

Wharton

Financial
Institutions
Center

*Banking Markets: Productivity,
Risk, and Customer Satisfaction*

by
Gerald R. Faulhaber

95-14

THE WHARTON FINANCIAL INSTITUTIONS CENTER

The Wharton Financial Institutions Center provides a multi-disciplinary research approach to the problems and opportunities facing the financial services industry in its search for competitive excellence. The Center's research focuses on the issues related to managing risk at the firm level as well as ways to improve productivity and performance.

The Center fosters the development of a community of faculty, visiting scholars and Ph.D. candidates whose research interests complement and support the mission of the Center. The Center works closely with industry executives and practitioners to ensure that its research is informed by the operating realities and competitive demands facing industry participants as they pursue competitive excellence.

Copies of the working papers summarized here are available from the Center. If you would like to learn more about the Center or become a member of our research community, please let us know of your interest.

Anthony M. Santomero
Director

*The Working Paper Series is made possible by a generous
grant from the Alfred P. Sloan Foundation*

Banking Markets:
Productivity, Risk and Customer Satisfaction ¹

Abstract: A structural model is developed which incorporates bank decisions on productivity, risk taking and customer satisfaction into an equilibrium model of banking markets. This structural model is estimated directly for 219 large US banks, 1984-1992. The results are: (i) banks differ widely in their ability to manage risk; (ii) there are substantial inefficiencies due to demand/capacity mismatches; (iii) greater customer satisfaction correlates with greater profitability, principally due to higher levels of demand; (iv) very large bank-specific effects that previous research discovered appear to have been largely captured in the structural model.

Key Words: banking, structural estimation, quality, productivity, efficiency, risk

Gerald R. Faulhaber is Professor of Management and of Public Policy and Management at the Wharton School. This research was supported by a grant from the Wharton Financial Institutions Center funded by the Sloan Foundation and Deloitte & Touche. The author wishes to thank Larry Brown, Kermit Daniel and Dan Raff, as well as participants in the Financial Institutions Center seminars for their comments and suggestions. Jalal Akhavein and Tony Cai provided invaluable research assistance.

BANKING MARKETS: PRODUCTIVITY, RISK, AND CUSTOMER SATISFACTION

1. Introduction

The past fifteen years have seen substantial changes in banking markets. Advances in computing and telecommunications have driven these global changes, but they have been most pronounced in US markets. A voluminous literature has developed, in part as a result of these changes, which has generally addressed two issues: (i) bank productivity, including economies of scale, economies of scope, and X-efficiency; (ii) bank risk-taking and its costs. Are banks the right size? Do they have the appropriate product scope? Are they using “best practice” techniques? How does size affect the efficiency of risk-taking? For a definitive review of this work, the reader is referred to Berger, Hunter, and Timme (1993). Studies of economies of scale and scope have dominated the literature. The consensus of the literature is that for large banks, scale economies are largely nonexistent and scope economies are small, a result confirmed in this paper. Bank risk was examined in McAllister and McManus (1993), where an inventory model to assess scale economies associated with risk pooling. Hughes and Mester (1994) model bank managers’ risk preferences, using the leverage ratio as their risk measure. Additionally, Hughes and Mester (1993) assess the impact of the “too-big-to-fail” doctrine on bank risk-taking, a matter we consider in this paper.

More recent innovations in the literature have focused on (i) the use of the profit function rather than the cost function, to help identify both demand side and supply side effects (see Berger, Hancock, and Humphrey (1993)); and (ii) the use of frontier methods to identify possible X-inefficiency. A series of papers has explored various methods for frontier estimation: data envelopment analysis, efficient frontier analysis, “thick” frontier analysis, and “distribution-free analysis”. Berger (1993) provides a discussion and analysis of these various methods and their differences in the context of banking. Kaparakis, Miller, and Noulas (1994) use a stochastic frontier approach and find that bank size, in particular the number of branches, has a negative impact on short-run efficiency. Mester (1994) uses cost frontier methods to assess the efficiency of banks in the Federal Reserve’s third district, finding significant X-inefficiencies. Grabowski, Rangan, and Rezvanian (1994) use a nonparametric frontier approach to assess the effect of deregulation on bank efficiency, finding a decline over the years 1979-1987, principally focused on technical rather than allocative inefficiency. Overall, these papers have identified potential X-inefficiencies as substantially greater than either scale or scope effects, and have thus turned scholarly

attention toward identifying possible sources for these large efficiency differences among banks. For example, Mester (1993) has examined efficiency differences between stock and mutual savings and loan institutions, finding that the mutual form of ownership is associated with greater efficiency.

In this paper, three questions are addressed: (i) the cost of bank risk is assessed, using banks' stock market betas as a broad measure of total risk². In particular, the effect of size on risk-taking is analyzed, as well as the decomposition of risk on a product basis; (ii) productivity losses due to mismatches between demand and capacity is modeled and estimated; (iii) the effect of customer satisfaction, or quality, on bank profitability is measured. Each of these issues is new (in varying degrees) to the literature. Additionally, the paper provides strong confirmation of previous research results on economies of scale and scope, and it provides significant new evidence on the X-efficiency/frontier analysis issue.

There are several features of this paper, other than the questions addressed, that make this paper unique:

- Most of the previous empirical work employed reduced form estimation: a cost function or a profit function is specified and its parameters are estimated. In this paper, a full structural model is developed: banks face long- and short-run cost function choices, they optimally choose capacity levels and risk levels with limited information, and they interact in markets, leading to an asymmetric Cournot-Nash equilibrium. Each bank's profit function is derived from the market equilibrium conditions, and its parameters are directly estimated (in structural, not reduced, form).
- The empirical analysis incorporates six bank products, and develops a unified approach to thinking about what a bank's products are. Specifically, this analysis includes off-balance-sheet items as an individual product line; to date, only Hasan, Karels, and Peterson (1994) and Jagtiani, Nathan, and Sick (1994) have examined this product line.
- The dataset on customer satisfaction is new to the literature, developed explicitly for this paper.
- Most of the previous empirical work examined one or a few very specific factors, attempting to measure their impact in isolation from other decisions and activities undertaken by the firm. In this paper, all of the relevant factors are integrated into a single model, so that all parameters are estimated simultaneously.

Developing an explicit integrated structural model of banking markets increases the complexity of the analysis and increases the difficulty of estimation. Further, this is not the standard approach in the existing empirical banking literature. The motivation to undertake this effort is two-fold. First, deriving the estimating equations from a model of maximizing agents, their opportunity and information sets, and their market interactions tightly ties the empirical estimation to the underlying economics. Second, with

this approach, specific hypotheses of bank behavior that can be reflected in the economic model can be explicitly estimated. For example, mismatches in bank demand and bank capacity, due (possibly) to management forecasting errors, are built into the economic model, so that allocative inefficiencies due to this potential mismatch can be explicitly estimated for each bank.

This second point is particularly important, given the current state of the empirical banking literature. Using the methods of frontier analysis, Berger and Humphrey (1991) have shown that there are large unexplained differences among the cost functions of banks. Berger, Hancock, and Humphrey (1993), derive similar but more powerful results using a bank profit function. Unfortunately, frontier methods (or fixed effects methods, as used in this paper) simply allow us to characterize bank-specific effects that are not captured in our models, but do not explain why it is that a bank is inefficient compared to other banks. We have thus reached a limit on what econometric analysis can tell us using generic functional forms and generic operating data (such as the *Call Report* data traditionally used for this purpose), if we wish to understand the underlying economics of banking markets. One ingenious approach to get beyond this limit is in Berger, Leusner, and Mingo (1994), in which a generic cost function approach is combined with a unique dataset on bank branches is used to estimate a specific type of potential inefficiency. This paper represents another approach: use generic banking data but with a richer structural model of banking markets, combined with a unique dataset (customer satisfaction data) which prises open yet another dimension for analysis. As researchers are better able to model and then observe potential causes of inefficiency, the fixed effects, which capture the unobserved variation among banks, become less important. The results of this paper take us a very long way toward that goal.

There are five distinct parts of our model of banks and banking markets. (i) **operating activities** each bank is assumed to produce (at most) six products: demand deposits, time deposits, commercial and industrial (C&I) loans, consumer loans, real estate loans, and off-balance-sheet items. The short-run cost of producing these six products is represented by a family of functions which capture potential complementarities among the products. The long-run cost function is the lower envelope of this family of cost functions. In each period, each bank chooses a short-run cost function based upon its estimates of demand for its products. These estimates may prove inaccurate, but its choice of technology is sunk (for one period) so that the bank incurs a cost penalty, operating off the long-run cost function. (ii) **risk**; a defining property of banks is that risk and risk management is at the core of their business. Each product, indeed each transaction, involves taking risk. We measure the overall risk of the bank using the “beta” of the bank’s stock, which captures not only credit and interest rate risk (such as would be associated with loans and securities) but all business risk (for example, expanding into new geographical markets). (iii) the **demand** for banking products; the market demand functions capture potential cross-elasticities among these six products. (iv) **customer satisfaction or quality**; banks may differ in their ability to provide

high-quality service, as defined by how satisfied their customers are³. For reasons explained below, we treat the ability to deliver high quality (or not) as *a characteristic* of a bank rather than as a *choice* the bank makes. (v) **competitive interactions** each bank having chosen a short-run cost function, all banks within each metropolitan statistical area (MSA) then play a Cournot-Nash quantity game (involving six markets); the output vector for each bank and price is determined as the equilibrium of this game. It is an asymmetric game, as banks in general will have chosen different short-run cost functions and different product risk levels, and therefore have different marginal costs. This will result in different output vectors for different banks in the same MSA.

The paper is organized as follows. Section 2 lays out the five distinct components of the model. Also included in this section is our treatment of what a bank's products are. This issue has generated some controversy in the literature, and our approach differs from any in the literature thus far; it is closest in spirit to the user cost method, outlined in Berger and Humphrey (1992). Section 3 uses these results to derive the estimating equations. In Section 4, the data is described, including how various anomalies in the several datasets were treated. Section 5 contains the results of the estimation, and Section 6 lists our conclusions as well as possible future research along the lines developed in this paper.

2. The Model

The model consists of four distinct components, which are described in turn.

Cost of Operating activities A bank can offer up to six products:

1. Demand deposits;
2. Time deposits;
3. Commercial and Industrial (C&I) loans;
4. Consumer loans; this category includes credit card accounts as well as the usual consumer credit loans such as automobile loans, etc.
5. Real estate loans; this includes home mortgages (though not mortgaged-backed securities) and commercial real estate loans;
6. Off-balance-sheet items; we include in this category all counterparty guarantees, such as letters of credit; we do not include futures contracts, such as swaps⁴.

Each bank's production of these six outputs in a time period is measured by the outstanding stock of each output on the bank's balance sheet at the end of the period (for products 1-5, and analogously for product 6). In the short run, production involves both fixed and variable costs; for each period, a bank must plan for and invest in capacity before demand is realized, based upon, among other things, its demand forecasts. Each bank faces a family of short-run cost functions of the form⁵

$$C(\mathbf{q}; F, \mathbf{c}) = F + \sum_{i=1}^6 c_i q_i^{2a_{ii}} + \sum_{i=1}^5 \sum_{j=i+1}^6 \mathbf{n}_{ij} (q_i q_j)^{a_{ij}}, \quad (1)$$

where \mathbf{v} and $\boldsymbol{\alpha}$ are symmetric matrices common to the entire family of short-run cost functions. The bank chooses a specific member of this family by selecting \mathbf{a} and $\mathbf{c} = (c_1, \dots, c_6)$, which choice variables completely characterize the menu of short-run cost functions. In turn, the lower envelope of this family of short-run cost functions defines the long-run cost function for every output vector in the positive orthant, $\mathbf{q} \in \mathbb{R}^6$, there exists a short-run cost function that (i) is equal to the long-run cost function $\mathbf{a}\mathbf{q}$; (ii) is tangent to the long-run cost function $\mathbf{a}\mathbf{q}$; and (iii) lies everywhere above the long-run cost function.

The long-run cost function is of the form

$$C(\mathbf{q}) = \sum_{i=1}^6 \sum_{j=i}^6 \mathbf{l}_{ij} (q_i q_j)^{g_{ij}} \quad (2)$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ are symmetric 6×6 matrices. For each vector $\tilde{\mathbf{q}}$ in the positive orthant, there exists a unique short-run cost function (F, \mathbf{c}) defined by the total condition (i):

$$C(\tilde{\mathbf{q}}; F, \mathbf{c}) = C(\tilde{\mathbf{q}}), \text{ or}$$

$$F(\tilde{\mathbf{q}}; \mathbf{c}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{v}) = \sum_{i=1}^6 \sum_{j=i}^6 \mathbf{l}_{ij} (\tilde{q}_i \tilde{q}_j)^{g_{ij}} - \sum_{i=1}^6 c_i \tilde{q}_i^{2a_{ii}} - \sum_{i=1}^5 \sum_{j=i+1}^6 \mathbf{n}_{ij} (\tilde{q}_i \tilde{q}_j)^{a_{ij}} \quad (3)$$

and the tangency conditions (ii):

$$\nabla C(\tilde{\mathbf{q}}; F, \mathbf{c}) = \nabla C(\tilde{\mathbf{q}}), \text{ or}$$

$$c_i = \frac{\tilde{q}_i^{-2a_{ii}}}{2a_{ii}} (2\mathbf{b}_{ii} \mathbf{l}_{ii} \tilde{q}_i^{2g_{ii}} + \sum_{j \neq i} [\mathbf{b}_{ij} \mathbf{l}_{ij} (\tilde{q}_i \tilde{q}_j)^{g_{ij}} - \mathbf{a}_{ij} \mathbf{n}_{ij} (\tilde{q}_i \tilde{q}_j)^{a_{ij}}]), \quad i = 1, \dots, n. \quad (4)$$

We assume the second-order conditions are satisfied.

The intuition here is straightforward. Banks choose a short-run cost function from the available menu; if they expect relatively low demand, they choose a low fixed cost, high variable cost technology. If they expect relatively high demand, they choose a high fixed cost, low variable cost technology. The available menu of short-run cost functions is defined by its lower envelope, which is therefore the long-run cost function.

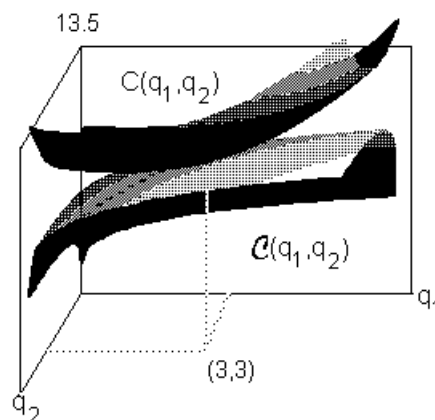
To help visualize these short-run and long-run cost functions, an example for two products is worked out in detail and shown graphically in Figure 1. The parameters of this example are:

$$\mathbf{a} = \begin{pmatrix} 1.2 & .5 \\ .5 & 1.2 \end{pmatrix}, \quad \mathbf{g} = \begin{pmatrix} .2 & .15 \\ .15 & .2 \end{pmatrix}, \quad \mathbf{I} = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}, \quad \mathbf{n}_{1,2} = -.06$$

The long-run cost function is completely determined by the parameters \mathbf{a} and γ . All members of the family of short-run cost functions share parameters \mathbf{a} and \mathbf{v} . We select a particular member of the family of short-run cost functions by specifying an output vector $\tilde{\mathbf{q}} = (3,3)$ at which the short-run function is tangent and equal to the long-run cost function. We then solve for F, \mathbf{c} by solving equations (3) and (4), which results in:

$$c_1 = c_2 = 0.071, F = 10.297$$

With these parameter values, the long-run cost function exhibits economies of scale and diseconomies of scope. The short-run cost function exhibits diseconomies of scale for reasonably large values \mathbf{q} and economies of scope. This is illustrated in the graph of the two functions, below. Note that the cost functions are equal and tangent at (3,3).



Input Prices It will be noted that both our long-run and short-run cost functions include the output vector but do not include input prices, as theory demands that they should. Only if input prices are assumed constant across banks and time periods is their omission justifiable. Since all variables are in constant (1982) dollars, inflation is not a source of variation in input prices. However, other time-related factors could cause variation in input prices, and more obviously, different sections of the country are likely to have different wage rates, leading to different input prices. In the estimation described below, however, both time dummies and (high-wage) geographic area dummies are not significantly different than zero, suggesting that such sources of variation are not present (or a perfectly canceled out by some other unobserved effect). Therefore, we maintain the assumption that variations in input prices are small enough so that we may safely assume input prices are relatively constant and can therefore be ignored.

Cost of Risk Banks are unique in that the management of risk is the core of their business. Yet our ability to measure risk is much less than our ability to measure, say, the cost of processing checks. In this paper, we use *market* measures of risk, based on the theory of financial markets, in order to measure the *aggregate* risk of the bank. Our risk measure is the bank's (or its holding company's) stock market beta, the measure of risk that stockholders bear (and must be compensated for) derived from the Capital Asset Pricing Model. Beta measures the covariance of a stock's price performance with that of the aggregate market, fully reflecting the effects of stockholders' portfolio diversification.

Costs are normally denominated in dollars per unit time, while the stock market beta is a dimensionless quantity. Since we wish to treat risk cost like any other cost of the bank, we develop a transform of the stock market beta which is a cost of risk commensurate with all other costs. We quantify the bank's total risk cost as the difference between current earnings and what the earnings of the bank must be to achieve the same value of the firm if the bank carried no risk. Let the value of the bank be V , the earnings of the bank in future period t be the random variable Π_t . Then the market discount rate of the bank is s , where s satisfies

$$V = E \left[\sum_{t=1}^{\infty} \frac{\Pi_t}{(1+s)^t} \right],$$

and E is the expectation operator. The discount rates incorporates the stock market's assessment of risk β .⁷ Define the *uniform expected earnings* as \bar{p} , the uniform certain level of earnings which would yield

the same value of the firm if discounted at s :

$$V = \sum_{t=1}^{\infty} \frac{\bar{p}}{(1+s)^t}.$$

Now define the *risk-free earnings* as \tilde{p} , the uniform certain level of earnings which would yield the same value of the firm if it were completely free of risk, and the earnings discounted at r_f :

$$V = \sum_{t=1}^{\infty} \frac{\tilde{p}}{(1+r_f)^t}, \text{ or}$$

$$\tilde{p} = \bar{p} \cdot \left(\frac{r_f}{r_f + b(r_m - r_f)} \right)$$

Now the risk cost is defined to be the additional earnings $\bar{p} - \tilde{p}$, that investors demand because of the risk that the bank carries. This is simply

$$R = \bar{p} \cdot \left(\frac{b(r_m - r_f)}{r_f + b(r_m - r_f)} \right); \text{ normalizing by expected earnings,}$$

$$\frac{R}{\bar{p}} = \left(\frac{b(r_m - r_f)}{r_f + b(r_m - r_f)} \right) = \text{risk / earnings ratio.} \quad (5)$$

Aggregate risk for a bank is comprised of risks from many individual transactions, each of which is associated with a specific product. We assume the most straightforward reduced form model that decomposes total bank risk into risk by product:

$$R^k = \sum_{i=1}^6 \bar{r}_i^k q_i^k. \quad (6)$$

A more general functional form involving higher order terms would no doubt conform more closely to our view of risk. However, we are estimating *bank-specific* coefficients in this study, so the per-bank sample size is quite limited, thereby restricting the parameters that can usefully be estimated. As a result, we must be content with capturing first-order effects only.

Demand Total market demand for these products is interdependent and is given by the inverse demand system⁹:

$$p_i = A_i \prod_{j=1}^6 Q_j^{h_{ij}}, \quad i = 1, \dots, 6. \quad (7)$$

where the matrix \mathbf{h} is symmetric. This demand system has the property that the self- and cross-flexibilities $\frac{Q_j}{P_i} \frac{\partial P_i}{\partial Q_j} = \mathbf{h}_{ji}$ are constant. The more familiar self- and cross-elasticities can be obtained from the flexibilities by inverting the flexibility matrix, and they are (of course) constant as well: $\mathbf{h}^{-1} = \mathbf{e}$, the elasticity matrix.

Each firm $k = 1, \dots, m$ supplies a portion $\mathbf{q}^k = (q_1^k, \dots, q_6^k)$ of this total demand, with $\mathbf{Q} = \sum_{k=1}^m \mathbf{q}^k$.

Customer Satisfaction/Quality Banks may differ in their ability and willingness to provide service quality that leads to higher satisfaction of their customers. Recent reports in the trade press suggest that customer satisfaction may be a highly profitable strategy for banks. “Financial marketers appear to have overlooked the fundamental truth that the longer an institution keeps a customer, the more profitable the customer becomes. [Banks need]...to maximize their satisfaction with [the] institution,” according to Vavra (1995). The inclusion of customer satisfaction in this analysis is unique among econometric studies, and is merited not by its intrinsic interest but also by the current interest among practitioners in quality.

The quality measure we use is an index derived from an extensive and long-standing survey of commercial bank customers conducted by Greenwich Associates, Inc., a market research firm specializing in the commercial banking sector. The index merges survey responses into a single number, normalized between 1 and 100, with a higher index number corresponding to greater customer satisfaction.

The details of the survey and the construction of the index from the survey instrument are discussed below in Section 4. We note here the key features of the index that validate its use in this study as a measure of customer satisfaction: (i) actual bank customers are surveyed to assess their opinions regarding their degree of satisfaction with their banking relationships. (ii) The survey has been conducted annually for over twenty years; the index is therefore based on a set of questions and responses that have been consistent over the period of our study. (iii) The survey results are considered informative by the banks themselves, as attested to by the fact that Greenwich Associates’ principal source of income over this period has been the sale of the survey results to individual banks. The validity of the results has thus passed a “market test.”

The quality measure may be modeled as either a choice variable of the bank, or a characteristic of the bank over which it has little or no control in the short term. In this paper, we treat quality as a characteristic of the firm rather than a choice variable. If the technology for providing quality in banking were well-understood and completely diffused throughout the industry, then modeling it as a choice

variable would be appropriate. In our view, this is not the case. While banks were subject to the pervasive regulation that characterized the industry pre-1980, it appears that quality was defined more by the brand of toasters the bank gave depositors for opening accounts than by how well banks determined and then provided what customers actually wanted. However, as competition intensified over the last fifteen years, many banks have sought to differentiate themselves as high-quality providers as a means of competitive advantage. Pursuing this strategy requires a bank to actually learn how to provide high-quality service; this strategy may require major changes within the bank, which may not easily be imitated. Thus, the diffusion of the technology of providing high-quality service could be rather slow, so that over the time period of our study, banks had either acquired the technology or they hadn't. If this is an accurate description of the diffusion of the technology of high-quality service, then treating quality as a characteristic rather than a choice variable is more appropriate, and that is our rationale for so treating it.

There are several empirical implications of this assumption. First, if quality is a choice variable, then in equilibrium we would expect to see returns equalized across quality choice; high-quality banks would be no more nor no less profitable than low-quality banks. However, if quality is a characteristic of banks, determined by the speed of diffusion of the technology, there may indeed be rents to providing high-quality service. Second, if quality is a choice variable, then in equilibrium we would expect that higher quality necessarily results in higher cost. If quality depends upon the relatively slow diffusion of a new technology, then adopters of the new technology need not have higher costs; in fact, they could have lower costs, if the technology both increases quality and reduces costs¹⁰.

Unfortunately, the quality index constructed from the Greenwich Associates' survey applies only to commercial customers. This suggests that the quality we measure affects only products associated with this market segment. The product most clearly associated with this segment is C&I loans¹¹. Higher quality can affect both demand and costs; if we denote the quality index by q , then we modify the short-run cost function (1) and the long-run cost function (2) as follows (assuming C&I loans are product 3):

$$C(\mathbf{q}; F, \mathbf{c}) = F + \sum_{i \neq 3} c_i q_i^{2a_{ii}} + \sum_{i \neq 3} \sum_{j \neq i} n_{ij} (q_i q_j)^{a_{ij}} + X^d \left(c_3 q_3^{2a_{33}} + \sum_{j \neq 3} n_{3j} (q_3 q_j)^{a_{3j}} \right)$$

$$C(\mathbf{q}) = \sum_{i \neq 3} \sum_{\substack{j \geq i \\ j \neq 3}} I_{ij} (q_i q_j)^{g_{ij}} + X^d \sum_{j=1}^6 I_{3j} (q_3 q_j)^{g_{3j}}$$

The sign of the exponent d determines whether increasing quality increases or decreases cost. Similarly, the demand system (7) can be modified to reflect quality as well. Since only C&I loans (product 3) are affected, this reduces to:

$$p_3 = X^V P_3 \prod_{j=1}^6 Q_j^{h_{3j}}$$

Assuming $V > 0$, increasing quality increases customers' willingness to pay. To reduce notation in the following exposition, we will suppress the explicit appearance of αX in the cost functions demand functions.

In equilibrium, the market may respond to quality in two ways: banks may charge higher prices for higher quality service, more business may come to higher quality banks, or both. Thus, the effect of higher quality service on bank profitability is potentially threefold: the ability to charge a price premium, a higher level of demand, and a change in costs. If quality is correctly viewed as a bank characteristic, then we have no prediction of its effect on profitability. If quality is a very slowly diffusing innovation which banks are unable to adopt easily, then we would not expect that high quality would lead to lower profits, but it could certainly lead to higher profits, especially during the period of slow diffusion where quality rents may be possible.

Products and Prices There is some controversy within the literature as to what a bank's outputs are. The asset approach takes bank deposits as inputs and bank loans as outputs. The user cost approach identifies as outputs those items for which revenues exceed costs, and inputs as those assets for which costs exceed revenues. In this analysis, we adopt the user cost method, but with the special focus that it is the net economic revenue, rather than the accounting revenue, that is the appropriate measure. We take the measured quantities of each of the six products to be its stock in dollars. This then focuses attention on what the price of the six products specified above really is. The price of a loan that the bank makes to a customer is obvious: it is the interest rate, denominated in dollars per dollar-year. The appropriate unit of loan quantity is again obvious: the total amount of loans outstanding in a given year.

More difficult is the pricing of demand deposits, and it is to this that we now turn. There are several types of pecuniary transactions associated with demand deposits: (i) interest payments to depositors, usually paid if a substantial minimum balance is held; (ii) various fees charged depositors, such as ATM usage, returned checks, per-check fees, etc. On balance, the total net revenue to a bank from these fees is very small and can be negative. What is missing from the pecuniary transactions is the opportunity cost of demand deposits to depositors. By keeping funds in a demand account, depositors are giving up earning returns (in an equivalently low-risk instrument of equivalent liquidity) in exchange for the transaction and depository services provided by the bank. In turn, the bank gets the use of this money for loan purposes while paying minimal interest costs. Assume that the next best alternative for the bank's customer is 90-day Treasury bills (highly liquid, very safe), then the opportunity cost to the depositor is the interest rate on the 90-day T-bills. The total price that the customer pays, including opportunity cost, is the rate for

90-day T-bills plus bank account fees minus bank interest payments. A similar calculation applies to time deposits, in which the customer is giving up potentially higher interest payments on T-bills for the convenience of readily available insured deposits. The implicit price is thus the 90-day T-bill rate minus the bank interest rate on time deposits; this calculation generally yields a positive price paid by time depositors for this service, though generally not as high a price as for demand deposits, for which the bank typically provides more service.

If these opportunity costs are counted as a price the customer pays the bank, then “revenues” are attributed to the bank that do not actually show up on its books. Again, we use the opportunity cost approach to handle this. The bank uses these “free” (in the pecuniary sense) funds in order to make loans, funds it would have to pay for if it did not have a deposit base to draw upon. Therefore, the opportunity cost “revenues” imputed from depositors should be imputed to borrowers. Thus, the price of a loan to a borrower is not the quoted interest rate, but that interest rate less the opportunity cost of funds collected from depositors. Bankers are familiar with this concept in the form of “net interest income;” interest from borrowers less the cost of funds. This approach differs in detail but not in spirit, and it permits the total revenues of the bank to equal the booked revenues. This pricing approach can be viewed as allocating booked revenues to products based on an opportunity cost view of banking.

3. Market Equilibrium and the Estimating Equations

In this section, the various parts of the model are brought together. The model of the banking market is developed and the equilibrium of the multiproduct game is derived. Using these equilibrium conditions, equations which depend only upon observables and parameters are derived; these are the equations to be used to estimate the model parameters.

Information and Timing of Play Prior to the beginning of each period, the managers of each bank must decide on the short-run cost function to deploy for the coming period, which requires that they commit to a fixed cost F^k prior to observing actual demand conditions. This amount is fixed in the short run and cannot be changed until the beginning of the next period.² Bank managers are assumed to know α , γ , λ , ν , and equations (3) and (4), but they do not know next period’s demand. In order to optimally choose the short-run cost function, banks must forecast demand, choosing the cost function which minimizes their expected cost over the forecast distribution.

A model of how banks can optimally do this is presented in Appendix A. Two key features of this model are: (i) since banks use private information in their forecasts, their optimal choice of short-run cost functions will differ; (ii) bank k ’s choice of cost function can be expressed as a choice of “planned-for”

demand $\tilde{\mathbf{q}}^k = (\tilde{q}_1^k, \dots, \tilde{q}_n^k)$, which is not necessarily equal to expected demand. If banks are making unbiased forecasts, then if costs are nonlinear it will be efficient to choose a planned-for demand not equal to its expected demand. Thus, any mismatches between capacity and demand may be an efficient response to forecast uncertainty.

On the other hand, bank managers may exhibit a bias in their capacity planning. Banks which are consistently over-optimistic will tend to have excess capacity in most periods, while under-optimistic banks will tend to be short on capacity. If such mismatches of demand and capacity are due to bias, then these mismatches represent allocative inefficiency. An objective of this analysis is to determine the extent of allocative inefficiency due to this source.

We denote the total bias of the bank by m^k ; our assumption is thus:

$$m^k = \frac{\tilde{q}_i^k}{q_i^k}, \text{ for } i=1, \dots, 6, \text{ and all time periods.} \quad (8)$$

The bias m^k is a bank-specific parameter to be estimated.

Market Equilibrium We assume that the number of banks in the market is fixed. In the case at hand, we identify metropolitan statistical areas (MSA) as the relevant banking markets. At the beginning of the period, demand is revealed. All banks in the market then play a Cournot-Nash quantity game, now with full knowledge of demand \mathbf{Q} , in which they offer \mathbf{q}^k (not necessarily equal to $\tilde{\mathbf{q}}^k$) in the market.

The timing of the game is illustrated in Figure 1.

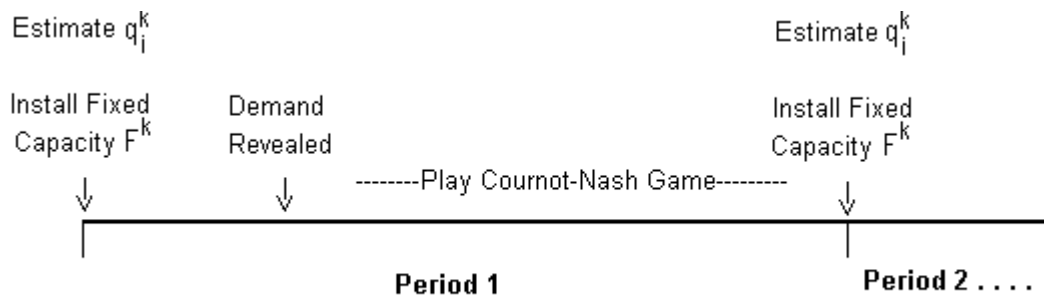


Figure 1

The form of the game suggests two questions: (i) is the Cournot-Nash assumption reasonable for these markets? and (ii) are banks the only players in these six product markets? In fact, the Cournot quantity game with a fixed number of players is uniquely suited to banking markets. Entry into these markets is restricted (though not prohibited) by government regulation, so that the “no entry “ assumption is a reasonable though not perfect approximation to reality, especially for larger banks. Prices tend to be market-determined, and are often strongly affected by capital market activity over which banks have little effect. By choosing capacity, such as size of branch network, extent of trading activities, number of credit officers hired and trained, banks are choosing quantities to offer the market, coincident with the Cournot quantity game.

More troubling is our assumption that the banks in our sample are the only players in the game. In fact, there are many more players in these markets, including smaller banks, non-bank financial institutions, and commercial paper markets. In virtually all of the products in this model, banks in our sample compete with institutions not in our sample. Lack of available data on these non-bank institutions constrains all researchers from addressing this issue. The results of this paper, as well as the results of virtually the entire empirical banking literature, must be understood in light of this deficiency.

Another concern is our assumption that the relevant market is the MSA. For retail products and many middle market products this is a good assumption. However, for some services such as corporate loans the market may well be national. This suggests that the market share of banks is less than that used here, so that price-cost margins would also be less.

Bank k chooses the quantity vector \mathbf{q}^k to maximize *economic* profit. Using equations (1) and (6), this can be written as:

$$\max_{\mathbf{q}^k} \mathbf{p}^k = \sum_{i=1}^n p_i(\mathbf{Q})q_i^k - F^k - \sum_{i=1}^n c_i^k (q_i^k)^{2a_{ii}} - \sum_{i=1}^{n-1} \sum_{j=i+1}^n n_{ij} (q_i^k q_j^k)^{a_{ij}} - \sum_{i=1}^n \bar{r}_i^k q_i^k \quad (9)$$

The first order conditions are:

$$\frac{\mathcal{J}\mathbf{p}^k}{\mathcal{J}\mathbf{q}_i^k} = MR_i^k - MC_i^k = p_i \left(1 + \sum_{j=1}^n s_{ji} \mathbf{f}_{ji} \mathbf{q}_j^k\right) - MC_i^k = 0,$$

where

$$s_{ji} = \frac{p_j Q_j}{p_i Q_i}, \quad \mathbf{f}_{ji} = \frac{Q_i}{p_j} \frac{\mathcal{J}p_j}{\mathcal{J}Q_i} = \mathbf{h}_{ji}, \quad \mathbf{q}_j^k = \frac{q_j^k}{Q_j},$$

which are, respectively, the share of product j total revenues relative to product i total revenues, the flexibility of price j with respect to quantity i (constant for the assumed functional form of the demand system), and the market share of firm k for the j^{th} product.

This first-order condition can be put into this more familiar form:

$$\frac{p_i - MC_i^k}{p_i} = \sum_{j=1}^n s_{ji} h_{ji} q_j^k \quad (10)$$

Now the marginal cost of product i for firm k consists of two parts: the operations cost and the risk cost:

$$MC_i^k = 2a_{ii} c_i (q_i^k)^{2a_{ii}-1} + \sum_{j \neq i}^n a_{ij} n_{ij} (q_i^k)^{a_{ij}-1} (q_j^k)^{a_{ij}} + \bar{r}_i^k.$$

solving out the full FOC for the coefficient c_i^k we obtain:

$$c_i^k = \frac{p_i (1 - \sum_{j=1}^n s_{ji} h_{ji} q_j^k) - \sum_{j \neq i}^n a_{ij} n_{ij} (q_i^k)^{a_{ij}-1} (q_j^k)^{a_{ij}} - \bar{r}_i^k}{2a_{ii} (q_i^k)^{2a_{ii}-1}} \quad (11)$$

This equation specifies what c must have been for firm k , based on the observations \mathbf{p} and \mathbf{q}^k , as well as the parameters α , \mathbf{v} , and η . We denote this relationship $\mathbf{c}^k(\mathbf{p}, \mathbf{q}^k; \alpha, \mathbf{v}, \eta)$.

Equation (3) shows that F^k also depends upon $\tilde{\mathbf{q}}^k$, so equation (8) can be used to express F^k as a function of the bank-specific parameters m^k and observables \mathbf{q}^k . Bank k 's profit can be expressed as a function of parameters and observables:

$$\mathbf{p}^k = \mathbf{p} \cdot \mathbf{q}^k - \sum_{i=1}^n c_i^k (q_i^k)^{2a_{ii}} - \sum_{i=1}^n \sum_{j=i}^n n_{ij} (q_i^k q_j^k)^{a_{ij}} - F(m^k \mathbf{q}^k; \mathbf{c}^k, \alpha, \gamma, \lambda, \mathbf{v}) - \sum_{i=1}^k \bar{r}_i^k q_i^k. \quad (12)$$

Estimating Equations Economic profit cannot be observed from the books of the bank, as it includes risk costs, among other possible costs not accounted for. We assume in this work that risk costs are the only costs that do not appear on banks' books, so that economic profit is equal to accounting earnings (which is observable) less risk cost: $\pi^k = \hat{\mathbf{p}}^k - R^k$. R^k in turn depends upon market observables r_f , r_m , and \mathbf{b}^k , as well as the growth rate g^k . In addition, we also include a dummy variable B^k for each bank. These are the fixed effect coefficients, designed to assess any bank-specific effects not picked up by the parameters of the model.

There are thus two estimating equations, one for risk and the other for operational activities. We use actual bank earnings $\hat{\mathbf{p}}^k$ as a proxy for uniform expected earnings:

$$R^k = \hat{\mathbf{p}}^k \left(\frac{\mathbf{b}^k (r_m - r_f)}{r_f + \mathbf{b}^k (r_m - r_f)} \right) = \sum_{i=1}^6 \bar{r}_i^k q_i^k \quad (13)$$

$$\hat{\mathbf{p}}^k = \mathbf{p} \cdot \mathbf{q}^k - \sum_{i=1}^n c_i^k (q_i^k)^{2a_{ij}} - \sum_{i=1}^6 \sum_{j=i}^6 n_{ij} (q_i^k q_j^k)^{a_{ij}} - F(m^k \mathbf{q}^k; \mathbf{c}^k, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{v}) + B^k \quad (14)$$

Everything in these two equations is either observable ($\hat{\mathbf{p}}^k, r_f, r_m, \mathbf{b}^k, \mathbf{p}, \mathbf{q}^k$), is a parameter ($\boldsymbol{\alpha}, \boldsymbol{\gamma}, \mathbf{v}, \boldsymbol{\lambda}, \mathbf{m}, \mathbf{B}$), or is derived from observables and parameters (\mathbf{c}^k, F^k). Therefore, we can estimate it using nonlinear methods. Note that these two equations are not simultaneous.

Fixed Effects and Frontier Methods Much of the recent work in the empirical banking literature has used frontier methods to examine bank efficiency differences that are not explained by the parameters of the model. The principal feature of all frontier methods in bank panel data is the derivation of the “efficient frontier”, either cost or profit, from consistent differences in bank-specific residuals. The idea is that banks with consistently high (for profit) residuals identify the “best practice frontier,” and that banks with lower residuals are off this frontier and are thus X-inefficient, in the sense of Leibenstein (1966).

These methods are best viewed as consisting of two parts: (i) the identification of consistent differences among banks based on their average residual or similar measure; (ii) the inference from these differences that banks with the most favorable average residuals represent a “best practice frontier.” There is no question that the recent work has been extremely valuable in that it has uncovered very large systematic differences among the costs and profitability of banks; these differences are so large as to dwarf potential scale and scope economies (see Berger and Humphrey (1991)). However, this value inheres in the first part of the analysis. The second part, in which inferences are made about the best practice frontier on the basis of these differences, appears to be more speculative. The assumption of frontier analysis is that the observed differences among firms are due to “superior management of resources” (Berger (1993)). However, differences in banks’ abilities to manage resources should in principle be observable, if the relevant variables are included in the model. Significant observed fixed effects are best viewed as a measure of our ignorance, as an indication that researchers need to look harder for possible omitted variables. Therefore, our approach in this paper is to focus exclusively on the systematic differences among banks, and not on inferences regarding “best practice” frontiers.

To do so, we use a fixed effects model, in which a dummy variable is included for each bank. This is the function of the variable B^k in equation (14). In this analysis, the role of B^k is to measure the extent to which bank-specific factors which we have not included in the model are important in determining bank profitability.

4. The Data

Three different datasets were employed in this study: *operating data*, *capital market data*, and *customer satisfaction data*.

Operating Data This data is taken from the Report of Condition and Income (“Call Report”), which all insured banks operating in the US are required to file with Federal regulators. We use quarterly data, from 1984 to 1992 inclusive, for all banks with over \$1 billion in assets in 1984. The variables in this dataset include:

quantities, end-of-period balance sheet entry for each bank for each of the six products listed;

revenues, total period net revenue for each bank for each of the six products listed;

earnings, income before extraordinary items and after taxes for each bank.

All quantities are expressed in thousands of 1982 dollars.

This data is collected from the banks themselves by the FDIC, which maintains the dataset. It is *accounting data*.

Capital Market Data This data is taken from the CRSP (Center for Research in Stock Prices) dataset, which contains end-of-day prices of stocks traded on major exchanges and other securities, such as bonds. The variables in this dataset include:

market rate of return, this is the total return (dividends plus capital gain) for the NYSE.

risk-free rate of return, this is the interest rate on 90-day Treasury bills.

beta; the \mathbf{b} of each bank for each quarter was computed using the actual price data together with the market rate and the risk-free rate.

This data is collected from US stock markets by the Center for Research in Stock Prices at the University of Chicago. It is *market* data.

The mapping from banking institutions listed on stock exchanges to banks listed in the Call Report is not trivial. The reporting unit for the Call Report is a state bank; often, this bank is part of a holding company that operates banks in several states. In turn, this company may be owned by another holding company, whose assets in principle could include non-banking firms. It is generally the “highest” holding company that is listed on a stock exchange. Even the mapping from “high holding companies” reported to the FDIC and stock exchange listing is not trivial; after extensive detective work, 219 of the banks with greater than \$1 billion in 1984 assets were identified for which their high holding company was listed on a stock exchange. The holding companies often held more than one bank in the sample, and may well have held other firms, all of which contributed to their risk. In this analysis, the high holding company’s beta was imputed to all its subsidiaries.

Not all banks had data which covered the entire period. Indeed, many banks were merged into new banks during this period. Since there is nothing in the model that is inherently dynamic, these mergers did not constitute a problem; as of the date of the merger, the old banks dropped from the dataset and the new one was inserted. The total number of data points (banks \times quarters) is 6190.

Customer Satisfaction Data This data is taken from an ongoing survey conducted by Greenwich Associates, Inc., a marketing research firm that has been conducting surveys of commercial customers of US banks since 1972. The survey is designed to elicit from customers (i) their degree of satisfaction with the banks they do business with; and (ii) the specific factors and attributes important to them, and how their banks fared on these items. From this extensive survey data, Greenwich Associates constructed a single index, expressly for this study, to measure overall customer satisfaction for each bank. The details of the survey methods and the construction of the index is contained in Appendix B. The key points about this index: (i) it is scaled to range from 0 to 100, with a mean of 50; higher scores correspond to greater customer satisfaction; (ii) only commercial customers are surveyed; thus, the quality index only applies to commercial products, in particular C&I loans; (iii) due to data limitations, only 112 banks (from our larger sample) are included, during the period 1985 to 1992. The surveys are conducted annually (at most), so quarterly data is not available. The total number of data points (bank \times years) is 476.

5. Empirical Results

The empirical results are presented in three parts. First, the results of the *risk estimation* are presented, as equation (13) can be estimated by itself. Next, the results of the *operating results estimation* are

presented, and last the results of the *quality estimation*. While these last two are nominally simultaneous, certain regularities in the data permit their separation.

Risk Estimation The risk cost/earnings ratio (on the left-hand side of (13)) was computed¹³ for each bank in each quarter for which data was available. The empirical cumulative distribution of the mean (across all quarters) of this ratio is plotted below:

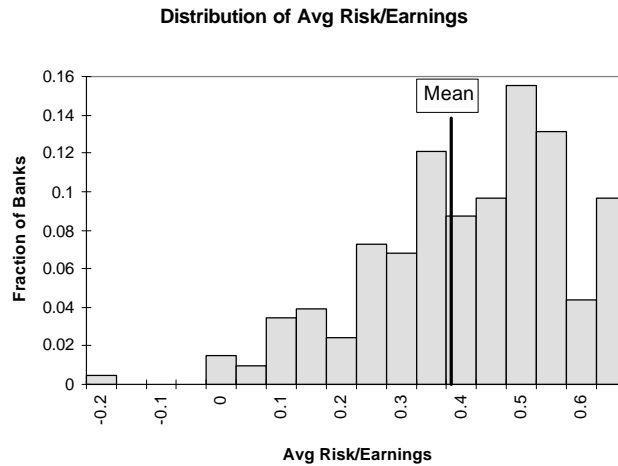


Figure 2

On average, the cost of risk accounts for 38% of earnings. If we use actual earnings as a proxy for uniform expected earnings, then the cost of risk is about 3% of booked cost. More interesting is the rather substantial spread of risk/earnings; for some banks risk accounts for 2/3 of their booked earnings, indicating that accounting earnings overstates their economic profit by a factor of three. Other banks apparently are able to manage risk more successfully, achieving negative b 's and leading to an economic profit greater than accounting earnings. This suggests that the ability to manage risk well is a scarce resource in banking.

The period of study was a turbulent one for banking markets, with very good times alternating with very bad times. These time patterns are evident in the plot of the mean and standard deviation of the risk cost-earnings ratio by year:

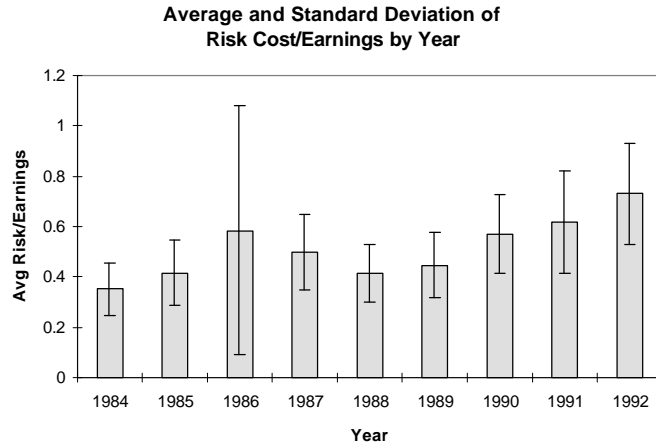


Figure 3

The effect of the S&L crisis is clearly evident in the increase in mean risk-earnings ratio for 1986, and particularly the large one-time increase in the standard deviation of the distribution across banks. The progressively greater riskiness of all banks into the early 1990s is evident in the mean risk; the fact that this increasing riskiness applied to all banks is suggested by the relatively stable standard deviation over this period.

Perhaps even more interesting is the relationship between size and risk. For banks under \$1 billion in assets, McAllister and McManus (1993) found that increasing size permitted lower risk costs for banks resulting from inventory economies. Our findings, based on banks over \$1 billion in assets and a stock market-based risk measure, are overall that the risk-earnings ratio is increasing in bank size. A linear regression of annual risk-earnings ratio on bank revenues yields a positive coefficient that is significant at the 99% level. . However, the relationship between size and risk is a complex one; in order to understand the fine structure of the data, the nonparametric curve-fitting Trewess method (Velleman, 1980) was used. The data and results of these analyses:

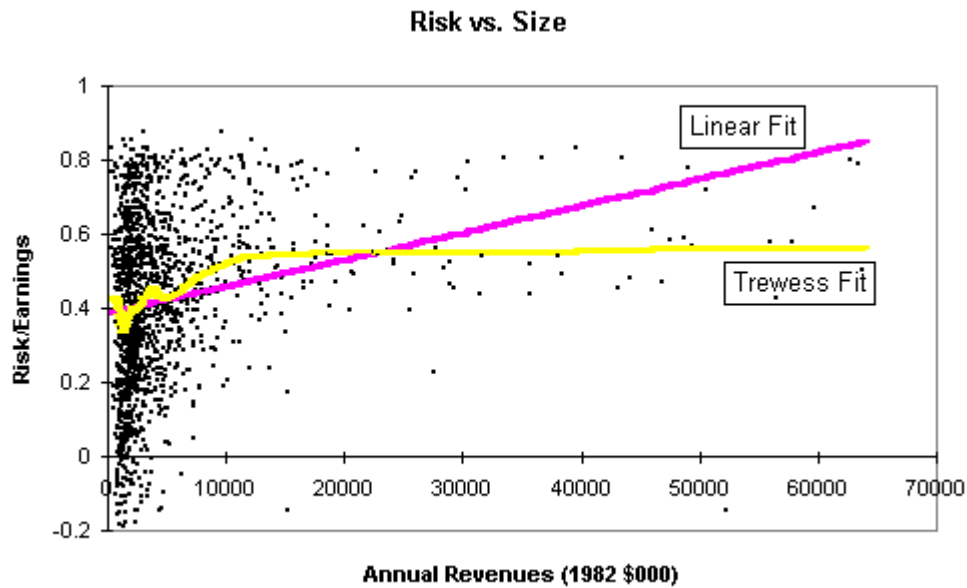


Figure 4

The Trewess fit yields interesting results: risk decreases with size for very small banks (< \$1.5 M in revenues) and then increases sharply with size up to about \$15 M in revenues, and is then relatively constant for larger banks.

The decrease of risk for very small banks is consistent with the results of McAllister and McManus (1993). The increase in risk with size is more puzzling. There are at least three possible explanations of the positive and significant risk-size relationship: (i) it is possible that as banks increase in size and layers of management, effective control of risk in the field is more difficult. (ii) It could also be that large banks become large by management's overly aggressive growth, which drives them to utilize excess capacity with riskier business. (iii) It may also be an optimal response to the so-called "too big to fail" doctrine; if bank managers and depositors of large banks are protected from losses, they are encouraged to undertake riskier actions. Equity holders are not so protected, so these risky actions are reflected in what equity investors are willing to pay for the shares of such banks.

The first two hypotheses would suggest that risk would increase with size without limit, which is not what we observe.¹⁴ The "too big to fail" hypothesis, however, is consistent with our observations; for banks between \$2.5 M and \$18 M there is some probability that in the event of a failure they will be bailed out, and this probability is increasing in size. The higher this probability, the more willing the bank managers are to engage in risky behavior, as they may avoid exposure to this risk via a bailout. Banks above \$18 M are virtually assured of a bailout, so that increasing size does not yield increases in risk-taking; these

banks are already at the maximum risk they wish to take, given the costs of bailouts. Note that this result is consistent with that of Hughes and Mester (1993).

Operating Results The estimation of equation (14) presents several problems. The first problem is the number of parameters and the estimating equation's nonlinearity. The full model has 99 structural parameters $(\alpha, \gamma, \lambda, \nu)$, 219 intercept dummies (B^k) and 219 slope dummies (m^k) . The full model is estimated using an iterative procedure; first, the dummy variables are assumed known and the 99 parameters are estimated using nonlinear least squares. Second, the estimated structural parameters are assumed known, the intercept dummies set equal to the bank average residuals, and the slope dummies are estimated using nonlinear least squares. These dummy values are then used to repeat the first step; the procedure is continued until convergence of all parameters is achieved.

The second problem is that while the operating results are available for 6190 data points, the quality results are only available for a subset of size 473, a reduction of over an order of magnitude. The loss of degrees of freedom involved in this reduction of sample size, especially with a nonlinear equation, is substantial. The full model was therefore estimated in four configurations: (i) the full 6190 sample, without quality; (ii) an annual sample of 1533 data points,⁵ without quality; (iii) the reduced sample of 473, without quality; and (iv) the reduced sample of 473 *with* quality. The structural parameter estimates were nearly identical for all four configurations; however, the standard errors of the estimates differed substantially, with the largest sample size yielding rather good estimates of many parameters. The stability of the parameter estimates suggested a sequential estimation strategy: estimate the operating parameters using the 6190 dataset, then reduce the fitted values to the 473 dataset to estimate only the quality parameters. In this section, therefore, the results of the operating estimation is reported.

Initial estimates of the full model yielded insignificant results for both the long-run and the short-run cost coefficients for the following cross-terms (refer to Section 2 for product numbers): (1,2), (1,3), (1,4), (2,4), (2,5), (2,6), (3,4), (3,5), (3,6), (4,5), (4,6).

Consequently, these terms were dropped in subsequent estimation. The overall regression results are

Table 1

R^2	0.850842
Adjusted R^2	0.849516
F-statistic	641.8394
log likelihood	-49045.23

Scale Effects Ray scale economies are defined in the usual way. For a scalar h and any output vector q , the long-run scale elasticity is defined as

$$S_{LR} = \frac{\sum_{i=1}^6 \sum_{j=i}^6 2g_{ij} l_{ij} (h^2 q_i q_j)^{g_{ij}}}{\sum_{i=1}^6 \sum_{j=i}^6 l_{ij} (h^2 q_i q_j)^{g_{ij}}},$$

with {constant returns, increasing returns, decreasing returns} to scale as S_{LR} is {=,<,>} 1. It is clear from this formula that if all $g_{ij} \geq 0.5$, then there cannot be increasing returns to scale. If the inequality is strict for at least one coefficient, then there are decreasing returns to scale. Inspection of the left-hand side of Table 2 shows that this is indeed the case. All coefficients are either not significantly different from 0.5, or they are significantly greater than 0.5 (g_{11} and possibly g_{23}).

The scale analysis of the short-run cost functions is complicated by the fact that the fixed costs lead to a weak form of scale economies: “spreading the fixed cost.” However, if we examine only variable costs, the left-hand side of Table 2 shows highly significant increasing marginal cost, with no coefficient significantly less than 0.5 and all the direct coefficients significantly above this number.

Table 2

	Estimate	Std. Error	p-Value*		Estimate	Std. Error	p-Value*
g_{11}	0.923659	0.114651	0.000110896	a_{11}	1.626377	0.548304	0.019996674
g_{22}	0.533569	0.126315	0.39521833	a_{22}	1.666517	0.489423	4.87509E-06
g_{33}	0.383381	0.12275	0.171063035	a_{33}	2.115984	0.175854	2.72262E-20
g_{44}	0.546865	0.038757	0.113318185	a_{44}	2.056322	0.780928	0.023159515
g_{55}	0.024372	51.18033	0.495912859	a_{55}	1.683526	0.889139	0.091606246
g_{66}	0.319622	0.215415	0.201215351	a_{66}	3.123559	0.525182	3.02229E-07
g_{16}	0.459534	0.07297	0.289609966	a_{15}	0.254758	14.18534	0.493103571
g_{23}	0.620312	0.090278	0.091344774	a_{16}	1.061438	0.143995	4.88542E-05
g_{56}	0.473472	0.126587	0.417008016	a_{23}	1.190735	0.04016	5.83302E-65
				a_{56}	-0.755504	17103787	0.5

* probability that the true value of the coefficient lies on the opposite side of 0.5 than the point estimate.

The short-run scale parameters are all substantially greater than 0.5, indicating significant increasing marginal costs in the short run. This appears consistent with the results of Kaparakis, Miller, and Noulas (1994).

Scope Economies A necessary and sufficient condition for long-run cost complementarities between products i and j is that $I_{ij} < 0$, and for short-run cost complementarities, $n_{ij} < 0$. The estimation results are not consistent with the existence of cost complementarities:

Table 3

	Coefficient	Std.Error	p-value		Coefficient	Std.Error	p-value
I_{15}	0.09333	0.154031	0.5446	n_{14}	0.460942	136.5392	0.9973
I_{16}	-0.01545	0.032797	0.6376	n_{15}	4.59E-07	1.58E-06	0.7707
I_{23}	0.001395	0.003181	0.6609	n_{16}	1.05E-07	9.50E-08	0.267
I_{56}	0.016984	0.05546	0.7594	n_{36}	0.157856	25732798	1.0000

None of the cross-terms are significantly different from zero; these results are consistent with the hypothesis that there are no cost complementarities, either short-run or long-run, in banking. Note that the existence of short-run fixed costs implies that simply spreading these fixed costs across several product lines does give rise to a weak form of scope economies.

Demand Elasticities Also estimated is the matrix of demand self- and cross-flexibilities η . Only six of the 21 flexibilities are significant at the 10% level, so that the results are suggestive rather than significant.

In order to present the results in a more familiar format, the inverse $\epsilon = \eta^{-1}$ of self- and cross-elasticities is shown in Table 4.

Table 4 - Elasticity Matrix

	1	2	3	4	5	6
1	0.421107	0.29976	-0.15746	0.305835	0.007331	0.194665
2	0.29976	0.658294	0.134493	-0.10661	-0.23474	0.089791
3	-0.15746	0.134493	0.02365	-0.06503	-0.06634	0.485121
4	0.305835	-0.10661	-0.06503	0.275833	-0.06198	0.405255
5	0.007331	-0.23474	-0.06634	-0.06198	0.067438	0.157416
6	0.194665	0.089791	0.485121	0.405255	0.157416	0.280526

Bias Parameters The bank-specific bias parameter represents the ratio of planned-for demand to expected demand for each bank. The theory of the model suggests that *ex post* costs are minimized if this ratio is

one, and costs increase as this ratio increases or decreases. A bias parameter is separately estimated for each bank. Of the 219 banks in the sample, 218 had sufficient data to estimate m^k . The distribution is shown below:

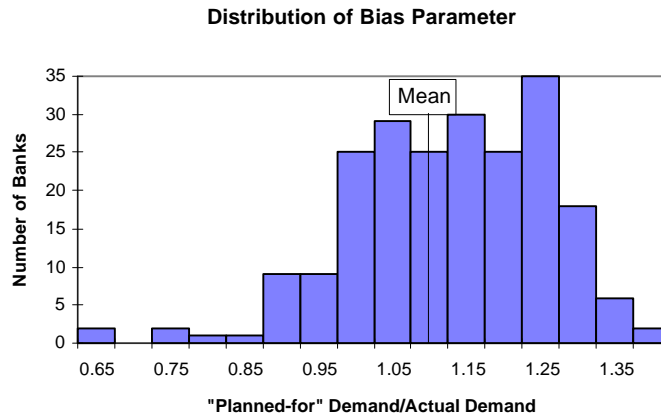


Figure 5

The mean ratio of planned-for demand to expected demand is 1.10, indicating substantial overcapacity on average. While a few banks are seriously pessimistic in their forecasting, most banks appear rather aggressive in installing capacity to meet future needs. 76% of the banks had bias parameters significantly greater than 1 at the 90% level.

Before concluding that US banks are deficient in capacity planning, two questions must be answered. First, is this mismatch of demand and capacity an efficient response to forecast uncertainty? In Appendix A, a model of optimal capacity choice is developed in which banks may optimally choose a capacity level that does not correspond to expected demand. A few key results from the Appendix A analysis: (i) this planned-for demand need not equal expected demand; under simplifying assumptions, it is optimal to provide for {more than, less than, the same as} the expected quantity, as $a_{ii} \{>, <, =\} 0.5$; and (ii) under reasonable bounding assumptions about a_{ii} and forecast error, the theoretical difference between planned-for demand and expected demand is quite small, most likely less than 1/2%. A testable hypothesis of this paper is that banks have consistent biases in their capacity planning; some banks overestimate how much demand they will serve while others underestimate. The results of Appendix A suggest that such “efficient” over (or under) capacity is likely to be very small. Therefore, our empirical finding of substantial mismatches between capacity and demand is almost surely the result of inefficient bias, and not an efficient response to forecast uncertainty.

It is also possible that mismatches between demand and capacity come about because the underlying process is non-stationary, and thus particularly costly to forecast well. However, the results of Appendix A suggest that even a simple-minded forecasting method seems to outperform the bank average, suggesting that it is not the difficulty of the forecasting problem that is leading to these mismatches.

Second, does this mismatch matter? It could be that such mismatches result in only a very small increase in cost, so it would not be optimal to invest much effort into getting the forecast right. A direct test of this would be to examine the ratio of short-run to long-run costs as computed from the model. Unfortunately, the fact that many of the parameters of the nonlinear model are not statistically significant leads to this ratio being more noise than information; in particular, this calculated ratio is less than one for about 35% of the observations, contrary to the theory. In a less direct method of assessing the cost of bias, the ratio of total cost to total revenues was regressed against a quadratic function of the bias parameter (assuming orthogonality with other coefficients), with the following results:

Table 5 - Cost/Revenue vs. Bias

	Coefficient	Std. Error	T-Statistic
Constant	1.622114	0.166863	9.721210
m	-1.465286	0.314564	-4.658150
m^2	0.743318	0.147404	5.042740

While the theory suggests that this function should be minimized at $m = 1$, there is nothing in the data that forces this to be the case. Therefore, it is surprising that the minimum of this empirically fitted function actually does occur at 0.986, with an optimal value of the cost/revenue ratio of 0.9, as shown in Figure 6.

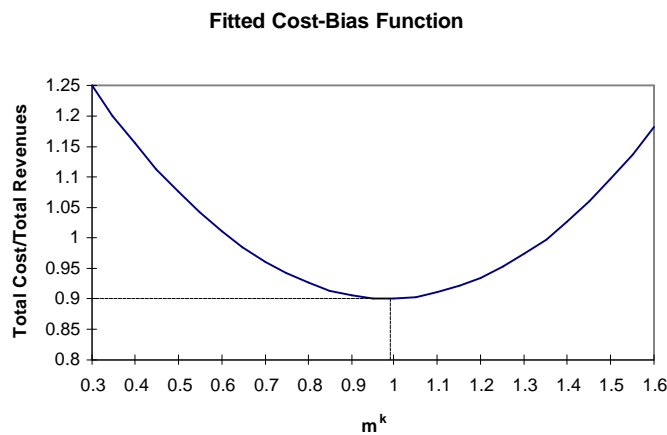


Figure 6 - Fitted Cost/Revenue vs. Bias

This suggests that biased forecasts can be costly, especially those of large magnitude. On balance, however, the quadratic model applied to the bias parameter estimates yields the result that the average cost increase associated with biased capacity planning amounts to 2.2% of costs (with standard deviation of 2%). While this appears small relative to the 10% average over-forecasting, it is in fact a rather large number compared to the average bank margin of 7.8% (earnings/costs). On average, over 25% of bank earnings are absorbed by the costs of the allocative inefficiency of demand/capacity mismatches.

However, allocative inefficiency is not the only possible cost of poor forecasting. Since most banks seem to err on the side of overcapacity, banks may increase their risk as a result. Excess capacity and low marginal cost may tempt bank managers to lower their credit standards to find enough customers to “fill the pipeline.” This hypothesis implies that the bias parameter and the risk/earnings ratio will be positively correlated. The data does not confirm this hypothesis, however; the relationship is actually somewhat negative.

It is also possible that excess capacity could translate into higher quality; more branches, for example, could lead to greater convenience for customers, who might then be willing to pay a price premium and/or give more business to the bank. For C&I loans, we have a direct measure of quality, which is not correlated with the bias parameter. For all other products, the bias parameter shows either no correlation with prices and quantities or else a small negative correlation. Thus, higher capacity is not correlated with quality, either measured directly or via prices and quantities. The bias parameter thus represents allocative inefficiency.

Fixed Effects Past research using fixed effects and frontier methods showed that there was a significant amount of unexplained cost/profit variation among banks. A primary objective of the structural modeling of this paper is to explain as much of this variation as possible. To the extent this objective is met, any remaining fixed effects should be small and/or insignificant. In fact, we find that very few of the estimated fixed effects in this model are statistically significant. The distribution of the t -values for the fixed effect coefficients is shown below in Figure 7.

Distribution of Fixed Effect p-Values

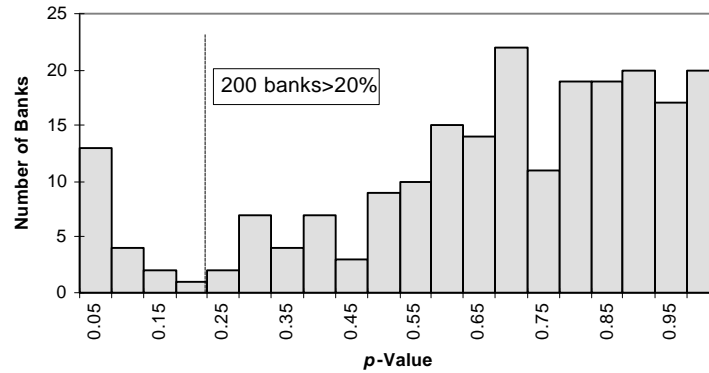


Figure 7

Only 13 of the 219 fixed effect coefficients are significant at the 5% level, and only 19 are significant at the 20% level. It is likely that the most significant thirteen fixed effect coefficients do represent bank differences, but the remaining banks (96% of banks) show no variation.

In view of the economic significance of fixed effects/frontier methods found by the recent literature, this result is quite strong. It suggests that the unobserved efficiency differences which caused substantial variations among banks have by and large been captured in the structural model presented here. The inefficiencies in risk management and capacity planning observed in the structural model appear to reflect virtually all of the inefficiencies that were manifested in previous research via fixed effects/frontier methods.

To test the hypothesis that it is the structure of the model that accounts for the relatively insignificant fixed effect coefficients, a simple reduced form model, with fixed effects⁶, was estimated:

$$C = \sum_{j=1}^6 x_j q_j + \sum_{k=1}^{219} \bar{B}^k .$$

The reduced form estimation was carried out using the quarterly dataset and the same cost and quantity data used in the structural estimation. In the reduced form model, the fixed effects coefficient \bar{B}^k are much more significant than in the structural model; 128 banks (58%) had coefficients significant at the 20% level in the reduced form model, compared with 19 banks (10%) in the structural model. The distributions of p-values of B^k (structural model) and \bar{B}^k (reduced form) are shown in Figure 8:

Distribution of Fixed Effect p-Values

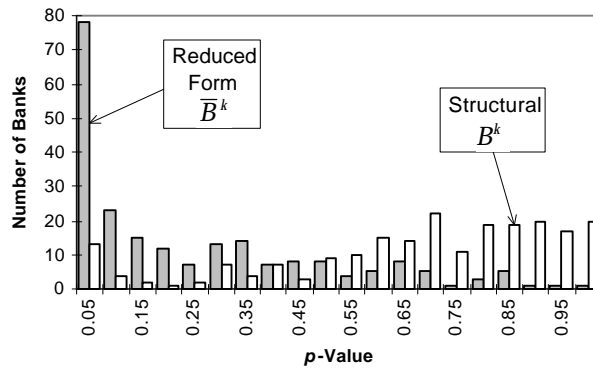


Figure 8

This comparison¹⁷ between the two distributions of p -values makes it clear that the very substantial significance of the fixed effects coefficients in the reduced form model by and large fades away in the structural model.¹⁸ What is unexplained variation in the reduced form model is explained using the structural model.

Other Estimation Problems It will be recalled that the cost function did not include input prices, as theory would suggest. This formulation is valid as long as input prices are constant across time and across banks. In order to test the time-independence assumption, the model was estimated using year dummies; all year coefficients were zero. This suggests that using constant dollars captures the full effect of time in the operating results estimation, in contrast to the risk estimation in which year is important.

However, there may still be variation in input prices across regions of the country. While the cost of capital and other materials is not likely to vary by region, it is possible for both labor costs and land costs to vary considerably. Two regions of the US suggest themselves as high-cost candidates: New York and California. To test for such differences in costs, the model was estimated using dummy variables to represent location in one of these regions. Chicago was also included as a relatively low-cost region. All three regional coefficients were zero. These results are consistent with the hypothesis that input prices can be safely ignored.

Finally, the nonlinear and iterative nature of the estimation is cause for concern that the parameter estimates may depend upon the starting points. Although the size of the problem precluded extensive testing, all results were verified by starting the estimation process at least three different initial values, with no effect on the outcome. It should be recalled, however, that the variance-covariance matrix derived

from nonlinear regressions is at best an approximation to the true matrix. Therefore, the quoted standard errors should be viewed with this reservation in mind.

Quality Estimation The fact that many banks are willing to pay Greenwich Associates for customer satisfaction survey data year after year suggests that quality is important to banks. There are several ways in which the effect of quality on a bank is realized: cost, price, and demand.

Cost As discussed above, the operating results estimation was completed on the full 6190-observation dataset. The fitted model was used to evaluate two terms: those relating to costs of C&I loans (C_3) and all other costs (C_{-3}):

$$C(\mathbf{q}) = F + \sum_{i \neq 3} c_i q_i^{2a_{ii}} + \sum_{i \neq 3} \sum_{j \neq i} n_{ij} (q_i q_j)^{a_{ij}} + X^d \left(c_3 q_3^{2a_{33}} + \sum_{j \neq 3} n_{3j} (q_3 q_j)^{a_{3j}} \right)$$

$$C(\mathbf{q}) = \mathbf{q} (C_{-3} + X^d C_3)$$

This equation was estimated on the 476-observation dataset, with the following results:

Table 6 - Cost-Quality Estimation

	Coefficient	Std. Error	T-statistic	p-value
q	1.769108	0.049293	35.88932	0.0000
d	0.011379	0.003201	3.554510	0.0000

R^2	0.702536	Adjusted R^2	0.701908
log likelihood	-4469.949	F-statistic	1119.470

Clearly, quality increases cost but not very much. The elasticity parameter is statistically significant but not economically significant; doubling quality only increases costs by 1%. For example, increasing quality from the median level of 51 to the 75th percentile level of 59, a 15% change, increases costs by on 0.15%.

Price High quality firms are often able to command a price premium for their services. Four-star hotels, high-end jewelers, and other retailers seem to earn rents from the higher prices they are able to charge. In some cases, though, higher quality commands only a very small price premium, but does lead to higher volume. Some department stores well-known for quality stay price-competitive but handle higher volume with greater efficiency.

The model $\log p_3 = q + f \log X$ was estimated, with the following results:

Table 7 - Price-Quality Estimation

	Coefficient	Std. Error	T-statistic	p-value
q	0.039351	0.019669	2.000693	0.0460
f	0.010535	0.005031	2.093846	0.0368

R^2	0.009165	Adjusted R^2	0.007074
log likelihood	972.0862	F-statistic	4.384192

Clearly, there are many factors other than quality which affect price. It is the case that quality does affect price positively (as is expected); again, the result is statistically significant but not economically significant; doubling quality permits a mere 1% price premium.

However, though both the quality elasticity of cost and of price are small, they are statistically significant and it is of interest to note that they are essentially equal. The small increase in cost from higher quality is recovered in the equally small increase in the price premium the bank can charge for quality; margins do not suffer at all with higher quality.

Quantity Does quality affect the quantity demanded? The estimation of this relationship is substantially more difficult. A simple regression of quantity on quality does indeed lead to a strongly positive and significant elasticity; however, this may merely mean that large banks are better at providing good service. In order to sort out whether quantity leads to greater quality (economies of scale in quality provision) or quality leads to greater quantity (a market response to better service), several models were estimated that included proxies for size as well as the quality measure.

The literal interpretation of the model suggests that the only reason banks differ in size is because of their capacity choices. Thus, the first model uses the bias parameter as a size proxy:

$\log q_3 = q_0 + q_1 \log X + q_2 \log m^k$. The second model uses total revenues as a size proxy in linear form: $q_3 = q'_0 + q'_1 X + q'_2 R$. The third model (unreported here) uses a combination of

q_1 , q_2 , and q_6 in quadratic form, with equally strong results. The estimated coefficients from the first two models:

Table 8 - Quantity-Quality Estimation

	Coefficient	Std. Error	T-statistic	p-value
Model 1				
q_0	5.954614	0.793794	7.501460	0.0000
q_1	0.980932	0.203279	4.825551	0.0000
q_2	-3.058770	0.510000	-5.997588	0.0000
Model 2				
q'_0	-24122.57	6444/394	-3.743187	0.0002
q'_1	607.5021	125.4451	4.842772	0.0000
q'_2	8.252161	0.266152	31.00547	0.0000

Using different proxies for size all lead to strongly positive significant elasticities, ranging from 0.16 to 0.98.

This set of results is consistent with the hypothesis that though banks cannot command a price premium for quality, it is a means of attracting and keeping customers. Once a bank adapts itself to a high-quality technology, they can realize greater demand at about the same cost.

If quality is effectively free and it generates a demand flow, then why don't all banks provide high quality service to their customers? The previous discussion argued that the technology for providing quality service involved a difficult adaptation by the bank and so is diffusing slowly in the industry. Successful adopters can achieve the benefits at almost no cost, so profits can be earned during this diffusion period.

Profit The above results suggest that quality should have a positive impact on profit. We test this directly by estimating a simple profit-quality relationship. Once again, size must be controlled for. Both revenues and the bias parameter were used as size controls, with similar results. We report on the model

$$p' = q_0 + q_1X + q_2R:$$

Table 9 - Profit-Quality Estimation

	Coefficient	Std. Error	T-statistic	p-value
q_0	-542.0245	441.7871	-1.226891	0.2205
q_1	16.04065	8.599727	1.865251	0.0628
q_2	0.065855	0.018246	3.609369	0.0003

Quality has the expected sign and is significant at the 6% level.

6. Conclusions

A model of banking markets is developed which explicitly incorporates risk, customer service quality, capacity planning, short-run and long-run cost functions, and market equilibrium. The model is estimated in structural form, using data from operating results, capital markets, and customer surveys. The principal findings are:

- (i) Banks differ widely in their ability to manage risk; larger banks take on relatively more risk; on average, risk cost accounts for 38% of bank earnings.
- (ii) There are substantial inefficiencies due to demand/capacity mismatches. On average, banks are over-optimistic by 10% in the demand they plan for, and this cost them about 2.2% of total costs. This is substantially more than can be justified by “optimal overshooting” in the face of planning uncertainty, and amounts to over 25% of average bank margin.
- (iii) Greater customer satisfaction correlates with greater profitability, principally due to greater demand; the effect of quality on cost and price is minimal.
- (iv) Bank-specific fixed effects are relatively insignificant; the very significant bank-specific effects that previous research discovered appear to have been largely captured and directly estimated in the structural model.
- (v) Confirmation of the results of previous research that there are no significant long-run economies of scale or scope.

The paper employs methodological innovations as well:

- (i) The Capital Asset Pricing Model is used to measure bank risk, thereby capturing all risk in a measure based on market behavior.
- (ii) The ability of banks to satisfy their customers is examined for the first time in this literature, and it is shown to be a source of profitability; a new dataset is introduced to study this phenomenon.
- (iii) Structural model estimation is used to expand the range of questions that can be empirically addressed; it is also used to explain and directly estimate inefficiencies hitherto captured only by inference in fixed effects models.

-- References --

- Berger, Allen N. (1993) "'Distribution-Free' Estimates of Efficiency in the US Banking Industry and Tests of the Standard Distributional Assumptions," *J. Productivity Analysis* **4**, 261-292.
- Berger, Allen N. and David B. Humphrey (1991) "The Dominance of Inefficiencies Over Scale and Product Mix Economies in Banking," *J. Monetary and Financial Econ* **28**, 117-148.
- Berger, Allen N. and David B. Humphrey (1992) "Measurement and Efficiency Issues in Commercial Banking," in Z. Griliches, ed., *Measurement Issues in the Service Sectors*, National Bureau of Economic Research (University of Chicago Press, Chicago, IL).
- Berger, Allen N., Diane Hancock, and David B. Humphrey (1993) "Bank Efficiency Derived from a Profit Function," *J. Banking and Finance* **17**, 317-347.
- Berger, Allen N., William C. Hunter, and Stephen G. Timme (1993) "The Efficiency of Financial Institutions: a Review and Preview of Research Past, Present, and Future," *J. Banking and Finance*, **17**, 221-249.
- Berger, Allen N., John H. Leusner, and John J. Mingo (1994) "The Efficiency of Bank Branches," Wharton Financial Institutions Center Working Paper 94-27.
- Economides, Nicholas (1993) "Quantity Leadership and Social Inefficiency," *Intl J. Industrial Organization*, **11**, 219-237.
- Grabowski, Richard, Nanda Ragan, and Rasoul Rezvanian (1994) "The Effect of Deregulation on the Efficiency of US Banking Firms," *J. Economics and Business* **46**, 39-54.
- Hassan, Kabir M., Gordon V. Karels, and Manfred O. Peterson (1994) "Deposit Insurance, Market Discipline, and Off-Balance Sheet Banking Risk of Large US Commercial Banks," *J. Banking & Finance*, **18**, 575-593.
- Hughes, Joseph P. And Loretta J. Mester (1993) "A Quality and Risk-Adjusted Cost Function for Banks: Evidence on the "Too-Big-To-Fail" Doctrine," *J. Productivity Analysis* **4**, 293-315.
- Hughes, Joseph P. and Loretta J. Mester (1994) "Bank Managers' Objectives" Federal Reserve Bank of Philadelphia Working Paper 94-8.

- Jagtiani, Julapa, Alli Nathan, and Gordon Sick (1994) "Scale Economies and Cost Complementarities in Commercial Banks : On- and Off-Balance-Sheet Activities," New York University Stern School Working Paper S-94-7.
- Kaparakis, Emmanuel I., Stephen M. Miller, and Athanasios G. Noulas (1994) "Short-run Cost Inefficiency of Commercial Banks: A Flexible Stochastic Frontier Approach," *J. Money, Credit, and Banking* **26**, 875-893.
- Leibenstein, Harvey (1966) "Allocative vs. 'X-Efficiency'," *American Economic Review*, **56**, 392-415.
- Mester, Loretta J. (1993) "Efficiency in the Savings and Loan Industry," *J. Banking and Finance* **17**, 267-286.
- Mester, Loretta J. (1994) "How Efficient Are Third District Banks?" *Business Review*, (Jan/Feb) 3-18.
- McAllister, Patrick H. And Douglas McManus (1993) "Resolving the Scale Efficiency Puzzle in Banking," *J. Banking and Finance* **17**, 389-405.
- Pulley, Lawrence B. and David B. Humphrey (1993) "The Role of Fixed Costs and Cost Complementarities in Determining Scope Economies and the Cost of Narrow Banking Proposals," *J. Business* **66**, 437-462.
- Westmore, Jill L. and John R. Brick (1994) "Commercial Bank Risk: Market, Interest Rate, and Foreign Exchange," *J. Financial Research* **17**, 585-596.
- Vavra, Terry G. (1995) "Selling After the Sale," *Bank Marketing*, **27**, 27-30.
- Velleman, Paul F. (1980), "Definition and comparison of robust nonlinear data smoothing algorithms," *J. Amer. Statist. Assoc.*, **75**, 609-615.

Appendix A

A Model of Optimal Capacity Choice Based on Demand Forecasting Bank k has available to it certain *public* information I and certain *private* information I^k regarding their demand for the next period. Using this information, bank k estimates the cumulative distribution function of its next-period demand as $G(\mathbf{x}; I, I^k) = G^k(\mathbf{x})$. The parameters depend upon both public and private information, so they need not be the same for all banks.

Bank k chooses the technology $(\tilde{F}^k, \tilde{\mathbf{c}}^k)$ that minimizes its expected cost (assuming risk neutrality):

$$(\tilde{F}^k, \tilde{\mathbf{c}}^k) = \arg \min_{F, \mathbf{c}} \int_{-\infty}^{\infty} \left(F + \sum_{i=1}^6 c_i x_i^{2a_i} + \sum_{i=1}^5 \sum_{j=i+1}^6 \mathbf{n}_{ij} (x_i x_j)^{a_{ij}} \right) dG^k(\mathbf{x}) \quad (\text{A15})$$

To simplify the analysis, we consider the case in which all products are independent, both in the economic sense ($\mathbf{n}_{ij} = \mathbf{0}$) and in the statistical sense ($dG^k(\mathbf{x}) = g^k(\mathbf{x}) = \prod g_i^k(x_i)$). In this case, the short-run and long-run cost functions are (removing redundant double subscripts)

$$C(\mathbf{q}; F, \mathbf{c}) = F + \sum c_i q_i^{2a_i}, \quad C(\mathbf{q}) = \sum I_i q_i^{2g_i}.$$

For each $\mathbf{q} > \mathbf{0}$, there is a corresponding (F, \mathbf{c}) , derivable from the conditions

$$C(\mathbf{q}; F, \mathbf{c}) = C(\mathbf{q}), \quad \nabla C(\mathbf{q}; F, \mathbf{c}) = \nabla C(\mathbf{q}):$$

$$c_i = \frac{I_i g_i}{a_i} q_i^{2(b_i - a_i)}, \quad F = \sum (I_i q_i^{2g_i} - c_i q_i^{2a_i}) = \sum \left(I_i q_i^{2g_i} \left[\frac{a_i - g_i}{a_i} \right] \right). \quad (\text{A16})$$

Evaluating the expected cost in the single-product case, we obtain

$$\int (F + \sum c_i x_i^{2a_i}) dG^k(\mathbf{x}) = F + \sum c_i \int x_i^{2a_i} g_i^k(x_i) dx_i = \sum \left(I_i q_i^{2g_i} \left[\frac{a_i - g_i}{a_i} \right] + \frac{I_i g_i}{a_i} q_i^{2(b_i - a_i)} \int x_i^{2a_i} g_i^k(x_i) dx_i \right).$$

Differentiating this with respect to q_i and setting the result equal to zero yields

$$2I_i g_i q_i^{2g_i - 1} \left(\frac{a_i - g_i}{a_i} \right) = 2I_i g_i q_i^{2g_i - 2a_i - 1} \left(\frac{a_i - g_i}{a_i} \right) \int x_i^{2a_i} g_i^k(x_i) dx_i$$

so that the optimal ‘‘planned-for’’ demand is

$$\bar{q}_i^k = \left(\int x_i^{2a_i} g_i^k(x_i) dx_i \right)^{\frac{1}{2a_i}}. \quad (\text{A17})$$

The actual technology (F, c) can be determined by substituting (A3) into (A2).

Clearly, for $a_i = 0.5$ (short-run constant marginal cost), the optimal capacity to install is that which corresponds with the mean demand: $\bar{q}_i^k = \bar{q}_i^k$. It is easy to show that for $a_i = 1$ (short-run increasing marginal cost), the optimal capacity corresponds to planned-for demand greater than the mean demand:

$\bar{q}_i^k = \left(\text{Var } q_i^k + (\bar{q}_i^k)^2 \right)^{\frac{1}{2}} > \bar{q}_i^k$, which is increasing in $\text{Var } q_i^k$. In addition, planned-for demand is an

increasing function of a_i , since $\frac{\partial \bar{q}_i^k}{\partial a_i} = \left(\int x_i^{2a_i} g_i^k(x_i) dx_i \right)^{\frac{1-2a_i}{2a_i}} \cdot \int x_i^{2a_i-1} g_i^k(x_i) dx_i > 0$, for all $a_i >$

0, assuming the support of g_i^k is strictly positive. We can therefore conclude that in the independent product case, it is optimal to provide for {more than, less than, the same as} the expected quantity, a_i {>,<,<=} 0.5.

If banks are making unbiased demand forecasts (in the sense that on average expected demand equals actual demand), then observed mismatches between actual demand and installed capacity may be efficient in that the mismatch may be an efficient response to forecast uncertainty and the nonlinear cost function. On the other hand, such mismatches could also be due to consistently biased demand forecasts, which would be inefficient. This could occur if, for example, bank managers are consistently over-optimistic about next period demand, consistently installed excess capacity as a result, and did not learn from past mistakes. Is it possible to distinguish between efficient mismatches of demand and capacity and inefficient mismatches?

Some simple numerical examples suggest that efficient mismatching of demand and capacity is unlikely to be significant in practice. In the empirical results below, we find the average a_i to be 2.04. To bound the coefficient of variation of the forecast distribution, we analyzed the simplest possible forecast method: next period's demand will equal last period's demand. Applying this method to the data for all six products yields an overall empirical coefficient of variation of 0.085. Assuming a normal forecast distribution, this leads to planned-for demand 1.1% larger than expected demand. Of course, this simplest possible forecast method should be easy to beat; if bank forecasters can reduce the coefficient of variation by half (a very modest target, it would seem), then planned-for demand is only 0.3% larger than expected demand. In the case at hand, efficient mismatching of demand and capacity are not significant; in the following, mismatching of demand and capacity will be attributed to inefficiently biased forecasts.

Appendix B

The Greenwich Associates' Dataset

Greenwich Associates is a strategic research and consulting firm for financial service providers, offering over 60 programs in commercial banking, investment management, stockbroking, bond dealing, investment banking, and foreign exchange dealing. The firm has been in business and conducting financial market research since 1972.

Each year Greenwich Associates conducts 32,000 interviews with senior financial officers at corporations in 19 countries. The data collected from these interviews captures qualitative information on the satisfaction of clients with their current providers of financial services.

Research in the US corporate credit market is split into three programs:

Large Corporate Banking research (covering companies with over \$500 million in sales) has over 1200 respondents. Nearly 75% of the population is covered and the research is conducted annually.

Middle Market Banking (companies with sales of \$50-500 million) interviews officials at over 3000 companies. Roughly half of the population is covered and the research is conducted every odd numbered year.

Commercial Banking (companies with sales of \$5-50 million) analyzes the performance of over 300 banks on a state-by-state basis. Approximately 20,000 interviews are conducted, giving a coverage rate of 20% of the population. Regions covered rotate in a two year cycle.

Bank clients are asked to describe their relationship with the specific bank; the types of relationship are:

Customer: This means that the corporation has been a customer of the bank over the period covered by the survey.

Principal Bank: The bank is one of the corporation's three or four most important banks.

Lead Bank: The bank is the corporation's main or first bank.

Data collected for this study from the Large Corporate and Middle Markets was based on relationships of type "Principal Bank," while data from Commercial Banking was based on relationships of type

“Customer.” So the Competitive Situation Report for Large Corporate Banking, for example, would show what percentage of clients from relationship type “Principal Bank” indicated a positive response to a question.

As might be expected, the survey instrument is quite large, and varies from year to year as well as from segment to segment. In order to ensure as much uniformity across the sample as possible, eight specific questions were selected. These questions appear in all survey instruments in all years in all market segments. On the basis of a principal components analysis, Greenwich Associates avers that these questions capture the key elements of customer satisfaction. In summary form, these questions are:

A) How Do You Rate This Bank in:

- i-willingness to lend
- ii- competitive loan pricing
- iii- cash management capabilities
- iv- international service capabilities

B) How Do You Rate This Bank’s Account Officers in:

- i-convincing bank to meet credit needs
- ii- prompt follow-up
- iii- knowledge of cash management services
- iv- advice on corporate finance

For each question, the respondent had five choices:

- 1- Poor 2- Fair 3- Good 4- Very Good 5-Excellent

For each of the items above, the average percent of respondents who rated the bank a “4” or “5”, referred to as “Above Average,” is recorded, and constitutes the core data of this study.

The dataset consists of observations on 112 banks over several years between 1985 and 1992, resulting in a total of 476 usable observations. Each observation consists of data drawn from the three market segments, each of which has:

- 1) the number of customer respondents;
- 2) for each of the eight questions, the percentage of all respondents who rated the bank “Above Average”; and
- 3) the average revenue for respondents in this segment (from Dun and Bradstreet).

The overall index is constructed as follows:

- For each market segment (Large Corporate, Middle Market, and Commercial), the average of (2) above for all eight questions was computed. In each segment, this measure was Z-score normalized.
- The three normalized indices were combined by weighting them by the average size of clients in each market and the number of clients each bank had in each market and then adding them together.
- This produced an index with mean zero and a unit standard deviation. The range of the index was from -4.23 to 4.40.

This index was then re-scaled, so that -4.23 was mapped to “1” and 4.4 was mapped to “100.” This yields a resulting index ranging from 1 to 100, with a mean of 50.

-- Notes --

¹ The extensive literature on the efficacy of bank mergers is included in this category as well.

² Not all bank risk is captured in the shareowners' perspective; the fact that failed banks are frequently bailed out transfers some default risk from shareowners to taxpayers.

³ There are, of course, many ways in which quality can be defined. Operational measures, such as mishandled checks, data entry errors, etc., are used by virtually all banks to monitor the quality they provide their customers. Customer-based measures, such as overall satisfaction, are more subjective and perhaps less precise, as they involve interview methods. However, it is our view that quality is in the eye of the customer, not the process engineer; consequently, we opt for the latter measure: customer satisfaction is quality, for our purposes.

⁴ Omitting futures contracts from this analysis may bias the risk results, in that such contracts are often used to hedge a bank's other assets.

⁵ For ease of exposition, we postpone consideration of the measure of quality and how it affects the short- and long-run cost functions until the section below on Customer Satisfaction.

⁶ Using markets to measure risk of banks, rather than accounting ratios, regulatory measures, or "industry-watcher" ratings, has both strengths and weaknesses. Its strengths:

Market risk measures are *forward-looking* rather than based on historical evidence.

Market measures are determined via transactions among buyers and sellers who have a high stake in getting their information right.

Market measures incorporate *all* the risk of a bank: market risk, business risk, asset risk, liability risk, and so forth. It also measures *net* risk, in that risks undertaken which are appropriately hedged are not counted (as opposed to certain regulatory measures).

Market measures are easily observed and do not rely on judgments of "experts" who may be unreliable or not disinterested. Such measures are difficult to "fudge."

On the other hand, there are weaknesses as well:

Internal measures of risk within banks, at the transaction level, the departmental level, or even the corporate level, may not easily translate into an external market measure. Therefore, disaggregation of market measures to operating units and validation against internal risk measures will prove difficult.

Market measures are only available for institutions with tradable securities, which in many cases may be holding companies. The performance of banks held by such firms may not be adequately reflected in the risk measures of the parent.

Banks typically disclose significantly less information than non-bank firms, and therefore investors have less information on which to base their trades of bank stocks.

To the extent that Federal bailouts shield shareowners from some default risk, and debtholders from almost all such risk, some idiosyncratic risk may not be captured in this market measure.

On balance, we conclude that using the bank's stock market beta is likely to be the most reliable measure of aggregate bank risk available.

⁷ Let r_f be the risk-free rate (say, on short-term US Treasury bills) and r_m be the rate of return on the market as a whole (say, the S&P 500). The bank's \mathbf{s} satisfies the Capital Asset Pricing Model equation: $\mathbf{s} = r_f + \mathbf{b}(r_m - r_f)$.

⁸ If we were willing to forego bank-specific coefficient estimates for total-sample estimates, then it is obvious a richer model could be estimated. However, we suspect that the interesting variation is among banks rather than among products; we prefer an approximate solution to the more interesting problem rather than a more precise solution to the less interesting problem.

⁹ Again, we postpone discussion of quality and how it affects the demand functions until the section below on Customer Satisfaction.

¹⁰ By way of analogy, consider the technology shift that occurred in Japanese automobile manufacturing firms in the 1970s. To a first approximation, this shift could be characterized as from a "do it fast, fix your mistakes later" model to a "do it right the first time" model. At the time of its adoption by Japanese firms, it appears that this new technology permitted both higher quality and lower cost, enabling them to compete well against the then-dominant American firms. It might be thought that this new approach would be quickly adopted by American firms as they lost domestic market share. Such was not the case; it appears that successfully adopting this technology requires substantial changes in work and management practices which American firms were only able to effect in the 1990s, which implies a diffusion delay of nearly two decades.

¹¹ and possibly real estate loans as well. For expositional purposes, we shall assume it affects only C&I loans.

¹² The notion that firms make capital choices prior to making price and output choices is common in both theory and practice. For example, Economides (1993) reviews this point, quoting various recent papers.

¹³ In order to smooth out variations in market-wide observables, we estimated the difference $r_m - r_f$ from the data as 0.088, which we use throughout. In addition, several methods for estimating \mathbf{g} were tried, including a linear trend and an exponential trend of the actual data. In both cases, a substantial number of negative growth rates, as well as growth rates that led to a negative value of the firm, were encountered. As a result, the growth rate was not used in computing the risk cost.

¹⁴ The second hypothesis is disconfirmed below.

¹⁵ The annual sample was included to determine if the estimation was sensitive to the elimination of non-end-of-year data. Some have suggested that banks are less serious about quarterly results than they are about end-of-year results, and may even view such intermediate results as "window dressing," shading their estimates optimistically. Our results do not support such allegations.

¹⁶ Berger (1993) compared using fixed effects (his "Method 2") to measure X-inefficiency with using average residuals ("Method 1"). While recognizing the limitations of average residuals, he finds them

superior to fixed effects in that the latter appeared to capture scale effects which should be picked up by the model parameters, at least in his study. We do not find that result in this analysis. Nevertheless, Figure 8 was re-calculated using the p -values of average residuals, and the result is essentially the same: far greater significance for the reduced form model than the structural model.

¹⁷ Comparison of the significance of the fixed effects coefficients with the significance of fixed effects (or average residuals or frontiers) from previous research is not possible, as none of the studies of which the author is aware report the standard errors of these estimates.

¹⁸ The difference in the significance of the fixed effects in the two models is also seen when each model is tested for the hypothesis that *all* fixed effects are zero (clearly rejected in both cases). For the structural model, the resulting F -statistic is 1.28; for the reduced form model, the F -statistic is 6.69.