing BHCs are identified as those in the highest quartile of market-value efficiency. For comparison, means for BHCs in the lowest efficiency quartile are given in braces. Note, for 2010, the null hypothesis that the composed error term is two-sided could not be rejected. Hence, for 2010, the value-maximizing path coincides with the utility-maximizing path for all BHCs in the sample.

Sample	2007		2003		2010
	Utility-Max. Expansion Path Scale econs Cost elast No. of obs.	**Value-Max. Expansion Path** Scale econs Cost elast 25% most effic {25% least effic}	**Utility-Max. Expansion Path** Scale econs Cost elast No. of obs.	**Value-Max. Expansion Path** Scale econs Cost elast 25% most effic {25% least effic}	**Utility-Max. and Value-Max Expansion Path** Scale econs Cost elast No. of obs.
Full Sample	1.1626 0.8640 n = 219	1.2358 0.8138 {1.1296 0.8883}	1.2081 0.8357 n = 349	1.2905 0.7862 {1.1642 0.8653}	1.2695 0.7963 n = 220
< $2 billion	1.1290 0.8881 n = 99	1.1348 0.8827 {1.1373 0.8840}	1.1794 0.8532 n = 220	1.1950 0.8404 {1.1563 0.8683}	1.2103 0.8315 n = 103
$2 billion – $10 billion	1.1653 0.8609 n = 80	1.2014 0.8369 {1.1320 0.8855}	1.2052 0.8346 n = 83	1.2509 0.8068 {1.1807 0.8501}	1.3000 0.7754 n = 75
> $10 billion	1.2403 0.8107 n = 40	1.2854 0.7848 {1.1738 0.8543}	1.3506 0.7536 n = 46	1.4813 0.6980 {1.3237 0.7643}	1.3601 0.7474 n = 42