Hughes, Joseph P.; Mester, Loretta J.

Working Paper

The performance of financial institutions: Modeling, evidence, and some policy implications

Working Paper, No. 2018-05

Provided in Cooperation with:
Department of Economics, Rutgers University

The Performance of Financial Institutions: Modeling, Evidence, and Some Policy Implications

Joseph P. Hughes
Rutgers University

and

Loretta J. Mester
Federal Reserve Bank of Cleveland
and
The Wharton School, University of Pennsylvania

June 25, 2018

Prepared for the *Oxford Handbook of Banking, 3rd edition*

Abstract. The unique capital structure of commercial banking – funding production with demandable debt that participates in the economy's payments system – affects various aspects of banking. It shapes commercial banks' comparative advantage in providing financial products and services to informationally opaque customers, their ability to diversify credit and liquidity risk, and how they are regulated, including the need to obtain a charter to operate and explicit and implicit federal guarantees of bank liabilities to reduce the probability of bank runs. These aspects of banking affect a bank's choice of risk versus expected return, which, in turn, affects bank performance. Banks have an incentive to reduce risk to protect their valuable charters from episodes of financial distress, and they also have an incentive to increase risk to exploit the cost-of-funds subsidy of mispriced deposit insurance. These are contrasting incentives tied to bank size. Measuring bank performance and its relationship to size requires untangling cost and profit from decisions about risk versus expected return because both cost and profit are functions of endogenous risk-taking. This chapter gives an overview of two general empirical approaches to measuring bank performance and discusses some of the applications of these approaches found in the literature. One application explains how better diversification available at a larger scale of operations generates scale economies that are obscured by higher levels of risk-taking. Studies of commercial banking cost that ignore endogenous risk-taking find little evidence of scale economies at the largest banks, while those that control for this risk-taking find large scale economies at the largest banks – evidence with important implications for regulation.

Keywords: Bank, Efficiency, Risk, Cost, Profit, Scale Economies, X-Inefficiency

Direct correspondence to

Joseph P. Hughes, Professor, Department of Economics, Rutgers University, New Brunswick, NJ 08901, jphughes@rci.rutgers.edu

Loretta J. Mester, President and CEO, Federal Reserve Bank of Cleveland, 1455 E. 6th Street, Cleveland, OH 44114, Loretta.Mester@clev.frb.org

The authors thank the editors Allen Berger, Phillip Molyneux, and John Wilson for helpful comments.

1 Introduction

The commercial banking industry is undergoing disruption from other types of financial intermediaries. To understand the future evolution of the industry in the wake of such disruption, it is important to take a step back and ask: What do commercial banks do? What are the key components of commercial banking technology that allow them to do it? And what determines whether banks do it efficiently?

Banks' ability to ameliorate informational asymmetries between borrowers and lenders and to manage risks is the essence of bank production. The literature on financial intermediation suggests that commercial banks, by screening and monitoring borrowers, can help solve potential moral hazard and adverse selection problems caused by the imperfect information between borrowers and lenders. Commercial banks are unique in issuing demandable debt that participates in the economy's payments system. This debt confers an informational advantage to banks over other lenders in making loans to informationally opaque borrowers. In particular, the information obtained from checking account transactions and other sources allows banks to assess and manage risk, write contracts, monitor contractual performance, and, when required, resolve nonperformance problems. Bhattacharya and Thakor (1993) review the modern theory of financial intermediation, which takes an informational approach to banking.

That commercial banks' liabilities are demandable debt also gives banks an incentive advantage over other intermediaries. The relatively high level of debt in a bank's capital structure disciplines managers' risk-taking and their diligence in producing financial services by exposing the bank to an increased risk of insolvency. The demandable feature of the debt, to the extent that it is not fully insured, further heightens performance pressure and safety concerns by increasing liquidity risk. These incentives tend to make banks good monitors of their borrowers. Thus, banks' unique funding by demandable debt that participates in the economy's payments system gives banks both an incentive advantage and an informational advantage in lending to firms too informationally opaque to borrow in public debt and equity markets. The uniqueness of commercial bank production, in contrast to the production of other

types of lenders, is derived from the special characteristics of banks' capital structure: the funding of informationally opaque assets with demand deposits.[1] Calomiris and Kahn (1991) and Flannery (1994) discuss the optimal capital structure of commercial banks.

But banks' ability to perform efficiently – to adopt appropriate investment strategies, to obtain accurate information concerning their customers' financial prospects, and to write and enforce effective contracts – depends in part on the property rights and legal, regulatory, and contracting environments in which they operate. Such an environment includes accounting practices, chartering rules, government regulations, and the market conditions (e.g., market power) under which banks operate. Differences in these features across political jurisdictions can lead to differences in the efficiency of banks across jurisdictions.[2]

Banks' unique funding by demand deposits motivates key components of the legal and regulatory environments that influence managerial incentives for risk-taking and efficiency. Banks' participation in the payments system leads to their regulation and, in particular, to restrictions on entry into the industry. The need to obtain a charter to open a bank confers a degree of market power on banks operating in smaller markets and, in general, permits banks to exploit valuable investment opportunities related to financial intermediation and payments. Government regulation and supervision of banks promotes their safety and soundness in order to protect the payments system from bank runs that reduce bank lending and threaten macroeconomic stability. Protecting the payments system frequently involves deposit insurance. To the extent that the insurance is credible, it reduces depositors' incentive to run banks when they have fears about banks' solvency. Consequently, it reduces banks' liquidity risk and, to the extent it

[1] Berlin and Mester (1999) find empirical evidence of an explicit link between banks' liability structure and their distinctive lending behavior. As discussed in Mester (2007), relationship lending is associated with lower loan rates, less stringent collateral requirements, a lower likelihood of credit rationing, contractual flexibility, and reduced costs of financial distress for borrowing firms. Banks' access to core deposits, which are rate inelastic, enables banks to insulate borrowers with whom they have durable relationships from exogenous credit shocks. Mester, Nakamura, and Renault (2007) also find empirical evidence of a synergy between the liability and asset sides of a commercial bank's balance sheet, showing that information on the cash flows into and out of a borrower's transactions account can help an intermediary monitor the changing value of collateral that a small-business borrower has posted.

[2] Demirgüç-Kunt, Kane, and Laeven (2007) use a sample of 180 countries to study the external and internal political features that influence the adoption and design of deposit insurance, which, in turn, affects the efficiency of the domestic banking system.

is underpriced, gives banks the incentive to take additional risk for higher expected return. Moreover, if uninsured creditors also believe they will be protected from losses because government regulators will treat the bank as too big to fail, the risk-taking incentive is heightened further.

2 Banking Technology and Performance

2.1 Banks' Risk Menu and Conflicting Incentives for Risk-Taking

Mispriced deposit insurance and too-big-to-fail policies can create a cost-of-funds subsidy that gives banks an incentive to take additional risk.[3] But banks also have an incentive to avoid risk to protect their valuable charters from episodes of financial distress. Distress involves liquidity crises resulting from runs by uninsured depositors, regulatory intervention in banks' investment decisions, and even the loss of the charter when distress results in insolvency. As discussed in Hughes and Mester (2013b), Marcus (1984) finds that banks with high-valued investment opportunities maximize their expected market value by pursuing lower-risk investment strategies that protect their charters and thereby preserve their ability to exploit these opportunities. On the other hand, banks with low-valued investment opportunities maximize their expected value by adopting higher-risk investment strategies that exploit the cost-of-funds subsidy of mispriced deposit insurance (Keeley, 1990). Mid-range risk strategies do not maximize value. These dichotomous investment strategies, as well as other sources of risk-taking and risk-avoidance, fundamentally shape production decisions and must be taken into account when modeling bank production. Herring and Vankudre (1987) offer a similar analysis of these dichotomous investment strategies. Hughes, Mester, and Moon (2016) provide evidence of these dichotomous value-maximizing strategies in banks' capitalization, and Hughes and Moon (2018) find similar evidence in credit risk strategies.

[3] FDIC (2014) summarizes some of the estimates of the subsidy found in the literature.

The risk environment banks face can be characterized by a frontier of expected return and return risk, which shows a bank's menu of efficient investment choices.[4] In Figure 1 from Hughes and Mester (2013b), a smaller bank's menu of investment choices is given by the lower frontier. Consider a smaller bank that operates at point A.[5] To illustrate scale-related diversification, suppose a larger bank is created by scaling up the assets of this smaller bank. In principle, the larger bank can obtain better diversification of its assets, which reduces credit risk, and better diversification of its deposits, which reduces liquidity risk. Thus, the larger bank can efficiently produce the expected return of the smaller bank (point A) with less return risk (point A′). In fact, the larger bank will likely take advantage of its better diversification and produce a different (and perhaps more complicated) mix of financial services. Nonetheless, the risk-expected-return frontier of the larger bank lies above that of the smaller bank because the larger bank has a better menu of investment choices resulting from improved diversification.

[Insert Figure 1]

Textbooks point to better diversification, which reduces the costs of risk management, as a key source of scale economies. The link between better diversification and scale economies is apparent when comparing a larger bank operating at point A′ with one operating at point B. A larger bank operating at point A′ has the same expected return but lower risk than the smaller bank operating at point A, while a larger bank at point B operates with the same return risk as the smaller bank but obtains a higher expected return. At point B, the better diversification of deposits allows the larger bank to economize on liquid assets without increasing liquidity risk, while the better diversification of loans allows it to economize on equity capital without increasing insolvency risk. Thus, its expected return for the same risk as the smaller bank is higher.

---

[4] For expository purposes, in this discussion we are assuming that only the first two moments of the distribution of returns matter for bank production. More generally, however, higher moments, such as skewness and kurtosis, can be expected to influence, for example, calculations of value-at-risk and the choice of investment strategies that minimize the probability of financial distress or that exploit the federal safety net. Thus, risk resulting from higher moments likely plays an important role in bank production.

[5] To simplify the discussion, we assume that the smaller bank operates efficiently; therefore, point A lies on the frontier rather than beneath it. See Hughes and Mester (2013b) for an analysis of how inefficiency is related to scale economies in banking.

Better diversification, though, does not necessarily mean that the larger bank operates with less risk; rather, it means the larger bank experiences a better risk-expected-return frontier. Heightened competition and lower-valued growth opportunities in the larger bank's markets or lower marginal costs of risk management might induce the larger bank to choose to produce its output with more risk in order to obtain a higher expected return – say, the strategy at point C or point D.

A bank's risk-taking is also influenced by external and internal mechanisms that discipline bank managers. Internal discipline might be induced or reduced by organizational form, ownership and capital structure, governing boards, and managerial compensation. External discipline might be induced or reduced by government regulation and the safety net, capital market discipline (takeovers, cost of funds, stakeholders' ability to sell stock), managerial labor market competition, outside blockholders of equity and debt, and product market competition.[6] This operating environment can also create agency conflicts that influence managers' incentives to pursue value-maximizing risk strategies. Managers whose wealth consists largely of their undiversified human capital tend to avoid riskier investment strategies that maximize the value of banks with poorer investment opportunities. However, the presence of a diversified outside owner of a large block of stock might encourage the board of directors to put in place a compensation plan that overcomes managers' risk aversion and encourages value-maximizing risk-taking (Laeven and Levine, 2009; Cheng, Hong, and Scheinkman, 2015).

Thus, in order to measure the efficiency of bank production, it is important to account for bank risk-taking and the optimality of this risk-taking.

2.2 The Empirical Measurement of Banking Technology and Performance

There are two broad approaches to measuring technology and explaining performance: non-structural and structural. Using a variety of financial measures that capture various aspects of

[6] La Porta, Lopez-de-Silanes, and Shleifer (2002) examine banking systems in 92 countries and find that government ownership is correlated with poorer countries and countries with less developed financial systems, poorer protection of investors' rights, more government intervention, and poorer performance of institutions. They also find that government ownership is associated with higher cost ratios and wider interest rate margins. Aghion, Alesina, and Trebbi (2007) provide evidence that democracy has a positive impact on productivity growth in more advanced sectors of the economy, possibly by fostering entry and competition.

performance, the non-structural approach compares performance among banks and considers the relationship of performance to investment strategies and other factors such as characteristics of regulation and governance. For example, the non-structural approach might investigate technology by asking how performance measures are correlated with such investment strategies as growing by asset acquisitions and diversifying or focusing the bank's product mix. It looks for evidence of agency problems in correlations of performance measures and variables characterizing the quality of banks' governance. While informal and formal theories may motivate some of these investigations, no general theory of performance provides a unifying framework for these studies.

The structural approach is choice-theoretic and, as such, relies on a theoretical model of the banking firm and a concept of optimization. The older literature applies the traditional microeconomic theory of production to banking firms in much the same way as it is applied to non-financial firms and industries. The newer literature views the bank as a financial intermediary that produces informationally intensive financial services and takes on and diversifies risks – unique, essential aspects of financial intermediation that are not generally taken into account in traditional applications of production theory.[7] For example, the traditional theory defines a cost function by a unique cost-minimizing combination of inputs for any given level of outputs. Thus, the cost function gives the minimum cost of any given output vector without regard to the return risk implied by the cost-minimizing input vector. Ignoring the implied return risk may be appropriate for non-financial firms, but for financial institutions, return risk plays an essential role in maximizing the discounted flow of expected profits. First, return risk influences the rate at which future expected profits are discounted. Second, return risk affects the expected cost of financial distress. The bank with high-valued investment opportunities may find the level of risk associated with

[7] This framework often guides the choice of outputs and inputs in the bank's production structure. For example, as discussed in Mester (2008), the traditional application of efficiency analysis to banking does not allow bank production decisions to affect bank risk, which rules out the possibility that scale- and scope-related improvements in diversification could lower the cost of borrowed funds and induce banks to alter their risk exposure. Also, much of the traditional literature does not account for the bank's role in producing information about its borrowers in its underwriting decisions when specifying the bank's outputs and inputs. An exception is Mester (1992), who directly accounted for banks' monitoring and screening role by measuring bank output treating loans purchased and loans originated as separate outputs entailing different types of screening, and treating loans held on balance sheet and loans sold as separate outputs entailing different types of monitoring.

the cost-minimizing vector too high. If so, it may choose to reduce the credit risk of the given output vector by adding more labor and physical capital to improve its credit evaluation and loan monitoring capabilities. In doing so, it trades higher cost (lower profit) for lower profit risk to reduce the expected costs of financial distress and the discount rate on its expected cash flow, thus maximizing its market value. This trade-off suggests that measuring bank performance by a cost metric or a profit metric that fails to account for endogenous risk-taking is likely to be seriously biased.

Notice in this example that risk influences the decision of how to produce a given output vector and, thus, must influence the cost of producing it. In Figure 1, when the risk-expected-return frontier for the larger bank is narrowly interpreted as showing different investment strategies for producing the same output vector – the scaled-up outputs of the smaller bank – it is clear that larger banks with higher-valued investment opportunities are likely to choose a lower risk-expected-return strategy, say, point B or A′, than banks with lower-valued opportunities, say, point C or D. Since the cost of producing the scaled-up output vector is likely to differ along the frontier, the value-maximizing input vector and, hence, the cost of the output, will be driven in part by risk considerations. And these risk considerations imply that ***revenue influences cost when risk matters***. How, then, can managers' preferences for these production plans and their implied risk be represented?

Letting the output vector be represented by $\boldsymbol{q}$, the input vector by $\boldsymbol{x}$, and equity capital by k, the technology for producing a given output vector is represented by the transformation function $T(\boldsymbol{x}, k; \boldsymbol{q}) \leq 0$. Points C and D in Figure 1 arise from different input vectors $(\boldsymbol{x}, k)$ that produce the given output $\boldsymbol{q}$. Let $\boldsymbol{z}$ represent the production plan and price environment. Managers' beliefs about how production plans interact with a given state of the world, s, to yield profit, π, imply a realization of profit, $\pi = g(\boldsymbol{z}, s)$, that is conditioned on the state of the world. And managers' beliefs about the probability distribution of states of the world imply a subjective distribution of profit that is conditional on the production plan: $f(\pi; \boldsymbol{z})$.

Under well-known restrictive conditions, this distribution can be represented by its first two moments, $E(\pi; \boldsymbol{z})$ and $S(\pi; \boldsymbol{z})$.[8]

The traditional literature on bank production and efficiency assumes that banks choose their production plan to minimize expected cost and maximize expected profit: Managers rank production plans by their *expected* profit and cost, the first moment of their subjective probability distribution of profit, $f(\pi; \boldsymbol{z})$, attached to each production plan. The newer research assumes that bank managers maximize the utility of their production plans. Rather than define the utility function over the first two moments, the newer literature defines it over profit and the production plan, $U(\pi; \boldsymbol{z})$, which is equivalent to defining it over the conditional probability distributions $f(\pi; \boldsymbol{z})$. Utility maximization is a more general objective that subsumes profit maximization and cost minimization (e.g., Hughes, 1999; Hughes, et al., 1999; Hughes, et al., 2000; and Hughes, Mester, and Moon, 2001; Hughes and Mester, 2013b). However, when higher moments of the profit distribution influence managers' preferences, managers may trade profit to achieve other objectives involving risk, say, value maximization. The model treats the choice of risk as endogenous.

Note, though, that the other objectives might reflect agency problems: Managers might take on too little risk in order to protect their jobs, or they might consume private benefits that reduce shareholder wealth. Thus, ***the utility-maximizing framework can explain inefficient as well as efficient production***. When the output vector is held constant, the utility-maximizing cost of output can be derived from the utility-maximizing input demands. This cost function accounts for the choice of whether to produce the particular output vector using a method that has lower risk and lower expected return or a method with higher risk and higher expected return (e.g., point B versus point C in Figure 1). This choice depends on differences in the value of investment opportunities. In this case, managers' ranking of production plans captures the profit and profit-risk environment they face.

[8] See Hughes, et al. (2000) for further discussion of this model.

How one gauges performance in structural models, then, depends on whether one views bank managers as ranking production plans by their first moments (i.e., minimum expected cost or maximum expected profit), or, more generally, by higher moments as well as the first moment, i.e., considerations involving risk. In the latter case, one would want to gauge the trade-offs between risk and expected return being made by banks where there is less of an agency problem between owners and managers – i.e., banks with strong corporate controls (see Hughes, Mester, and Moon, 2001). In both the structural and non-structural approaches, the performance metric and the specification of the performance equation reflect implicitly or explicitly an underlying theory of managerial behavior.

As a general specification of the structural and non-structural approaches, let y_i represent the measure of the i^{th} bank's performance. Let $\boldsymbol{z}_i$ be a vector of variables that capture key components of the i^{th} bank's technology (e.g., output levels and input prices) and let $\boldsymbol{\tau}_i$ be a vector of variables affecting the technology (e.g., the ratio of nonperforming to total loans). Jensen and Meckling (1979) add a vector, $\boldsymbol{\theta}_i$, of characteristics of the property-rights system, contracting, and regulatory environment in which the i^{th} firm operates (e.g., whether the country has a deposit insurance scheme and the degree of investor protection that exists) and a vector, $\boldsymbol{\phi}_i$, of characteristics of the organizational form and the governance and control environment of the i^{th} firm (e.g., whether the bank is organized as a mutual or stock-owned firm, the degree of product market concentration, and the number of outside directors on its board). When the sample of banks used in the estimation includes financial institutions located in environments with different property rights and contracting environments or with different governance and control structures, estimating this model permits one to investigate how these differences are correlated with differences in bank performance.

Allowing for random error, the performance equation to be estimated takes the form,

$$(1) \qquad y_i = f(\boldsymbol{z}_i, \boldsymbol{\tau}_i, \boldsymbol{\phi}_i, \boldsymbol{\theta}_i \mid \boldsymbol{\beta}) + \varepsilon_i.$$

The specification of the vectors $\boldsymbol{z}_i$ and $\boldsymbol{\tau}_i$ differs between the structural and non-structural approaches.

2.3 The Structural Approach to Bank Efficiency Measurement: Cost Minimization, Profit Maximization, X-Inefficiency, and Managerial Utility Maximization

The traditional *structural approach* usually relies on the economics of cost minimization or profit maximization, where the performance equation (1) denotes a cost function or a profit function. Occasionally, the structural performance equation denotes a production function. While estimating a production function might tell us if the firm is *technically efficient*, i.e., if managers organize production such that the firm maximizes the amount of output produced with a given amount of inputs (so that the firm is operating on its production frontier), we are more interested in *economic efficiency*, i.e., whether the firm is responding to relative prices in choosing its inputs and outputs to minimize cost and/or to maximize profit, which subsumes technical efficiency. Risk plays no explicit role in these performance functions, although some papers include one or more dimensions of risk in the estimation as control variables. See Berger and Mester, 1997 and 2003, and Mester, 2008, for further discussion. Including risk components as controls does not fully capture the trade-off between risk and expected return that banks face. While including risk, e.g., the variance of profit, in the cost function might control for the second moment of return, higher moments would not be taken into account, and these higher moments may be an important element in the bank's production decision. So the standard cost function conditioned on risk is unlikely to capture important considerations in banking production and value maximization. In addition, as discussed below, the assumptions of cost minimization and profit maximization underlying the standard structural approach have been tested and rejected by some papers in the literature. See, for example, Evanoff (1998), Evanoff, Israilevich, and Merris (1990), Hughes, et al. (1996, 2000), Hughes, Mester, and Moon (2001), and Hughes and Mester (2013b).

In the newer literature, the optimization problem is managerial utility maximization, where the manager ranks production plans not just by their first moment – expected profit – but also by higher moments, such as skewness and kurtosis risk, as well as variance risk, that characterize profit risk. The utility-maximizing cost function is derived from the profit function, conditioned on the output vector. Thus, the cost function includes arguments that characterize revenue. In Figure 1 the larger bank can

produce its scaled-up output vector with a menu of production plans that differ by their expected profit and profit risk. The utility-maximizing cost function captures the plan that maximizes managerial utility and, thus, reflects a risk-expected-return trade-off.

To specify the utility-maximizing performance equation (1), Hughes, et al. (1996, 1999, 2000) adapt the Almost Ideal Demand System to derive a *utility-maximizing* profit equation and its associated input demand equations. This profit function does not necessarily maximize profit, since it follows from managers' assessment of risk and risk's effect on asset value; it might also reflect managers' concerns about their job security. Profit maximization (cost minimization) can be tested by noting that the standard translog profit (cost) function and share equations are nested within the model and can be recovered by imposing the parameter restrictions implied by profit maximization (cost minimization) on the coefficients of this adapted system. Hughes, et al. (1996, 1999, 2000) and Hughes and Mester (2013b) test these restrictions in their applications and reject the hypothesis of profit maximization (and cost minimization).

Both newer and traditional performance functions can differ by the definition of cost they use: Accounting (cash-flow) cost excludes the cost of equity capital, while economic cost includes it. The challenge of specifying economic cost is in estimating the cost of equity capital. McAllister and McManus (1993) arbitrarily pick the required return and assume it is uniform across banks. Clark (1996) and Fiordelisi (2007) use the Capital Asset Pricing Model to estimate it. Fiordelisi (2007) describes the resulting profit function as "economic value added." Alternatively, the quantity of equity capital can be substituted for its price. In these cases of restricted cost and profit functions, the expense of equity capital is excluded from the empirical measure of cost and profit.

The traditional structural performance equation can be fitted to the data as an average relationship, which assumes that all banks are equally efficient at minimizing cost or maximizing profit, subject to random error, ε_i, which is assumed to be normally distributed. Alternatively, it can be estimated as a *frontier* to capture best observed practice and to gauge X-inefficiency, the difference between the best-observed-practice performance and achieved performance.

The literature has used four basic methods for estimating the frontier: stochastic frontier, the distribution-free approach, the thick frontier, and data envelopment analysis (DEA). Berger and Mester (1997) review the estimation methods and present evidence on scale economies, cost X-inefficiency, and profit X-inefficiency using the stochastic frontier and distribution-free methods.[9] Mester (2008) reviews the concept, measurement, and empirical literature on X-efficiency and Dijkstra (2017) catalogs many of the studies.

In the stochastic frontier method, the error term, ε_i, consists of two components: One is a two-sided random error that represents noise (v_i), and the other is a one-sided error representing inefficiency (μ_i). The stochastic frontier approach disentangles the inefficiency and random error components by making explicit assumptions about their distributions. The inefficiency component measures each bank's extra cost or shortfall of profit relative to the frontier – the best-practice performance observed in the sample. Leibenstein (1966) called the type of inefficiency that can result from poor managerial incentives or the failure of the labor market to allocate managers efficiently and weed out incompetent managers, *X-inefficiency*. Jensen and Meckling (1976) called such inefficiency *agency costs* and provided a theoretical model of managerial utility maximization to explain how, when incentives between managers and outside stakeholders are misaligned, managers may trade off the market value of their firm to enjoy more of their own private benefits, such as consuming perquisites, shirking, discriminating prejudicially, and taking too much or too little risk to enhance their control.

Let y_i denote either the cost or profit of firm *i*. The stochastic frontier gives the highest or lowest potential value of y_i given $\mathbf{z}_i$, $\boldsymbol{\tau}_i$, $\boldsymbol{\phi}_i$, and $\boldsymbol{\theta}_i$,

$$(2) \qquad y_i = F(\mathbf{z}_i, \boldsymbol{\tau}_i, \boldsymbol{\phi}_i, \boldsymbol{\theta}_i \mid \boldsymbol{\beta}) + \varepsilon_i,$$

where $\varepsilon_i \equiv \mu_i + v_i$ is a composite error term comprising v_i, which is normally distributed with zero mean, and μ_i, which is usually assumed to be half-normally distributed and negative when the frontier is fitted as

[9] Note that the literature often uses the term "best-practice performance" and sometimes calls it "potential performance." However, this is somewhat of an abuse of terms, since measured best-practice performance does not necessarily represent the best possible practice, but merely the best practice observed among banks in the sample (see Berger and Mester, 1997, and Mester, 2008).

an upper envelope in the case of a profit function and positive when the frontier is fitted as a lower envelope as in the case of a cost function. $\boldsymbol{\beta}$ are parameters of the deterministic kernel, $F(\mathbf{z}_i, \boldsymbol{\tau}_i, \boldsymbol{\phi}_i, \boldsymbol{\theta}_i | \boldsymbol{\beta})$, of the stochastic frontier. The i^{th} bank's inefficiency is usually estimated by the mean of the conditional distribution of μ_i given ε_i, i.e., $E(\mu_i|\varepsilon_i)$. The difference between best-observed-practice and achieved performance gauges managerial inefficiency in terms of either excessive cost – *cost inefficiency* – or lost profit – *profit inefficiency*. Expressing the shortfall and excess as ratios of their frontier (best observed practice) values yields profit and cost inefficiency ratios. While the fitted stochastic frontier identifies best-observed-practice performance of the banks in the sample, it cannot explain the behavior of inefficient banks. A number of papers have surveyed investigations of bank performance using these concepts: for example, Berger and Humphrey (1997), Berger and Mester (1997), and Berger (2007). A recent comprehensive survey and classification of these models is found in Dijkstra (2017).[10]

As discussed in Hughes, et al. (2000) and Mester (2008), since inefficiency is derived from the regression residual, selection of the characteristics of the banks and the environmental variables to include in the frontier estimation is particularly important. These variables define the peer group that determines best-practice performance against which a particular bank's performance is judged. If something extraneous to the production process is included in the specification, this might lead to too narrow a peer group and an overstatement of a bank's level of efficiency. Moreover, the variables included determine which type of inefficiency gets penalized. If bank location, e.g., urban versus rural, is included in the frontier, then an urban bank's performance would be judged against other urban banks but not against rural banks, and a rural bank's performance would be judged against other rural banks. If it turned out that rural banks are more efficient than urban banks, all else equal, the inefficient choice of location would not be penalized. An alternative to including the variable in the frontier regression is to measure efficiency based on a frontier in which it is omitted and then see how it correlates with efficiency. Several papers have looked at the correlations of efficiency measures and exogenous factors, including

[10] See Dijkstra (2017), Table 1, p. 29.

Mester (1993), Mester (1996), Mester (1997), and Berger and Mester (1997). Mester (1997) shows that estimates of bank cost efficiency can be biased if bank heterogeneity is ignored. See also Bos, et al. (2005) on the issue of whether certain differences in the economic environment belong in the definition of the frontier.

Since the utility-maximizing profit function explains inefficient as well as efficient production, it cannot be fitted as a frontier. To gauge inefficiency, Hughes, et al. (1996) and Hughes, Mester, and Moon (2001) estimate a best-observed-practice risk-return frontier and measure inefficiency relative to it. The estimated utility-maximizing profit function yields a measure of expected profit for each bank in the sample, and, when divided by equity capital, the expected profit is transformed into expected return on equity, $E(\pi_i/k_i)$. Each bank's expected (or predicted) return is a function of its production plan and other explanatory variables. When the estimation of the profit function allows for heteroscedasticity, the standard error of the predicted return (profit), σ_i, which is a measure of econometric prediction risk, is also a function of the production plan and other explanatory variables and varies across banks in the sample.[11] The estimation of a stochastic frontier similar to (2) gives the highest expected return at any particular risk exposure:

$$(3) \qquad E(\pi_i/k_i) = \alpha_0 + \alpha_1\sigma_i + \alpha_2\sigma i^2 - \mu_i + \nu_i ,$$

where ν_i is a two-sided error term representing noise, and μ_i is a one-sided error term representing inefficiency. A bank's *return inefficiency* is the difference between its potential return and its noise-adjusted expected return, gauged among its peers with the same level of return risk. (Note, however, that

[11] Note that the estimated profit (or return) function resembles a multi-factor model where the factors are the explanatory variables in the profit function. The regression coefficients can be interpreted as marginal returns to the explanatory variables, and the standard error of the predicted return, a function of the variance-covariance matrix of the estimated marginal returns, resembles the variance of a portfolio return. Hughes (1999) and Hughes, Mester, and Moon (2001) report that the regression of *ln*(market value of equity) on $ln(E(\pi_i/k_i))$ and $ln(\sigma_i)$ for 190 publicly traded bank holding companies has an R-squared of 0.96, which implies that the production-based measures of expected return and risk explain a large part of a bank's market value. For a regression of the market value of equity on $E(\pi_i/k_i)$ and σ_i, Hughes and Mester (2013b) report R-squared values of 0.99, 0.94, and 0.97 for samples of data from 2003, 2007, and 2010, respectively. These values of R-squared are significantly higher than those obtained by regressing the market value on the accounting net income before and after taxes.

if a bank's managers are taking too much or too little risk relative to the value-maximizing amount, this inappropriate level of risk is not taken into account by this measure of inefficiency.)

Koetter (2006) uses the model of managerial utility maximization and the associated measure of risk-return efficiency developed in Hughes, et al. (1996, 1999, 2000) to investigate the efficiency of universal banks in Germany between 1993 and 2004. Comparing the measure of return efficiency with cost and profit efficiency estimated by standard formulations, he finds evidence that *efficient* banks using a low-risk investment strategy score poorly in terms of standard profit efficiency measures, since they also expect lower profit.

Hughes, Mester, and Moon (2001) take this a step further by recognizing that the utility-maximizing choices of bank managers need not be value maximizing to the extent that there are agency problems within the firm and managers are able to pursue their own non-value-maximizing objectives. To identify the value-maximizing banks among the set of all banks, they select the quarter of banks in the sample that have the highest predicted return efficiency. These banks are the most likely group to be maximizing value or, at least, producing with the smallest agency costs. One can use this set of efficient banks to gauge characteristics of the value-maximizing production technology. For example, mean scale economies across this set of banks would indicate whether there were scale economies as banks expand output along a path that maximizes value. In contrast, mean scale economies across *all* banks would indicate whether there were scale economies as banks expand output along a path that maximizes managers' utility, but this can differ from the value-maximizing expansion path to the extent that managers are able to pursue their own objectives and these objectives differ from those of outside owners.

While the model of managerial utility maximization yields a structural utility-maximizing profit function that includes as special cases the standard maximum profit function and a value-maximizing profit function, it is, nevertheless, based on accounting measures of performance. An alternative model developed by Hughes and Moon (2003) gauges performance using the market value of assets. They develop a *utility-maximizing q-ratio function* derived from a model where managers allocate the potential (frontier) market value of their firm's assets between their consumption of agency goods (market-value

inefficiency) and the production of market value, which, given their ownership stake, determines their wealth. The utility function is defined over wealth and the value of agency goods and is conditioned on capital structure, outside blockholder ownership, stock options held by insiders, and other managerial incentive variables. The authors derive a utility-maximizing demand function for market value and for agency goods (inefficiency). Hence, their *q*-ratio equation is *structural* and, consequently, enjoys the properties of a well-behaved consumer demand function. The authors use these properties to analyze the relationship between value (or inefficiency) and the proportion of the firm owned by insiders, which is their opportunity cost of consuming agency goods.

2.4 The Non-structural Approach to Bank Efficiency Measurement

The *non-structural approach* to bank performance measurement usually focuses on achieved performance and measures y_i, in equation (1), by a variety of financial ratios, such as the return-on-assets, the return-on-equity, or the ratio of fixed costs to total costs. However, some applications have used measures of performance that are based on the market value of the firm (which inherently incorporates market-priced risk), such as Tobin's *q*-ratio (which is the ratio of the market value of assets to the book value of assets); the Sharpe ratio (which measures the ratio of the firm's expected return in excess of the risk-free return to the volatility of this excess return (where volatility is measured by the standard deviation of the excess return)); or an event study's cumulative abnormal return (CAR), which is the cumulative error terms of a model predicting banks' market return around a particular event. Other applications have measured performance by an inefficiency ratio obtained by estimating either a non-structural or structural performance equation as a frontier.

The non-structural approach focuses on the relationship of these performance measures to various bank and environmental characteristics, including the bank's investment strategy, location, governance structure, and corporate control environment. For example, the non-structural approach might investigate technology by asking how performance ratios are correlated with characteristics of the bank, such as asset acquisitions, the bank's product mix, whether the bank is organized as a mutual or stock-owned firm, and the ratio of outside to inside directors on its board. While informal and formal theories may motivate

some of these investigations, no general theory of performance provides a unifying framework for these studies.

For example, Hughes, et al. (2018) use stochastic frontier techniques to estimate banks' best-practice return on assets (ROA) achieved for a given level of ROA risk. The best-practice ROA gauges the potential return of a bank's investment opportunities, while the difference between a bank's best-practice ROA and its achieved ROA, adjusted for noise, eliminates the influence of luck (statistical noise) and gauges its systematic failure to achieve its potential return. ROA inefficiency measures a bank's effectiveness in exploiting its investment opportunities and adds an important performance measure to the standard measures of achieved ROA and risk-normalized ROA.

Assaf, et al. (2018) estimate cost and profit functions and use the distribution-free approach to gauge cost and profit efficiency in years preceding financial crises in the U.S. economy. They then ask how these measures of efficiency are related to performance during a financial crisis. They find that cost efficiency in periods of normal economic activity is related to improved ROA and return on equity (ROE) and reduced return risk and risk of failure during crises. However, profit efficiency yields limited benefits, perhaps because it could reflect higher returns from riskier investments.

Fiordelisi, Marques-Ibanez, and Molyneux (2011) investigate the intertemporal relationship of cost and revenue efficiency, capital, and risk in European banks during the period 1995-2007. Their evidence suggests that lower efficiency is associated with higher future bank risk and that increases in capital are related to future cost efficiency improvements. Moreover, higher efficiency is associated with higher future capitalization.

In an innovative study of bank performance, Egan, Lewellen, and Sunderam (2017) develop two measures of bank productivity: one focused on deposit-taking and the other on the returns generated by a bank's asset allocation. They study how the market-to-book ratio is related to the two measures and find that deposit productivity is associated with the majority of variation in bank value. Moreover, they find synergies between deposit-taking and lending. High deposit productivity is associated with high asset

productivity. These results are consistent with those of Mester, Nakamura, and Renault (2007), who find synergies between the liability and asset sides of a commercial bank's balance sheet.

Using the frontier methods in a non-structural approach, Hughes, et al. (1997) proposed a proxy for Jensen and Meckling's agency cost: a frontier of the market value of assets fitted as a potentially nonlinear function of the book-value investment in assets and the book value of assets squared. This frontier gives the highest potential value *observed* in the sample for any given investment in assets. For any bank, the difference between its highest potential value and its noise-adjusted achieved value represents its lost market value – a proxy for agency cost (*X*-inefficiency). Several studies have used either this systematic lost market value or the resulting noise-adjusted *q*-ratio to measure performance: Baele, De Jonghe, and Vander Vennet (2006), Hughes, et al. (2003), De Jonghe and Vander Vennet (2005), Hughes and Moon (2003), Hughes, et al. (1999), Hughes, Mester, and Moon (2001), Hughes and Mester (2013a and 2013b), Hughes, Mester, and Moon (2016).

Habib and Ljungqvist (2005) specified an alternative market-value frontier as a function of a variety of managerial decision variables, including size, financial leverage, capital expenditures, and advertising expenditures. Thus, the peer grouping on which the frontier is estimated is considerably narrower than the wide grouping based on investment in assets, and inefficient choices of these conditioning values are not accounted for in the measurement of agency costs.

2.5 Specifying Outputs and Inputs in Structural Models of Production

In estimating the standard cost or profit function or the managerial utility maximization model, one must specify the outputs and inputs of bank production. The intermediation approach (Sealey and Lindley, 1977) focuses on the bank's production of intermediation services and the total cost of production, including both interest and operating expenses. Outputs are typically measured by the dollar volume of the bank's assets in various categories. As mentioned above, an exception is Mester (1992), who, to account for the bank's screening and monitoring activities, measured outputs as loans previously purchased (which require only monitoring), loans currently originated for the bank's own portfolio, loans

currently purchased, and loans currently sold. Inputs are typically specified as labor, physical capital, deposits and other borrowed funds, and, in some studies, equity capital.

While the intermediation approach treats deposits as inputs, there has been some discussion in the literature about whether deposits should be treated as an output, since banks provide transactions services for depositors. Hughes and Mester (1993) formulated an empirical test for determining whether deposits act as an input or an output. Consider variable cost, *VC*, which is the cost of nondeposit inputs and is a function of the prices of nondeposit inputs, $\boldsymbol{w}$, output levels, $\boldsymbol{y}$, other variables affecting the technology, $\boldsymbol{\tau}$, and the level of deposits, *x*. If deposits are an input, then $\partial VC/\partial x < 0$: Increasing the use of some input should decrease the expenditures on other inputs. If deposits are an output, then $\partial VC/\partial x > 0$: Output can be increased only if expenditures on inputs are increased. Hughes and Mester's empirical results indicate that insured and uninsured deposits are inputs at banks in all size categories.

2.6 Specifying Capital Structure in Performance Equations

Typically, cost and profit functions are measured without considering the bank's capital structure, which results in a seriously mis-specified model that omits an important funding input: equity capital. However, the newer literature recognizes the importance of bank managers' choice of risk and capital structure to bank performance. Some of the first structural models to include equity capital as an input are Hancock (1985, 1986), McAllister and McManus (1993), Hughes and Mester (1993), Clark (1996), and Berger and Mester (1997).

As discussed in Hughes and Mester (1993), Berger and Mester (1997), Hughes (1999), and Mester (2008), a bank's insolvency risk depends not only on the riskiness of its portfolio but also on the amount of financial capital it has to absorb losses. Insolvency risk affects bank costs and profits through (1) the risk premium the bank has to pay for uninsured debt, (2) the intensity of risk management activities the bank undertakes, and (3) the discount rate applied to future profits. A bank's capital level also directly affects costs by providing an alternative to deposits as a funding source for assets.

Most studies use the cash-flow (accounting) concept of cost, which includes the interest paid on debt (deposits) but not the required return on equity, as opposed to economic cost, which includes the cost

of equity. Failure to include equity capital among the inputs can bias efficiency measurement. If a bank were to substitute debt for some of its financial equity capital, its accounting (cash-flow) costs could rise, making the less capitalized bank appear to be more costly than the more well-capitalized bank. To solve this problem, one can include the level of equity capital as a quasi-fixed input in the cost function. The resulting cost function captures the relationship of cash-flow cost to the level of equity capital, and the (negative) derivative of cost with respect to equity capital – the amount by which cash-flow cost is reduced if equity capital is increased – gives the shadow price of equity. The shadow price of equity will equal the market price when the amount of equity minimizes cost or maximizes profit. Even when the level of equity does not conform to these objectives, the shadow price nevertheless provides a measure of its opportunity cost. Hughes, Mester, and Moon (2001) find that the mean shadow price of equity for small banks is significantly smaller than that of larger banks. This suggests that smaller banks over-utilize equity relative to its cost-minimizing value, perhaps to protect charter value. On the other hand, larger banks appear to under-utilize equity relative to its cost-minimizing value, perhaps to exploit a deposit subsidy and the subsidy due to the too-big-to-fail doctrine. In both cases, these capital strategies, while not minimizing cost, may be maximizing value.

2.7 Specifying Output Quality in the Performance Equation

In measuring efficiency, one should control for differences in output quality to avoid labeling unmeasured differences in product quality as differences in efficiency. Controls for loan quality, e.g., nonperforming loans to total loans by loan category or loan losses, are sometimes included in the cost or profit frontier as controls (see Mester, 2008, for further discussion).[12]

[12] As discussed in Berger and Mester (1997), whether it is econometrically appropriate to include nonperforming loans or loan losses in the cost or profit function depends on the extent to which these variables can be treated as exogenous. If the main driver of losses is economic shocks (bad luck), the variables could be considered exogenous. If losses largely reflect management decisions (e.g., management is inefficient or has made a conscious decision to cut short-run expenses by cutting back on loan origination and monitoring resources), then it may not be appropriate to treat the variables as exogenous. To solve this problem, Berger and Mester (1997) use the ratio of nonperforming loans to total loans in the bank's state as the control variable. The state average would be nearly entirely exogenous to any one bank, but can control for negative shocks that affect bank output quality.

The variable, nonperforming loans, can also play a role as a quasi-fixed "input" whose quantity rather than price is included in the performance equation. As such, its "cost" – loan losses – is excluded from the performance metric, either cost or profit. Its price is the expected loan-loss rate. Hence, when the cost of nonperforming loans, i.e., loan losses, is excluded from the performance measure, a case can be made for including the level of nonperforming loans, and when the performance measure is net of loan losses, the logic suggests that the loss rate be included in the specification of the performance equation.

Hughes and Moon (2018) suggest that the relationship between bank performance and the nonperforming loan ratio can be decomposed into two parts: (1) nonperformance due to the inherent credit risk the bank assumes and (2) nonperformance due to the proficiency of the bank in assessing credit risk and monitoring loan performance. They use stochastic frontier techniques to estimate the minimum nonperformance ratio observed in the sample, conditional on the volume and composition of a bank's loans, the average contractual interest rate charged on these loans, and market conditions such as the average GDP growth rate and market concentration. This minimum ratio reflects the best-practice ratio – the ratio that a bank would experience if it were fully efficient at credit-risk evaluation and loan monitoring – and gauges the inherent credit risk of the loan portfolio. The difference between a bank's noise-adjusted observed nonperforming loan ratio and the best-practice minimum ratio reflects the elimination of the influence of luck (statistical noise) and gauges the bank's inefficiency at lending. Restricting the sample to publicly traded U.S. bank holding companies and gauging financial performance by market value at year-end 2013, they find that the ratio of nonperforming loans to total loans is, on average, negatively related to financial performance except at the largest banks. This positive association of financial performance with nonperforming loans at the largest banks can be refined by decomposing nonperformance into inherent credit risk and lending inefficiency: The results suggest that taking more inherent credit risk enhances market value at a larger group of large banks, not just the largest ones, while lending inefficiency is negatively related to market value at all banks. Thus, market discipline appears to reward riskier lending at large banks and discourage lending inefficiency at all banks. Hughes, et al. (2018) use this technique to compare the business lending and commercial real estate lending

performance of banks in three size groups: banks with assets of less than $1 billion (small community banks), banks with assets between $1 billion and $10 billion (large community banks), and banks with assets between $10 billion and $50 billion (midsize banks). They find that for business lending and commercial real estate lending, large community banks and midsize banks assume higher inherent credit risk and exhibit more efficient lending. They also find that, unlike small community banks, large community banks have financial incentives to increase lending to small businesses.

3 Applications of the Structural Approach

3.1 Performance in Relation to Organizational Form, Governance, Regulation, and Market Discipline

An increasing number of papers using structural models are exploring the importance of governance and ownership structure to the performance of banks. The structural model is first used to obtain a frontier-based measure of inefficiency. Then inefficiency is regressed on a set of explanatory variables.

Using confidential regulatory data on small, closely held commercial banks, DeYoung, Spong, and Sullivan (2001) use a stochastic frontier to measure banks' profit efficiency. They find banks that hire a manager from outside the group of controlling shareholders perform better than those with owner-managers; however, this result depends on motivating the hired managers with sufficient holdings of stock. They calculate an optimal level of managerial ownership that minimizes profit inefficiency. Higher levels of insider holdings lead to entrenchment and lower profitability.

Berger and Hannan (1998) consider the relationship of bank cost efficiency, estimated by a stochastic frontier, to product market discipline, gauged by a Herfindahl index of market power. They find that the reduced discipline of concentrated markets is associated with a loss of cost efficiency far more significant than any welfare loss due to monopoly pricing.

DeYoung, Hughes, and Moon (2001) use the model of managerial utility maximization developed by Hughes, et al. (1996, 2000) to estimate expected return and return risk. Using these values, they

estimate a stochastic risk-return frontier as in equation (3) to obtain each bank's return inefficiency. They consider how banks' supervisory CAMEL ratings are related to their size, their risk-return choice, and their return inefficiency. They find that the risk-return choices of efficient banks are not related to their supervisory rating, while the higher-risk choices of inefficient banks are penalized with poorer ratings. Moreover, the risk-return choices of large inefficient banks are held to a stricter standard than smaller banks and large efficient banks.

Two studies by Mester (1991, 1993) investigate differences in scale and scope measures for stock-owned and mutual savings and loans by estimating average cost functions. She finds evidence of agency problems at mutual S&Ls, as evidenced by diseconomies of scope, prior to the industry's deregulation, and evidence that these agency costs were lessened after the deregulation in the mid-1980s.

Using data for the period 1989-1996, Altunbas, Evans, and Molyneux (2001) estimate separate and common frontiers for three organizational forms in German banking: private commercial, public (government-owned) savings, and mutual cooperative banks. They argue that the same technology of intermediation is available to all so that the choice of technology is a management decision whose efficiency should be compared among all types of firms. The private sector appears to be less profitable and less cost efficient than the other two sectors. These results are especially clear in the case of the common frontier, but they are also obtained from the estimation of separate frontiers.

3.2 Uncovering Evidence of Scale Economies by Accounting for Risk and Capital Structure

Former Federal Reserve Chairman Alan Greenspan (2010) summarized the literature on scale economies in banking: "For years the Federal Reserve had been concerned about the ever larger size of our financial institutions. Federal Reserve research had been unable to find scale economies in banking beyond a modest-sized institution." But, in fact, many investigators, including some at the Fed, have found evidence of scale economies even at the largest financial institutions. This research includes, for example, Hughes, et al. (1996), Berger and Mester (1997), Hughes and Mester (1998), Hughes, Mester, and Moon (2001), Berger and Mester (2003), Bossone and Lee (2004), Feng and Serletis (2010), Wheelock and Wilson (2012), and Hughes and Mester (2013b), and Dijkstra (2017). Dijkstra (2017)

provides an extensive classification of papers investigating scale and scope economies by their techniques and findings.[13]

The Greenspan observation raises the fundamental question: Are scale economies in banking illusive or elusive? The investment strategies of many of the largest financial institutions constituted ground zero in the recent banking crisis, and their rescue under the too-big-to-fail doctrine has prompted some prominent policymakers to call for breaking up the largest banks. For example, Fisher and Rosenblum (2012) assert, "Hordes of Dodd-Frank regulators are not the solution; smaller, less complex banks are. We can select the road to enhanced financial efficiency by breaking up TBTF banks – now." Hoenig and Morris (2012) call for limiting the government safety net to the core activities of commercial banks, including lending, taking deposits, providing liquidity and credit intermediation services, and disallowing banks from doing certain non-core banking activities, including engaging in broker-dealer activities, making markets in derivatives or securities, trading derivatives and securities for their own account or their customers', or sponsoring hedge funds or private equity funds. Tarullo (2011), however, questions whether breaking up banks would lead to efficiency and suggests there is a trade-off between concerns for systemic risk and efficiency: "An additional concern would arise if some countries made the trade-off by limiting the size or configuration of their financial firms for systemic risk reasons at the cost of realizing genuine economies of scope or scale, while other countries did not. In this case, firms from the first group of countries might well be at a competitive disadvantage in the provision of certain cross-border activities." And Powell (2013) indicates that if the current regulatory reform agenda succeeds in substantially reducing the likelihood of bank failure and minimizing the externalities caused by a large bank failure, then in his view, this would be preferable to breaking up the banks, since such a break-up would "likely involve arbitrary judgments, efficiency losses, and a difficult transition."

While textbooks assert that scale economies characterize banking (e.g., Kohn, 2004, and Saunders and Cornett, 2010), these economies elude many empirical studies because the studies generally

[13]See Dijkstra (2017), Table 3, p. 38.

fail to account for the effects of endogenous risk-taking on banks' cost as bank size increases. Textbooks cite diversification as one component of the technology that generates scale economies. As discussed above, in Figure 1, the larger bank enjoys a better risk-expected-return trade-off and chooses its risk exposure on that improved frontier to maximize managerial utility, which is likely associated with *expected* shareholder value in the absence of severe agency problems. The increase in cost due to the larger output will depend on the investment strategy the larger bank chooses. For example, as a bank scales up its output and moves from point A to point A′, diversification has resulted in lower risk, and cost is likely to have increased less than proportionately than the increase in output. If risk-taking is costly, then the investment strategy at point C may result in, say, a proportional increase in cost compared to operating at point A, while the investment strategy at point D may imply a more than proportional increase in cost. Hughes (1999) contends that studies of how cost varies with output that ignore the effects of endogenous risk-taking on cost are likely to identify the technology as constant returns to scale when larger banks tend to produce at point C and as scale diseconomies when larger banks tend to produce at point D. To the extent that larger banks are generally more risky than smaller banks (Demsetz and Strahan, 1997), the naïve econometric investigation of banking cost that ignores endogenous risk-taking is likely to find that larger banks experience constant returns to scale or even scale diseconomies. Hughes, Mester, and Moon (2001) call the effect on cost from moving from point A to point A′, the *diversification effect* – diversification leads to a decline in risk for the same level of expected profit. They call the effect on cost of moving from point A′, which resulted from better scale-related diversification, to point C or D, the *risk-taking effect.*

Accounting for endogenous risk-taking – isolating the diversification effect – in estimating scale economies requires controlling for revenue as well as cost. While the traditional cost function does not incorporate any revenue terms, the utility-maximizing cost function incorporates revenue because it is derived from the utility-maximizing profit function, conditioned on the output vector, and as noted earlier, it reflects bank managers' choice of risk as well as expected return. In Figure 1, suppose that the smaller

bank chooses to produce its output vector with the investment strategy at point A and the large bank chooses to produce its output vector with the strategy at point D. Scale economies estimated in the neighborhood of point A refer to the increase in cost for a small proportional increase in outputs given the investment strategy at point A. If expanded output allows for better diversification that lowers costs for a given expected return, then the estimated scale economies would compare cost at point A to cost at point A′. In this way, it would isolate the diversification effect and avoid the bias of measuring scale economies at point D relative to point A.[14]

Hughes and Mester (2013b) estimate several traditional cost functions and the risk-return-driven cost function for U.S. bank holding companies in 2003, 2007, and 2010. In all three years, estimates derived from the traditional minimum cost functions, which do not take into account the banks' risk-expected-return choice, indicate modest scale economies or, in some cases, constant returns to scale. In contrast, the utility-maximizing cost function, which takes into account the banks' risk-expected-return choice, yields evidence of large scale economies that increase with the scale of the bank. For example, in 2007, for the smallest banks (with less than $0.8 billion in assets), estimated scale economies is 1.12, which means that a 10 percent increase in output levels is associated with an 8.8 percent increase in cost. For the largest banks (with greater than $100 billion in assets), estimated scale economies is 1.34, which means that a 10 percent increase in output levels is associated with a 7.5 percent increase in cost. Hughes and Mester (2013b) also find that the estimated expected profit and profit risk obtained from this model explain between 94 and 99 percent of the variation in U.S. bank market value between the years 2003 and 2010. Thus, these measures of performance and best-practice capture the capital market's assessment of banks' market-priced risk and performance.

[14] Demsetz and Strahan (1997) demonstrate that a larger scale of operations leads to better diversification of banking risk – in particular, bank-specific risk estimated from a multi-factor asset pricing model. To isolate this diversification effect, they regress bank-specific risk on asset size and find a small negative association. When they control for the many ways banks take risk, the relationship between risk and asset size becomes much more negative and statistically significant. They note that isolating the scale-related diversification effect requires controlling for differences in business strategies that influence risk exposure. Finding the effect of scale-related diversification on scale economies requires a similar approach to controlling for endogenous risk-taking.

Dijkstra (2017) also estimates the risk-return-driven cost function and standard formulations of the cost function for banks in countries in the Euro area for each year from 2002 through 2011 and compares estimates of scale economies based on these estimated functions. The results based on the standard formulations of the cost function show scale economies in the early part of the time period that turn into diseconomies around the start of the financial crisis in 2008. But estimates based on the risk-return-driven cost function indicate scale economies at banks of all sizes that rise throughout the sample period, with magnitudes similar to those in Hughes and Mester (2013b). In addition, Dijkstra's estimates based on the risk-return-driven cost function indicate economically significant scope economies.

Dijkstra finds that scale economies measured by the risk-return-driven cost function increase over the span of his data, 2002 through 2011. This finding is consistent with the substantial improvements in banking technology that have occurred over this period. Advances in information technology and communications have improved the efficiency of payments networks, credit evaluation, loan monitoring, risk management, and organizational control. In addition, deregulation of interstate banking in the U.S. has allowed banks to expand their scale to exploit cost economies related to scale and has improved takeover discipline in banking.[15] Dijkstra's estimates of scale economies for European banking increase from 1.160 in 2002 to 1.271 in 2011. Hughes, et al. (2000) estimate the risk-return-driven cost function and find scale economies of 1.146 for U.S. banks in 1990. Hughes, Mester and Moon (2001) estimate scale economies of 1.145 for U.S. banks in 1994. Hughes and Mester (2013b) estimate scale economies of 1.183 for U.S. banks in 2003 compared to 1.254 for U.S. banks in 2010. For the largest U.S. financial institutions with assets exceeding $100 billion, Hughes and Mester (2013b) estimate scale economies of 1.357 in 2003 and 1.432 in 2010.

Using the methods of Wheelock and Wilson (2012), Wheelock and Wilson (2017) compare estimates of scale economies for U.S. banks in 2006 to those in 2015. They find that 27 percent more banks experienced increasing returns in 2015 than in 2006 and that (p. 20) ". . . the largest four banks

[15] See Brook, Hendershott, and Lee (1998), who investigate how the relaxation of interstate banking restrictions in the U.S. improved bank performance.

have seen significant increases in returns to scale since 2006, suggesting that scale economies still provide an impetus to become even larger." These estimates of scale economies, which increase over time, are consistent with the hypothesis that technological improvements substantially reduce the average cost of financial products and services and, as Wheelock and Wilson (2017) observe, provide strong incentives to institutions to increase their scale to exploit these improving economies.[16]

Spreading overhead costs, such as those related to technology and regulatory compliance, over a larger scale is often cited as an important source of scale economies in banking. Kovner, Vickery, and Zhou (2014) investigate this issue and find evidence of operating cost economies when they control for a bank's investment strategy. Controlling for the bank's investment strategy when measuring the effect of an increase in scale on cost can be illustrated in our Figure 1. This would be equivalent to assessing the increase in cost when a bank moves from A to A′ rather than when it moves from point A to C or D, points that represent riskier investment strategies. When Kovner, Vickery, and Zhou estimate the cost elasticity without controlling for the investment strategy, they find that a 10 percent increase in assets implies a 9.93 percent increase in operating cost – essentially constant returns to scale. When they control for asset allocation, the cost elasticity falls to 9.79 percent. When they control for asset allocation, revenue sources, funding structure, and organizational complexity, the operating cost elasticity falls to 8.99 percent – evidence of scale-related operating cost economies. Moreover, operating scale economies increase with bank size, so that the largest financial institutions obtain the largest operating cost economies.

This evidence of large scale economies at the largest financial institutions suggests that breaking them up into smaller institutions with the goal of reducing the systemic risk they pose would reduce their competitiveness in global financial markets. Using their 2007 estimates, Hughes and Mester (2013b) consider breaking each of the 17 institutions that exceed $100 billion in consolidated assets in half to

[16] As described previously, cost studies based on the mis-specified cost function that fails to account for potentially costly endogenous risk-taking are likely to underestimate scale economies. Using data from the 1980s, such studies typically failed to find evidence of scale economies except at the smallest banks. On the other hand, data obtained from the 1990s were more likely to yield estimates of scale economies at larger banks. Thus, even these estimates suggest technological advances. For a detailed catalogue of these studies, see Dijkstra (2017).

create 34 banks with total assets equal to those of the 17 larger institutions. Holding product mix constant i.e., assuming the smaller institutions produce the same product mix as the larger ones, their costs are 23 percent higher. In a similar exercise, Wheelock and Wilson (2012), who also find large scale economies at banks of all sizes, scale back the four largest U.S. institutions in 2009 to a size of $1 trillion and increase their numbers so that the total assets of the smaller institutions equal those of the larger institutions. They find that the cost of the smaller institutions is approximately 19 percent higher. These two exercises suggest that breaking up the largest institutions into smaller institutions will limit their global competitiveness and provide incentives to produce their financial services offshore, where such limits are not operative.

A related issue in this literature questions whether the estimated scale economies at the largest financial institutions result from cost-of-funds subsidies due to banks being considered too big to fail. Hughes and Mester (2013b) present several pieces of evidence indicating that the large scale economies they find are not driven by a cost-of-funds subsidy derived from banks being considered too big to fail. First, they find large scale economies at small banks in their sample as well as at large banks. Second, when they re-estimate their model excluding banks with assets greater than $100 billion, and then calculate scale economies out of sample for the largest banks, their results are unchanged. Finally, they calculate scale economies for the largest banks if they faced the cost-of-funds of smaller banks. Again, their results are unchanged. Hughes and Mester (2013b) conclude that the underlying technology, not too-big-to-fail subsidies, accounts for the scale economies of the largest financial institutions. Davies and Tracey (2014) adopt the strategy used by Hughes and Mester (2013b) of replacing the observed price of borrowed funds with a price that seeks to eliminate the TBTF subsidy – one derived from a Moody's rating for each bank that assumes government assistance in financial distress. 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 too-big-to-fail subsidy. However, there is a critical difference between their methodology and that of Hughes and Mester (2013b). Hughes and Mester measure scale economies by substituting the pseudo

prices into the fitted measure of scale economies, which is derived from the 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 prices 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.

Kroszner (2016) considers differences in funding costs not only between large and small financial institutions but also between large and small firms in nonfinancial industries – industries where there is no possibility of government support during financial distress. He finds that funding costs are lower for large firms across many industries: from 84 basis points for energy to 5 basis points for utilities, with banks in the middle at 35 basis points. Using data from the period 2004 to 2013, Ahmed, Anderson, and Zarutskie (2015) find that the borrowing costs of nonfinancial firms as well as those of financial firms tend to decline with borrower size; moreover, financial firms exhibit lower borrowing costs that are less sensitive to size than those in several other industries. They suggest that size-related differences in borrowing costs may be partially influenced by higher liquidity and recovery rates of larger borrowers, rather than government subsidies related to size. Minton, Stulz, and Taboada (2017), which we discuss in the next section, study banks over the period 1987-2015 and find no evidence that larger banks benefit from too-big-to-fail subsidies.

4 Applications of the Non-Structural Approach

4.1 Measuring the Value of Investment Opportunities ("Charter Value")

The value of a bank's investment opportunities is often measured by Tobin's *q*-ratio; however, in the presence of agency costs, Tobin's *q*-ratio captures only the ability of the incumbent managers to exploit these opportunities. Ideally, the value of investment opportunities should be gauged

independently of the ability and actions of the current management. Hughes, et al. (1997) and Hughes, et al. (2003) propose a measure based on fitting a stochastic frontier to the market value of assets as a function of the book value of assets and variables characterizing the market conditions faced by banks in their local markets. These conditions include a Herfindahl index of market power and the macroeconomic growth rate. The fitted frontier gives the highest potential value of a bank's assets in the markets in which it operates. Thus, this potential value is conditional on the location of the bank and represents the value the bank would fetch in a competitive auction. Hughes, et al. (1997) define this value as the bank's "charter value" – its value in a competitive auction.

4.2 Measuring the Performance of Capital Strategies

Several papers have used the non-structural performance equation to examine the relationship between bank value and bank capital structure. Hughes, et al. (1997) regress performance measured by Tobin's *q*-ratio and market-value inefficiency on a number of variables characterizing bank production. Calomiris and Nissim (2007) regress the ratio of the market value of equity to its book value on a similar list of variables. De Jonghe and Vander Vennet (2005) apply the market-value frontier of Hughes, et al. (1997) to derive a noise-adjusted measure of Tobin's *q*, which they use to evaluate how leverage and market power are related to value. Hughes, Mester, and Moon (2016) regress performance based on the market value of assets and on the shortfall of the market value of assets, obtained from a market-value frontier, on banks' capital structure and variables that characterize banks' business strategy. All four studies find evidence that banks follow dichotomous strategies for enhancing value, as predicted by Marcus (1984): a strategy that entails lower risk and lower leverage and a strategy that entails higher risk and higher leverage.

4.3 Measuring the Performance of Business Strategies

A common result in the corporate finance literature is that nonfinancial conglomerates, which combine firms in different industries, trade at a discount. Klein and Saidenberg (2010) consider the extent to which such a diversification discount exists in commercial banking. Because banks' scope is generally limited to financial products and services, a diversification discount in banking would likely

result from organizational complexity rather than industry diversification. They use data from 1990-1994, a period before geographical restrictions were lifted by the Riegle-Neal Act when banks crossed state lines by forming a holding company to operate separate bank subsidiaries in other states. They find that bank holding companies with many subsidiaries have, on average, a lower Tobin's *q*-ratio than banks with fewer subsidiaries. They conclude that organizational complexity, in addition to scope-related diversification, may also contribute to the diversification discount reported in the corporate finance literature.

Brook, Hendershott, and Lee (1998) investigate how bank performance was affected by the Riegle-Neal Act's lifting of interstate branching restrictions, a liberalization that allowed bank holding companies to consolidate subsidiaries operating across state lines. They find evidence in support of their hypothesis that relaxation of the restrictions on interstate banking improved bank performance through better exploitation of scale economies and through enhanced market discipline due to more extensive takeover threats. Using event-study methodology, they investigate the reaction of bank stock prices to the passage of the legislation and find a statistically significant positive cumulative abnormal return (CAR). Underperforming banks whose management was least entrenched received the highest CAR.

Minton, Stulz, and Taboada (2017) investigate the relationship between bank valuation and asset size. Using Tobin's *q*-ratio and the market-to-book value of equity to measure valuation, they find a statistically insignificant relationship between bank asset size and valuation for banks with assets in the $10 billion to $50 billion range, but a significantly negative association between size and valuation for banks with assets greater than $50 billion. Their analysis suggests that this negative association is due to the volume of trading assets that a bank holds. Namely, they find that banks with more trading assets are valued less, and that larger banks tend to have larger volumes of trading assets relative to assets.

Cetorelli, Jacobides, and Stern (2017) use a new data set that details the organizational structure of U.S. bank holding companies. They map a bank's entry and exit across scope-related sectors and find that scope expansion is, on average, associated with worse financial performance measured by the Tobin *q* ratio and by ROE. Schmid and Walter (2014) adopt a strategy developed by Berger and Ofek (1995) to

look for evidence of a scope-related diversification discount in banking. They construct a measure of excess value that compares a bank's value to its value were its segments operating as standalones. They define these segments by categorizing a bank's assets into distinct industry segments. The value of each of these segments is given by the median ratio of the market value of assets per dollar of assets for banks that operate solely in a segment. The imputed value of a bank that operates in multiple segments were its assets divided into stand-alone segments is computed as the asset-weighted sum of values across the segments in which the bank holds assets. The authors find that, on average, the imputed stand-alone value of the portfolio is greater than that of the diversified firm. The diversification discount is approximately 11 percent in 2005 but statistically insignificant throughout the ensuing financial crisis.

The textbook definition of scope economies compares the cost of producing an output vector in a single multi-product firm with the cost of producing it in separate single-product firms. Mester (1991) proposes a technique of estimating scope economies from the multi-product cost estimation that does not require finding banks in the sample that produce only one type of output or setting the value of a particular type of output to zero in the estimated cost function when no firms in the sample produce at this level. Essentially, this technique indicates whether the marginal cost of a particular output increases or decreases with another output. If it decreases, there are scope economies between the two outputs. Hughes and Mester (1993) apply this technique to U.S. data and do not find evidence of overall scope economies. Dijkstra (2017) applies it to Euro-zone banks and finds evidence of scope economies.

4.4 Relationship of Ownership Structure to Bank Value

In an influential study, Morck, Shleifer, and Vishny (1988) hypothesized that managerial ownership creates two contrasting incentives: A higher ownership stake, first, better aligns the interests of managers and outside owners and, second, enhances managers' control over the firm and makes it harder for managers to be ousted when they are not efficient. Measuring performance by Tobin's *q*, these authors provide evidence that the so-called alignment-of-interests effect dominates the entrenchment effect at lower levels of managerial ownership, while the entrenchment effect dominates over a range of higher levels.

Studies that attempt to measure the *net* effect of the alignment and entrenchment effects on firm valuation cannot identify these effects individually – only their sum in the form of the sign of a regression coefficient or a derivative of a regression equation. Adams and Santos (2006) cleverly isolate the entrenchment effect by considering how the proportion of a bank's common stock that is controlled but not owned by the bank's own trust department is statistically related to the bank's economic performance. The voting rights exercised by management through the trust department enhance management's control over the bank but do not align their interests with outside shareholders' because the beneficiaries of the trusts, not the managers, receive the dividends and the capital gains and losses.

Caprio, Laeven, and Levine (2003) study the effect of ownership, shareholder protection laws, and supervisory and regulatory policies on the valuations of banks around the world. The authors construct a database of 244 banks across 44 countries. They measure performance by Tobin's *q*-ratio and by the ratio of the market value of equity to the book value of equity. They find evidence that (1) banks in countries with better protection of minority shareholders are more highly valued, (2) bank regulations and supervision have no significant effect on bank value, (3) the degree of cash-flow rights of the largest owner has a significant positive effect on bank value, and (4) an increase in ownership concentration has a larger positive effect on valuation when the legal protection of minority shareholders is weak.

Laeven and Levine (2009) consider a sample of large banks in 48 countries in 2001 and investigate how the cash-flow rights of the largest shareholder and various regulatory provisions affect the probability of insolvency. They find that the cash-flow rights of the largest shareholder are positively related to the risk of insolvency. They also find that when there is a shareholder with large cash-flow rights, deposit insurance and activity restrictions are associated with increased insolvency risk, but they are uncorrelated with insolvency risk when the bank is widely held.

Hughes, et al. (2003) examine U.S. bank holding companies and find evidence of managerial entrenchment among banks with higher levels of insider ownership, more valuable growth opportunities, poorer financial performance, and smaller asset size. When managers are not entrenched, asset acquisitions and sales are associated with reduced market value inefficiency. When managers are

entrenched, sales are associated with smaller reductions in inefficiency, while acquisitions are associated with greater inefficiency.

5 Conclusions and Policy Implications

Great strides have been made in the theory of bank technology in terms of explaining banks' comparative advantage in producing informationally intensive assets and financial services and in taking, diversifying, and offsetting a variety of risks. Great strides have also been made in explaining sub-par managerial performance in terms of agency theory and in applying these theories to analyze the particular environment of banking. In recent years, the empirical modeling of bank technology and the measurement of bank performance have begun to incorporate these theoretical developments and yield interesting insights that reflect the unique nature and role of banking in modern economies.

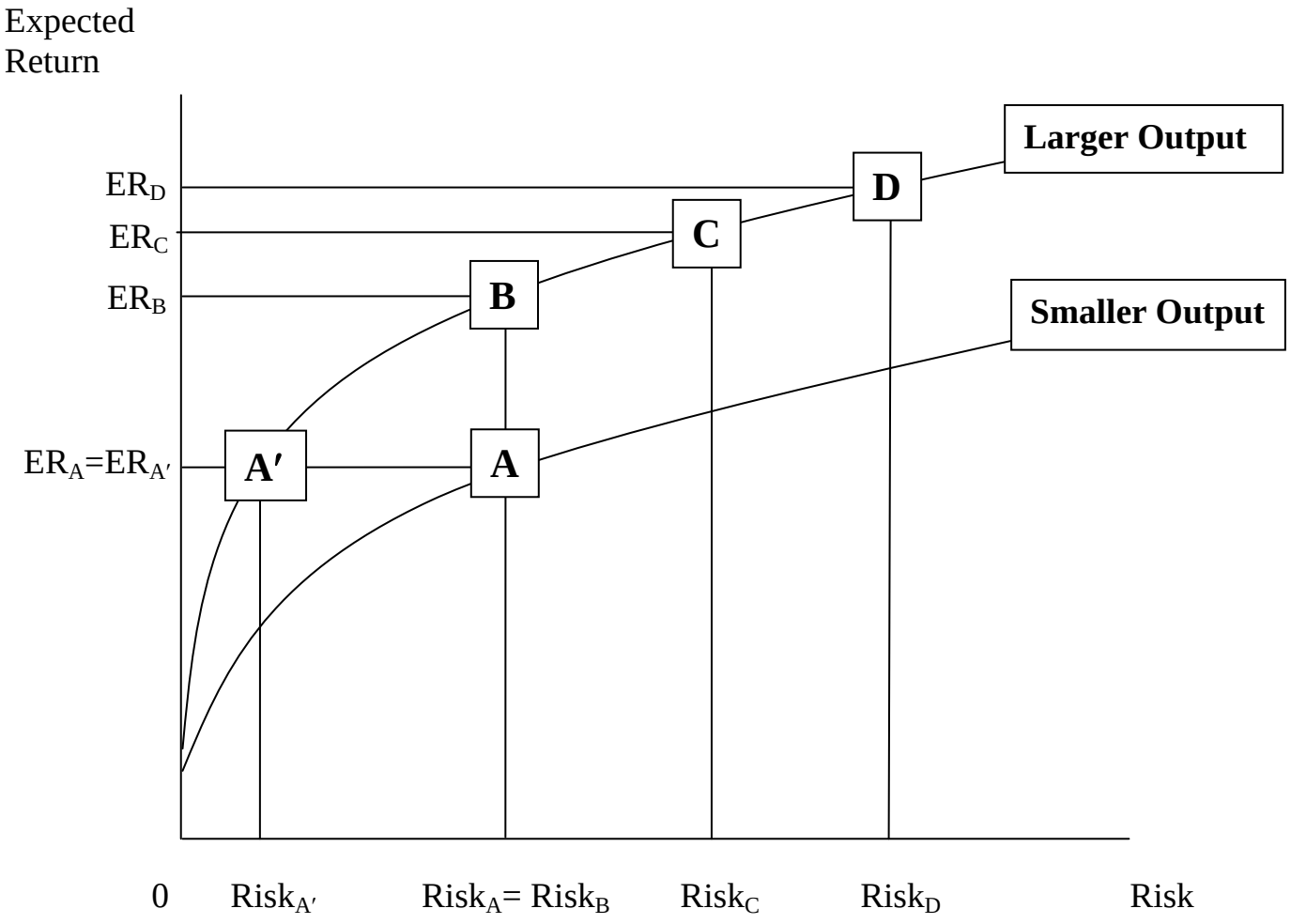
This new literature recognizes that the choice of risk influences banks' production decisions (including their mix of assets, asset quality, off-balance-sheet hedging activities, capital structure, debt maturity, and resources allocated to risk management), and so, in turn, affects banks' costs and profitability. Measures of bank performance should take account of this endogeneity. The estimation of structural models that incorporate managerial preferences for expected return and risk has uncovered significant scale economies in banking that increase with bank scale, a finding that differs from the earlier literature but accords with the consolidation of the banking industry that has been occurring worldwide. This finding of significant scale economies at the largest financial institutions also suggests that proposals to break up these institutions into smaller institutions in an attempt to ameliorate too-big-to-fail problems could limit their global competitiveness and provide them with incentives to produce their financial services offshore where such limits are not operative. Moreover, banks' strategies to avoid restrictions on scale, such as moving such activities into the less-regulated nonbank sector, may create new sources of systemic risk that are harder to supervise and regulate.

Performance studies based on structural models of managerial utility maximization, as well as those based on non-structural models of bank production, have incorporated variables designed to capture

incentive conflicts between managers and outside stakeholders. These studies have shown that factors associated with enhanced market discipline are also associated with improved bank performance and that improved bank performance is not necessarily associated with improved financial stability when the improved performance results from investment strategies that increase financial leverage and heighten credit risk. In short, the incentive of larger banks to take extra risk to exploit the federal safety net and increase their expected market value may undermine financial stability. These results suggest an important role for capital regulation and enhanced supervision of large financial institutions.

Figure 1

Scale-Related Diversification and Risk-Return Frontiers



Source: Hughes and Mester (2013b)

Bibliography

Adams, R.B. and Santos, J.A.C. (2006). Identifying the effect of managerial control on firm performance, *Journal of Accounting and Economics* **41**, 55-85.

Aghion, P., Alesina, A. and Trebbi, F. (May 2007). Democracy, technology, and growth, Working Paper, Department of Economics, Harvard University.

Ahmed, J.I., Anderson, C. and Zarutskie, R.E. (2015). Are the borrowing costs of large financial firms unusual? Finance and Economics Discussion Series 2015-024. Washington: Board of Governors of the Federal Reserve System, http://dx.doi.org/10.17016/FEDS.2015.024.

Altunbas, Y., Evans, L. and Molyneux, P. (2001). Bank ownership and efficiency, *Journal of Money, Credit, and Banking* **33**, 926-954.

Assaf, A., Berger, A.N., Roman, R.A. and Tsionas, M. (2018). Does efficiency help banks survive and thrive during financial crises? Working Paper.

Baele, L., De Jonghe, O. and Vander Vennet, R. (August 2006). Does the stock market value bank diversification? Working Paper No. 2006/402, Department of Financial Economics, Ghent University.

Berlin, M. and Mester, L.J. (1999). Deposits and relationship lending, *Review of Financial Studies* **12,** 579-607.

Berger, A.N. (2007). International comparisons of banking efficiency, *Financial Markets, Institutions and Instruments* **16**, 119–144.

Berger, A.N. and Hannan, T.H. (1998). The efficiency cost of market power in the banking industry: A test of the 'quiet life' and related hypotheses, *Review of Economics and Statistics* **80**, 454-465.

Berger, A.N. and Humphrey, D. B. (1997). Efficiency of financial institutions: International survey and directions for future research, *European Journal of Operational Research* **98**, 175-212.

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.

Berger, A.N. and Mester, L.J. (2003). Explaining the dramatic changes in performance of U.S. banks: Technical change, deregulation, and dynamic changes in competition, *Journal of Financial Intermediation* **12**, 57-95.

Berger, P.G. and Ofek, E. (1995). Diversification's effect on firm value, *Journal of Financial Economics* **37**, 39-65.

Bhattacharya, S. and Thakor, A. (1993). Contemporary banking theory, *Journal of Financial Intermediation* **3**, 2-50.

Bos, J.W.B., Heid, F., Koetter, M., Kolari, J.W., and Kool, C.J.M. (2005). Inefficient or just different? Effects of heterogeneity on bank efficiency scores, Deutsche Bundesbank Discussion Paper No. 2.

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

Brook, Y., Hendershott, R. and Lee, D. (1998). The gains from takeover deregulation: Evidence from the end of interstate banking restrictions, *Journal of Finance* **53**, 2185-2204.

Calomiris, C.W. and Kahn, C. M. (1991). The role of demandable debt in structuring optimal banking arrangements, *American Economic Review* **70**, 312-326.

Calomiris, C.W. and Nissim, D. (2007). Activity-based valuation of bank holding companies, Working Paper 12918, National Bureau of Economic Research.

Caprio, G., Laeven, L. and Levine, R. (2003). Governance and bank valuation, Working Paper 10158, National Bureau of Economic Research.

Cetorelli, N., Jacobides, M.G. and Stern, S. (2017). Transformation of corporate scope in U.S. banks: Patterns and performance implications, Staff Report no. 813, Federal Reserve Bank of New York (May).

Cheng, I.-H., Hong, H. and Schneinkman, J.A. (2015). Yesterday's heroes: Compensation and risk at financial firms, *Journal of Finance* **70**, 839-879.

Clark, J. (1996). Economic cost, scale efficiency and competitive viability in banking, *Journal of Money, Credit, and Banking* **28,** 342-364.

Davies, R. and Tracey, B. (2014). Too big to be efficient? The impact of too-big-to-fail factors on scale economies for banks, *Journal of Money, Credit, and Banking* **46**, 219-253.

De Jonghe, O. and Vander Vennet, R. (2005). Competition versus agency costs: An analysis of charter values in European banking, Working Paper, Ghent University.

Demirgüç-Kunt, A., Kane, E.J. and Laeven, L. (January 2007). Determinants of deposit-insurance adoption and design, NBER Working Paper No. 12862.

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.

DeYoung, R.E., Hughes, J.P. and Moon, C.-G. (2001). Efficient risk-taking and regulatory covenant enforcement in a deregulated banking industry, *Journal of Economics and Business* **53**, 255-282.

DeYoung, R., Spong, K. and Sullivan, R.J. (2001). Who's minding the store? Motivating and monitoring hired managers at small, closely held commercial banks, *Journal of Banking and Finance* **25**, 1209-1243.

Dijkstra, M.A. (2017). *Economies of Scale and Scope in Banking: Effects of Government Intervention, Corporate Strategy and Market Power,* Amsterdam University Press.

Egan, M., Lewellen, S. and Sunderam, A. (2017). The cross section of bank value, NBER Working Paper No. 23291.

Evanoff, D.D. (1998). Assessing the impact of regulation on bank cost efficiency, *Economic Perspectives,* Federal Reserve Bank of Chicago, **22,** 21-32.

Evanoff, D.D., Israilevich, P.R., and Merris, R.C. (1990). Relative price efficiency, technical change, and scale economies for large commercial banks, *Journal of Regulatory Economics*, **2**, 281-298.

FDIC. (2014). TBTF subsidy for large banks – Literature review (updated), Prepared for Thomas Hoenig, Vice Chair, Federal Deposit Insurance Corporation.

Feng, G. and Serletis, A. (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.

Fiordelisi, F. (2007). Shareholder value efficiency in European banking, *Journal of Banking and Finance* **31**, 2151-2171.

Fiordelisi, F., Marques-Ibanez, D. and Molyneux, P. (2011). Efficiency and risk in European banking, *Journal of Banking and Finance* **35**, 1315-1326.

Fisher, R. and Rosenblum, H. (2012). How huge banks threaten the economy, *Wall Street Journal*, April 4, 2012.

Flannery, M.J. (1994). Debt maturity and the deadweight cost of leverage: Optimally financing banking firms, *American Economic Review* **84**, 320-331.

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

Habib, M.A. and Ljungqvist, A. (2005). Firm value and managerial incentives: A stochastic frontier approach, *Journal of Business* **78**, 2053-2093.

Hancock, D. (1985). The financial firm: Production with monetary and nonmonetary goods, *Journal of Political Economy* **93**, 859-880.

Hancock, D. (1986). A model of the financial firm with imperfect asset and deposit liabilities, *Journal of Banking and Finance* **10**, 37-54.

Herring, R. J. and Vankudre, P. (1987), Growth opportunities and risk-taking by financial intermediaries, *Journal of Finance* **42**, 583-599.

Hoenig, T.M. and Morris, C.S. (2012). Restructuring the banking system to improve safety and soundness, manuscript, Federal Deposit Insurance Corporation.

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

Hughes, J.P., Jagtiani, J., Mester, L.J. and Moon C.-G. (2018). Does scale matter in community bank performance? Evidence obtained by applying several new measures of performance, Working Paper 18-11, Federal Reserve Bank of Philadelphia (March).

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., Mester, L.J. and Moon C.-G. (1999). The dollars and sense of bank consolidation, *Journal of Banking and Finance* **23**, 291-324.

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., Moon C.-G. and Pagano, M. (2003). Do bankers sacrifice value to build empires? Managerial incentives, industry consolidation, and financial performance, *Journal of Banking and Finance* **27**, 417-447.

Hughes, J.P., Lang, W., Moon C.-G. and Pagano, M. (1997). Measuring the efficiency of capital allocation in commercial banking, Working Paper 98-2, Federal Reserve Bank of Philadelphia (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. (1993). A quality and risk-adjusted cost function for banks: Evidence on the 'too-big-to-fail' doctrine, *Journal of Productivity Analysis* **4**, 293-315.

Hughes, J.P. and Mester, L.J. (2013a). A primer on market discipline and governance of financial institutions for those in a state of shocked disbelief, Chapter 2 in *Efficiency and Productivity Growth: Modelling in the Financial Services Industry*, ed. Pasiouras, F., John Wiley and Sons: West Sussex, U.K., pp. 19-47.

Hughes, J.P. and Mester, L.J. (2013b). Who said large banks don't experience scale economies? Evidence from a risk-return-driven cost function, *Journal of Financial Intermediation* **22**, 559-585.

Hughes, J. P., Mester, L.J. and Moon, C.-G. (2016). The two faces of equity capital in U.S. commercial banking: Market discipline working for and against financial stability, Department of Economics, Rutgers University, Working Paper 201611.

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.

Hughes, J.P. and Moon C.-G. (2003). Estimating managers' utility-maximizing demand for agency goods, Working Paper 2003-24, Department of Economics, Rutgers University.

Hughes, J. P. and Moon, C.-G. (2018). How bad is a bad loan? Distinguishing inherent credit risk from inefficient lending (Does the capital market price this difference?), Department of Economics, Rutgers University, Working Paper 201802.

Jensen, M.C. and Meckling, W.H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure, *Journal of Financial Economics* **5**, 305-360.

Jensen, M.C. and Meckling, W.H. (1979). Rights and production functions: An application to labor-managed firms and codetermination, *Journal of Business* **52**, 469-506.

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

Klein, P.G. and Saidenberg, M.R. (2010). Organizational structure and the diversification discount: Evidence from commercial banking, *Journal of Industrial Economics* **58**, 127-155.

Koetter, M. (2006). The stability of efficiency rankings when risk-preferences and objectives are different, Discussion Paper 08/2006, Series 2: Banking and Financial Studies, Deutsche Bundesbank.

Kohn, M. (2004). *Financial Institutions and Markets* (Oxford: Oxford University Press).

Kovner, A., Vickery, J. and Zhou, L. (2014). Do big banks have lower operating costs? *Economic Policy Review*, Federal Reserve Bank of New York, December, 1–27.

Kroszner, R. (2016). A review of bank funding cost differentials, *Journal of Financial Services Research* **49**, 151-174.

La Porta, R., Lopez-de-Silanes, F. and Shleifer, A. (2002). Government ownership of banks, *Journal of Finance* **57**, 265-301.

Leibenstein, H. (1966). Allocative efficiency vs. 'X-efficiency,' *American Economic Review* **5**6, 392-415.

Laeven, L. and Levine, R. (2009). Bank governance, regulation, and risk taking. *Journal of Financial Economics* **93**, 259–275.

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

McAllister, P.H. and McManus, D. (1993). Resolving the scale efficiency puzzle in banking, *Journal of Banking and Finance* **17**, 389-406.

Mester, L.J. (2008). Optimal industrial structure in banking, Chapter 5 in *Handbook of Financial Intermediation and Banking*, Boot, A. and Thakor, A. (eds.) Amsterdam: North-Holland/Elsevier, 133-162.

Mester, L.J. (First and Second Quarters 2007). Some thoughts on the evolution of the banking system and the process of financial intermediation, Federal Reserve Bank of Atlanta *Economic Review*, 67-75.

Mester, L.J. (1997). Measuring efficiency at U.S. banks: Accounting for heterogeneity is important, *European Journal of Operational Research* **98**, 230-242.

Mester, L.J. (1996). A study of bank efficiency taking into account risk-preferences, *Journal of Banking and Finance* **20**, 1025-1045.

Mester, L.J. (1993). Efficiency in the savings and loan industry, *Journal of Banking and Finance*, **17**, 267-286.

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

Mester, L.J. (1991). Agency costs among savings and loans, *Journal of Financial Intermediation* **1**, 257-278.

Mester, L.J., Nakamura, L.I. and Renault, M. (2007). Transactions accounts and loan monitoring, *Review of Financial Studies* **20**, 529-556.

Minton, B.A., Stulz, R.M. and Taboada, A.G. (2017). Are larger banks valued more highly? Fisher College of Business Working Paper Series, The Ohio State University, WP 2017-08.

Morck, R., Shleifer, A. and Vishny, R.W. (1988). Management ownership and market valuation: An empirical analysis, *Journal of Financial Economics* **20**, 293-315.

Powell, J.H. (2013). Ending “too big to fail,” Remarks at the Institute of International Bankers 2013 Washington Conference, Washington, DC.

Saunders, A., and Cornett, M. (2010). *Financial Institutions Management: A Risk Management Approach* (New York: McGraw-Hill Higher Education).

Schmid, M. and Walter, I. (2014). Firm structure in banking and finance: Is broader better? *Journal of Financial Perspectives* **2**, 65-75.

Sealey, C.W. and Lindley, J.T. (1977). Inputs, outputs, and a theory of production and cost at depository financial institutions, *Journal of Finance*, **32,** 1251-1266.

Tarullo, D. K. (2011). Industrial organization and systemic risk: An agenda for further research, Remarks at the Conference on the Regulation of Systemic Risk, Federal Reserve Board, Washington, DC.

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.

Wheelock, D. and Wilson, P. (2017). The evolution of scale economies in U.S. banking, Working Paper 2015-021C, Federal Reserve Bank of St. Louis.