Volume Title: Output Measurement in the Service Sectors

Volume Author/Editor: Zvi Griliches, editor

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-30885-5

Volume URL: http://www.nber.org/books/gril92-1

Conference Date: May 4-5, 1990

Publication Date: January 1992

Chapter Title: Measurement and Efficiency Issues in Commercial Banking

Chapter Author: Allen N. Berger, David B. Humphrey

Chapter URL: http://www.nber.org/chapters/c7237

Chapter pages in book: (p. 245 - 300)

# 7 Measurement and Efficiency Issues in Commercial Banking

Allen N. Berger and David B. Humphrey

Commercial banking is a very difficult service industry in which to measure output, technical change, or productivity growth. First, there is disagreement over which services banks produce and over how to measure them. In addition, banking services are often priced implicitly through below-market interest rates on deposit balances, making observed revenue flows inaccurate guides to choosing the important outputs to include in the analysis. Banking also remains a highly regulated industry in which substantial inefficiencies have been shown to exist. As a result, technical improvements that increase the productivity of the most efficient firms may not be well reflected in the industry as a whole. A further complication is that the deposit side of banking underwent substantial deregulation in the 1980s, including the lifting of effective interest rate ceilings on certain deposits and the creation of new types of accounts. The deregulation directly raised banking costs and shifted the optimal mix between the provision of services and the payment of interest to depositors. Measurement of cost changes and productivity gains must take these factors into account, including the possibility of a period of significant disequilibrium as banks attempted to adjust to deposit deregulation.

Despite these difficulties, it is important to analyze the banking industry, as it constitutes almost 20 percent of the U.S. finance, insurance, and real estate service sector of the national income accounts (net of owner-occupied housing). In addition, the externalities that banking generates through its roles as

Allen N. Berger is senior economist at the Board of Governors of the Federal Reserve System, and David B. Humphrey is the Fannie Wilson Smith Eminent Scholar in Banking at Florida State University.

The opinions expressed do not necessarily reflect those of the Board of Governors or its staff. The authors thank Frank Wykoff for his discussion of the paper. They also thank Tim Bresnahan, Dennis Fixler, Zvi Griliches, Diana Hancock, Stacie Humphrey, John Leusner, Jack Triplett, Kim Zieschang, and the other conference participants for helpful comments and suggestions and Alex Wolman for outstanding research assistance.

the nation's primary financial intermediary and conduit for monetary policy are considered to be important enough to require substantial government protection and supervision. Correspondingly, this paper attempts to meet the challenges mentioned above in measuring output and efficiency in U.S. banking in the 1980s. It is hoped that some of the methods used here will also be applicable to the study of other service sectors.

Section 7.1 analyzes some of the problems in defining and measuring bank output. Three methods of choosing which banking functions represent important outputs are evaluated: the asset, user cost, and value-added approaches. The value-added approach, which identifies the major categories of deposits and loans as the important bank outputs, is determined to be the most satisfactory for our purposes. Note that, although deposits are specified as outputs because of their associated service output to depositors, we also specify them as having input characteristics, since they provide much of the supply of funds to be invested in creating loan output.

Sections 7.2 and 7.3 examine inefficiency and technical change in banking. The striking degree of cost dispersion in banking, where some firms have average costs several times higher than others with similar scale and product mix, suggests that the standard assumption of equal efficiency underlying most analyses of technical change is invalid for banking. Cost function studies of technical change (e.g., Hunter and Timme 1986) or productivity measures that relate total industry output to inputs (e.g., the Bureau of Labor Statistics [BLS] labor productivity measure) may confuse changes in the minimum cost technology with changes in the deviations from that technology (i.e., inefficiency). We separate these elements here by estimating the change over time in both a cost frontier and the dispersion of industry costs from the frontier. Three methods of estimating a frontier are analyzed, and the thick-frontier method is chosen as most appropriate for the highly dispersed banking data. This method is applied to all 14,000 U.S. banks for the years 1980, 1984, and 1988, which roughly correspond to periods of pre-, mid-, and post-deregulation of the deposit side of banking, respectively. Shifts in the cost frontier over time are used to examine the effects of deregulation and technical change. Changes over time in the dispersion from the frontier are also evaluated. In this way, the standard approach to measuring productivity change is decomposed into two parts: the frontier shift and the change in dispersion from the frontier. It is found that the interest rate deregulation of the 1980s and the banking industry's response to it on balance increased costs in banking, but much of this increase benefited bank depositors through higher interest payments without a corresponding decrease in the provision of deposit services.

## 7.1 Defining and Measuring Bank Output

There is long-standing disagreement over exactly what it is that banks produce. Three alternative methods of choosing bank outputs are analyzed here,

the asset, user cost, and value-added approaches. It is argued that the value-added approach, which defines outputs as those activities that have substantial value added (i.e., large expenditures on labor and physical capital), is best for accurately estimating changes in bank technology and efficiency over time.

### 7.1.1 The Asset Approach to Defining Bank Output

Virtually all observers would agree that bank liabilities have some characteristics of inputs, because they provide the raw material of investable funds, and that bank assets have some characteristics of outputs as they are ultimate uses of funds that generate the bulk of the direct revenue that banks earn. Under the asset approach, banks are considered only as financial intermediaries between liability holders and those who receive bank funds. Loans and other assets are considered to be bank outputs; deposits and other liabilities are inputs to the intermediation process (see Sealey and Lindley 1977). For some large banks that primarily purchase their funds (with interest payments) from other banks and large depositors and turn these funds into loans, this is an adequate description of bank output. However, most banks do much more than purchase their funds—they also provide substantial services to depositors, but these services are not counted as output in the asset approach.

Mamalakis (1987) makes the useful distinction between the funds intermediation and deposit services of banks, of which the asset approach considers only the former. Intermediation services transform balance-sheet liabilities into assets and pay out and receive interest to cover the time value of the funds used in this capacity. Although some large banks tend to specialize in this function, most banks raise a substantial portion of their funds through produced deposits and provide liquidity, payments, and safekeeping services (as well as interest payments) to depositors to obtain these funds.

For some purposes, the asset approach is the most appropriate. For instance, in a study of loan costs or profitability, a reduced-form model in which the costs and different methods of raising funds are taken to be exogenous may be best. However, any study of banking output as a whole needs to consider a structural form in which the investable funds are an intermediate output of raising deposits, and the services are provided to depositors as partial payment to obtain these funds. The reduced-form asset approach excludes the important differences in service output that occur when the funds are raised via produced deposits versus purchased funds. Moreover, under current institutional arrangements, application of the asset approach to measure banking output often leads to contradictions. For example, consider a bank that produces deposits and sells virtually all its funds to a second bank, which makes commercial loans with these funds. If the two banks merge, there is no change in total banking output, ceteris paribus. However, under the asset approach, if both commercial and interbank loans are considered to be outputs, then measured output would be diminished by the merger because there would be no more interbank lending. If only commercial loans are considered to be out-

puts, then the bank that sells funds has no measured output, despite its production of deposit services and the fact that the second bank values the funds purchased.

### 7.1.2 The User Cost Approach

The user cost approach determines whether a financial product is an input or an output on the basis of its net contribution to bank revenue. If the financial returns on an asset exceed the opportunity cost of funds or if the financial costs of a liability are less than the opportunity cost, then the instrument is considered to be an financial output. Otherwise, it is considered to be a financial input. Hancock (1985a, 1985b) first applied the user cost approach to banking and Fixler and Zieschang (1990; chap. 6, this vol.) used it to determine the weights applied to bank asset and liability categories to derive indexes of bank output and prices.[1]

The user cost approach determines whether an asset or liability category contributes to the financial output of a bank. The operating costs involved in producing nonfinancial services associated with the asset or liability are not explicitly considered. However, under relatively standard assumptions, these operating costs (inclusive of a normal return on capital) are simply the dual of the user cost approach and are included implicitly. An optimizing bank earns (in financial revenue less operating costs) exactly its opportunity cost of funds at the margin on each asset and pays (in financial costs plus operating costs) exactly its opportunity cost at the margin on every liability.[2] Thus, to the extent that the user cost approach accurately measures marginal financial revenues and opportunity costs, its allocation is largely on the basis of excluded operating costs, which is almost the same as the basis of the value-added approach described below. However, there are some difficulties in measuring financial revenues and marginal opportunity costs that make the user cost approach to distinguishing outputs from inputs subject to significant measurement error and sensitive to changes in the data over time.

A problem with measuring the financial flows associated with balance-sheet items, particularly loans and demand deposits, is that there is some commingling of implicit revenues that cannot be easily disentangled. As discussed further below, banks frequently use compensating balances or pay below-market rates on deposits as a method of charging for bank services. Borrowers are often required to hold part of their loan funds as idle demand deposit balances, which means that some of a bank's earnings on a loan are implicit and

1. The user cost approach was pioneered by Donovan (1978) and Barnett (1980) in developing money supply indexes.

2. The underlying assumptions used here are fairly common in banking. If costs and revenues are separable across asset and liability categories and a bank holds securities that it perceives to be in infinite supply (e.g., Treasury Bills), then the quantities of asset and liability categories are adjusted until the marginal revenue less operating cost on every asset and the marginal revenue paid plus operating cost on every liability equal the security rate less its marginal operating cost. See Klein (1971) and Hannan and Berger (1991).

are earned by paying less than the opportunity cost of funds on deposits. Further implicit earnings accrue to the bank on a loan when additional balances are kept with the bank for liquidity, clearing, or timing purposes associated with spending the loan receipts. If the ratio of the compensating and conjunctive balances to loans were known, the implicit earnings could be allocated to loans in much the same way that the implicit losses on deposits from reserve requirements are calculated. However, this ratio is not known or estimated and these implicit revenues are instead allocated entirely to deposits. As a result, there is a bias toward finding loans to be inputs or to have a smaller output weight and toward finding demand deposits (where the balances are held) to be an output or to have a higher output weight.

Another difficulty is in adjusting opportunity costs for the important characteristics of bank assets and liabilities, including differences in credit risk, liquidity, and duration (maturity). Banks earn substantially higher rates for riskier, less liquid, and longer-term assets and pay substantially higher rates for deposits and other liabilities that are uninsured, have fewer liquidity features, and have longer terms to maturity. Theory requires that each dollar of bank liabilities or assets have the same marginal opportunity cost only *after* adjustment for these important characteristics. Therefore, the opportunity cost must be adjusted for each category or, equivalently, the financial return or cost of each category must be adjusted before applying a common opportunity cost.[3] In practice, these adjustments are difficult to make for every category, although there have been some attempts to do so.[4] When such adjustments are not made or fall short, the determination of outputs from inputs and the weights derived for an index of bank output is biased. The bias toward finding an asset to be an output or have a higher output weight is greater, the longer the maturity, the less the liquidity, and the greater the credit risk, because these characteristics increase the unadjusted rate earned on an asset but are not reflected in the opportunity cost as currently measured.[5] Thus, the matching of liability and asset durations to reduce interest rate risk, the holding of assets and liabilities with varying liquidity features, and the making of loans with different credit risks are all commonplace in banking but may not be well reflected in the application of the user cost approach.

A final difficulty is the apparent sensitivity for turning outputs into inputs and vice versa with slight changes in the data or assumptions. When Fixler and Zieschang (1990) switch the assumed opportunity cost for all balance-

3. As well, ex ante rates and opportunity costs are called for, but only ex post values are observed.

4. E.g., Hancock (1985a, 1985b) corrected for credit risks on loans by subtracting off historical average loan losses for each bank. Also, Fixler and Zieschang (chap. 6, this vol.) calculated opportunity costs that differ by bank, which may be viewed as a rough method of accounting for differences in risk, liquidity, and duration across institutions.

5. The biases go in the opposite direction for liabilities, because banks pay higher rates on longer maturity, less liquid, or riskier liabilities, and these higher rates are subtracted from a constant, unadjusted opportunity cost.

sheet items between the average interest rate on loans (the Bureau of Economic Analysis [BEA] method) and the average of interest rates on both loans and deposits (the United Nations Statistical Office [UNSO] method), a number of switches in sign of user cost occur.[6] In addition, nearly half of their financial categories (mostly the smaller categories) switch between inputs and outputs over a five-year period, even without changing the method of computing opportunity cost. One would expect banking technology to remain sufficiently constant that the determination of inputs and outputs should not change so often.

### 7.1.3 The Value-Added Approach

The value-added approach differs from the asset and user cost approaches in that it considers all liability and asset categories to have some output characteristics rather than distinguishing inputs from outputs in a mutually exclusive way. The categories having substantial value added, as judged using an external source of operating cost allocations, are employed as the important outputs. Others are treated as representing mainly either unimportant outputs, intermediate products, or inputs, depending on the specifics of the category. A significant difference from the user cost approach is that the value-added approach explicitly uses operating cost data rather than determining these costs implicitly as that part of the return or cost not accounted for by the difference between measured financial flows and marginal opportunity costs.

The application of the value-added approach here and in other recent cost studies of the banking industry (e.g., Berger, Hanweck, and Humphrey, 1987) identifies the major categories of produced deposits (demand, time and savings) and loans (real estate, commercial, installment) as important outputs, because they are responsible for the great majority of value added. Purchased funds (federal funds purchased, large CDs, foreign deposits, other liabilities for borrowed money) are treated as financial inputs to the intermediation process, because they require very small amounts of physical inputs (labor and capital). On the asset side, government securities and other nonloan investments are considered to be unimportant outputs, because their value added requirements are also very low.[7]

Table 7.1 shows the distribution of expenses for labor (salaries and fringe benefits) and capital (occupancy and furniture and equipment expenses) for the largest size class of banks reported in the Federal Reserve's Functional Cost Analysis (FCA) program for 1980, 1984, and 1988.[8] In 1988, the two

6. Fixler and Zieschang (chap. 6, this vol.) using a distance function approach, estimate opportunity costs that differ noticeably from the BEA and UNSO opportunity costs. However, the output and price indexes formed using the different opportunity costs were not very sensitive to these differences.

7. Government securities also often play an input role when they serve as required collateral on government deposits.

8. The FCA is a cost-allocation/accounting system that assigns direct and joint costs to a number of banking functions based on expert information and accounting rules of thumb. The FCA sample includes about 400–600 banks each year and is inclusive of all bank sizes except the largest.

**Table 7.1** **Distribution of Bank Value Added in 1980, 1984, and 1988 (%)**

| | Deposits | | Loans | | | |
|---|---|---|---|---|---|---|
| Year | Demand Deposits | Time & Savings | Real Estate | Commercial & Industrial | Installment | Total |
| 1980 | 37 | 10 | 3 | 11 | 10 | 71 |
| 1984 | 37 | 14 | 4 | 13 | 12 | 80 |
| 1988 | 36 | 12 | 4 | 14 | 12 | 78 |

*Source:* Board of Governors of the Federal Reserve System, FCA data.

*Note:* Data refer to banks with $200 million to $1 billion in deposits, the largest-size class in the FCA data.

major deposit functions shown absorbed 48 percent of bank value added; three major loan functions absorbed 30 percent, for a total of 78 percent.[9] Similar results are shown for the two earlier periods.[10]

The outputs identified using value added are similar to those used in the BLS measure of bank labor productivity, which uses a set of aggregate transaction flow data on major deposit and loan services, such as the number of checks written for demand deposits, the number of savings deposits and withdrawals for time and savings accounts, and the number of new loans for real estate, commercial, and installment loans (BLS 1989). Unfortunately, these flow data are not available for all banks. In the analysis below, the deflated values of deposit and loan balances are used as outputs for individual banks. The presumption is that these real dollar balances are proportionate to the underlying transactions and account maintenance service flows for the deposit categories and the transactions, credit evaluation, and monitoring service flows for the loan categories.[11] Note that, although real deposit balances are used to indicate bank service output, the interest costs on these deposits, which are associated with the role of deposits as providing the input of loanable funds, are specified as well. In the existing literature, deposits are generally treated as either an input or an output, but both characteristics are represented here.[12]

Despite the differences between the value-added and user cost approaches, the two methods do give similar results, in at least some cases. When we

9. Other bank functions, in declining order of importance, are trust (8 percent), credit cards (5 percent), and other data services (4 percent). The remaining 5 percent includes nonbanking activities (e.g., insurance), nondeposit funds, and safe deposit.

10. If the FCA data were not available, the value-added approach could be essentially replicated for any sample of banks by applying a statistical cost function to call report data. The coefficients of a regression of labor and capital expenses on the dollar volumes of assets and liabilities can substitute for the percentage of value added to determine the important bank outputs and their weights in an output index.

11. In support of this presumption, Humphrey (1992) showed that a cost-share weighted average of the deflated deposit and loan balances used here yields approximately the same growth rate as does the BLS index of bank transactions for the 1980s.

12. Comments by Frank Wykoff and Jack Triplett have helped us clarify our position on this issue, which is essentially an application of Mamalakis (1987).

apply our value-added weights to the same group of banks for the same time period (1984–88) as in Fixler and Zieschang (chap. 6, this vol.), we obtain a 7.6 percent annual growth rate, similar to their rate of about 8.8 percent.

### 7.1.4 Implicit Revenues versus Explicit Revenues in Banking

Much of the disagreement surrounding the choice of bank outputs can be traced to the fact that bank services are not priced in the same manner as services provided by other industries. In many cases, the pricing is implicit for institutional and regulatory reasons.[13] On the loan side, most of the revenue is explicit interest and fees. However, as discussed above, some implicit revenue is raised by business borrowers holding additional idle demand balances with the bank. On the deposit side, revenues from the compensating component of deposit balances, defined as the bank's earnings owing to payment of below-market interest rates, dominate explicit revenues.[14] As shown below, these implicit revenues currently account for over 80 percent of the revenue raised on deposits, although an unknown (but small) part of this figure is implicit revenue for loans.

This suggests that in banking, unlike other industries, explicit revenues are an unreliable guide to determining outputs or service flows. If banks paid market rates on all deposits and charged explicit fees for all deposit services, then this large explicit revenue flow would be convincing evidence that deposits provide substantial service output. Thus, much of the controversy regarding the treatment of deposits as an input or an output arises because the explicit revenues on deposits are relatively small. Another problem with the use of revenue data is that the proportion of revenue that is explicit is not constant. The deregulation of bank deposits caused a significant increase over the 1980s in the proportion of revenues that were explicit, from 5 percent in 1980 to 11 percent in 1984 to 18 percent in 1988.

Table 7.2 illustrates these points, showing estimated breakdowns of deposit revenues and costs for 1988. The implicit revenue from deposit balances is computed as follows:

$$(1) \qquad \text{Implicit revenue} = (1 - r_j/r_{FF})[\text{balance}_j \cdot (1 - RR_j)]\, r_{TB}$$

where $r_j$ = the average interest rate paid on balance $j$; $r_{FF}$ = the federal funds rate (a market rate); $\text{balance}_j$ = the value of the $j$th deposit balance; $RR_j$ = the

13. E.g., idle deposit balances for loan borrowers are negotiated on a case-by-case basis and provide a way to adjust the loan price without altering a very visible and comparable interest rate. For depositors, regulation forbids interest on demand deposits and formerly put ceilings on other deposit rates as well. The use of indirect pricing, such as minimum balance requirements at below market interest rates, also makes comparison shopping difficult.

14. A below-market interest rate on a deposit is equivalent to a zero-rate compensating balance on part of the deposits and a market rate on the remainder. E.g., if the actual rate paid is two-thirds of the market rate, then the implied zero-rate compensating balance is one-third of the total deposit balance. For a demand deposit, which has a zero interest rate, the entire balance is compensating.

**Table 7.2 Deposit Revenues and Costs for All U.S. Banks, 1988**

| Source of Revenues and Costs | Value (in billions of 1988 dollars) | |
|---|---|---|
| Revenues: | | |
| Value of compensating balances (implicit revenues) | | |
| Demand deposits | 26.7 | |
| Time & savings deposits | 14.5 | |
| Other deposits | 0.7 | |
| Total implicit revenues | | 41.9 |
| Explicit revenues from fees on deposits | | 9.4 |
| Total deposit revenue | | 51.3 |
| Allocated operating costs: | | |
| Demand deposits | 20.5 | |
| Time & savings deposits | 21.4 | |
| Other deposits | 5.5 | |
| Total operating costs allocated to deposits | | 47.4 |

*Source:* Revenues are calculated from the call report and costs are allocated from the FCA data. See Berger and Humphrey (1990, appendix table A1A) for more details.

reserve requirement on balance $j$; and $r_{TB}$ = the 90-day Treasury Bill rate, a standard-earnings credit rate applied to compensating balances. The first term in (1) compares the rate paid on deposits ($r_j$) with the market rate ($r_{FF}$) and determines the proportion of balance$_j$ that is purely compensating. This compensating component is then adjusted for nonearning required reserves ($1 - RR_j$) and evaluated using a standard earnings credit rate ($r_{TB}$), giving the implicit revenue flow. The top half of table 7.2 shows the estimated implicit and explicit revenues for deposits for all banks in the United States in 1988. Implicit revenues ($41.9 billion) account for 82 percent of the $51.3 billion in total deposit revenues. About two-thirds of the implicit revenue is generated from demand deposits; one-third is generated by time and savings deposits. The bottom half of the table shows the allocation of operating costs to the deposit categories using ratios of FCA costs for that year. The overall cost estimate of $47.4 billion is just a little below the estimated revenues of $51.3 billion. Note that the slightly higher revenues than costs may be expected because some of the revenues from demand deposits are actually implicit revenues on loans.

Two final conclusions are suggested by these data: First, the finding of a large amount of total (implicit plus explicit) revenue on both demand deposits and time and savings deposits supports the finding under the value-added approach that both types of deposits have output characteristics. Under the asset approach, neither of the deposit types is an output, and under the user cost approach as applied to date, usually only demand deposits are outputs.[15] Sec-

15. An exception is Fixler and Zieschang (chap. 6, this vol.), who do not distinguish among deposit categories, but rather include all deposits and some other purchased funds in a single output category.

ond, if one uses procedures applied to other industries to measure gross output in banking—namely, looking only at the explicit revenue flows—then deposit output is understated by about 80 percent. Moreover, the shifts over time from implicit to explicit pricing can give the false impression that the level of bank output is increasing, even if total revenues (implicit plus explicit) and total output may not have changed.

## 7.2 Inefficiency and Cost Dispersion in Banking

If all banks are approximately equally efficient, as is assumed in most bank cost studies, then it is appropriate to examine technical change and productivity growth over time using the data from either all banks or from a representative sample. However, if banks are not close to being equally efficient, then cost function measures of technical change, such as in Hunter and Timme (1986), or measures of average productivity change, such as the BLS labor productivity index, may confuse shifts in the minimum-cost technology with changes in the dispersion of bank costs away from the minimum-cost technology. We try to separate these elements by forming a thick-frontier cost function for relatively low-cost banks. In this section, we examine inefficiency and cost dispersion away from this frontier; in the following section, we examine changes in this frontier over time.

Banking costs show a striking degree of dispersion. In many cases, banks have costs that are several times higher than other banks with similar scale and product mix. This cost dispersion could be due to many factors, including simple inefficiency. Here, we estimate a thick-frontier cost function using data from banks in the lowest average cost quartile, compare it to a cost function for banks in the highest average cost quartile, and then decompose the difference. The residual that cannot be explained with the available variables is assumed to be a reasonable representation of inefficiency. Some evidence cited below supports this view, specifically (1) high-cost banks experienced much greater failure rates than low-cost banks, (2) the set of banks that were low cost were stable over 1980–88, and (3) low-cost banks consistently had the highest profits.

### 7.2.1 Cost Dispersion in Banking

Figure 7.1 shows for banks in branching and unit banking states the variation in average operating plus interest cost per dollar of assets by bank size class for 1988.[16] $AC_{MIN}$ shows the minimum cost per dollar intermediated; $AC_{Q1}$, $AC_{Q4}$, and $AC_{MEAN}$ are the average costs for the low-cost quartile, high-cost quartile, and overall mean, respectively.[17] The sample was divided into

16. Branching and unit banking states are treated separately here (as in other studies) because of the significantly different regulatory and competitive environments.

17. There were too few large banks in unit states in 1988 to form quartiles for the top two size classes, so only the mean is shown for these classes in figure 7.1B.

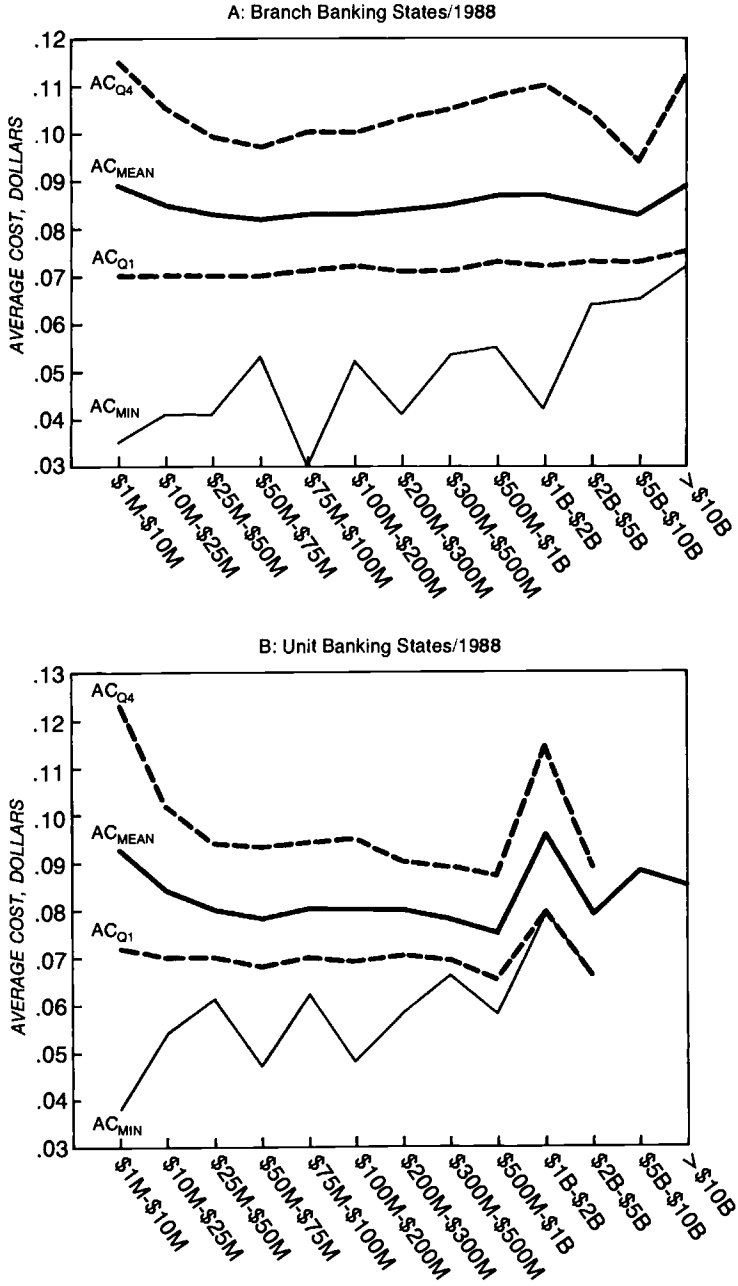


**Fig. 7.1 Average costs by size class and cost quartile**

size classes prior to forming the cost quartiles to ensure reasonable representation of all sizes of banks across quartiles and to limit the problem of dividing up the quartiles on the basis of a function of the dependent variables in the cost regressions below. For both branching and unit states, the data show large variations in average costs between the lowest-cost quartile ($AC_{Q1}$) and the highest-cost quartile ($AC_{Q4}$) for banks within the same size class where size and product mix variations are relatively small. To illustrate how important this dispersion is in light of the highly leveraged nature of banking, consider a typical bank, which has a 6 percent capital/asset ratio, earns 1 cent per dollar of assets, and has a return on equity (ROE) of 16.7 percent. An increase in costs of 3 cents, the typical difference between $AC_{Q1}$ and $AC_{Q4}$, will result in an ROE of $-33.3$ percent and wipe out equity capital in 3 years, all else equal.

The marked cost dispersion also appears to dominate the relatively small-scale and product mix economies in banking. For the 10,961 banks in branching states, the costs for the highest-cost quartile ($Q4$) are 36 percent higher on average than for the lowest ($Q1$); the maximum difference in costs across size classes (taken from the $AC_{MEAN}$ curve) is only 10 percent.[18] The size of these cost differentials strongly suggests that banks are not close to equally efficient, as is assumed in conventional banking studies.

### 7.2.2 Frontier Approaches to Measuring Inefficiency

With the exception of engineering-based analyses, production technologies are essentially unknown. As a result, inefficiencies must be measured relative to some cost or production "frontier" that is estimated from the data. Accordingly, measures of inefficiency reflect deviations of costs or input usage away from some minimal levels found in the data, rather than a true technology-based minimum. The difference among techniques found in the efficiency literature largely reflect differing maintained assumptions involved in estimating the location of the efficient or best-practice frontier.

The major difficulty in estimating a frontier cost or production function is in disentangling inefficiencies from random measurement error and luck. The econometric approach (e.g., Ferrier and Lovell 1990) estimates a frontier cost function where the (composed) error term includes both inefficiency and random error, which are assumed to be orthogonal to the regressors. The two components are separated by assuming that the inefficiencies are drawn from a half-normal distribution and the random errors are drawn from a normal distribution. Unfortunately, the location of the frontier is highly dependent on the actual shapes of the two distributions. As pointed out by Greene (1990) and Stevenson (1980), the half normal is rather inflexible and embodies an assumption that most observations are clustered near full efficiency, with

18. For the 1,844 banks in unit states, $Q4$ costs are 30 percent higher on average than $Q1$; the maximum difference across size classes is 20 percent.

higher degrees of inefficiency being decreasingly likely. This runs counter to the observed bank cost data (fig. 7.1), which suggest a relatively thick-tailed, unskewed distribution of costs.

The data envelopment analysis (DEA) approach (e.g., Aly, Grabowski, Pasurka, and Rangan 1990) avoids distributional assumptions by using linear programming techniques to estimate frontiers that connect the input requirements of the efficient firms. Unfortunately, it does so through the ad hoc assumption that there is no random error—all variation not in the inputs is treated as reflecting inefficiency. If random error does exist, it can have a large cumulative effect on aggregate inefficiency because this measure is determined by comparing the few fully efficient firms on the frontier with all other firms not on the frontier. As indicated in figure 7.1, the lowest-cost observations ($AC_{MIN}$) have costs far below both the mean ($AC_{MEAN}$) and the average of the lowest-cost quartile ($AC_{Q1}$), indicating that a substantial degree of random error may be present.

This paper views the measurement of inefficiencies from a different perspective and uses a set of ad hoc assumptions that are somewhat more intuitive and better justified by our data. Instead of trying to estimate a precise cost or production frontier edge, we estimate a "thick-frontier" cost function for the lowest average cost quartile of banks, where it may be reasonably assumed that banks are of greater than average efficiency. A cost function is also estimated for the highest-cost quartile, in which banks are presumably of less than average efficiency. Differences between these two cost functions are then divided between market factors (e.g., scale, product mix, branches) that are not easily attributable to inefficiency, and a residual, which we assume reasonably represents inefficiency. This inefficiency is then decomposed into several components. In the figures, these differences are roughly represented by the difference between the $AC_{Q1}$ and $AC_{Q4}$ lines. The exact maintained assumptions here are that the error terms within the lowest- and highest-cost quartiles reflect only randomly distributed measurement error and luck and that the differences between the lowest- and highest-cost quartiles reflect only market factors and inefficiencies.

A benefit of the thick-frontier approach is that it requires less specificity in the maintained statistical assumptions, and therefore is less likely to be substantially violated by the data. First, the assumption that the inefficiencies are uncorrelated with the regressors, maintained in the econometric approach, is not needed. Second, our assumption that the error terms for the quartiles satisfy standard regression properties seems no worse than (*a*) the econometric approach assumption that inefficiencies are from an arbitrary (half-normal) distribution, or (*b*) the DEA assumption that random error is zero. Third, even if the error terms within quartiles represent inefficiencies, rather than only random error as maintained, the thick-frontier approach remains a valid comparison of the average inefficiencies of high- and low-cost firms. Finally, as discussed below, the cost quartiles are quite stable over time and are inversely

related to long-term profits, both of which are consistent with the cost differences between quartiles reflecting long-term inefficiencies.[19]

### 7.2.3 Specification of the Thick Frontier

A separate equation is specified for each of three types of costs: physical operating costs, interest costs on produced deposits, and interest costs on purchased funds (which sum to total costs). This permits the use of known, exact prior information on which types of bank outputs affect which types of costs and also allows us to draw separate conclusions about inefficiencies in each of these three cost areas. As discussed in section 7.1, the specified bank outputs are two types of produced deposits—demand and time and savings deposits (DD and TS)—and three types of loans—real estate, commercial and industrial, and installment loans (RE, CI, and IN). Inputs are labor ($L$), physical capital ($K$), and purchased funds (PF).

There is one translog cost equation for each of the three types of cost and an input share equation for operating expenses:

$$\begin{aligned}\ln OC = {} & \alpha^1 + \sum_{i=1}^{5} \beta_i^1 \ln Y_i + \frac{1}{2}\sum_{i=1}^{5}\sum_{j=1}^{5} \beta_{ij}^1 \ln Y_i \ln Y_j + \lambda_B^1 \ln B \\ & + \frac{1}{2}\lambda_{BB}^1 \ln B \ln B + \sum_{i=1}^{5} \tau_{Bi}^1 \ln B \ln Y_i + \sum_{m=1}^{2} \gamma_m^1 \ln w_m \\ & + \frac{1}{2}\sum_{m=1}^{2}\sum_{n=1}^{2} \gamma_{mn}^1 \ln w_m \ln w_n + \sum_{m=1}^{2}\sum_{i=1}^{5} \rho_{mi}^1 \ln w_m \ln Y_i + \varepsilon^1; \end{aligned} \tag{2}$$

$$SOC_1 = \gamma_1^1 + \sum_{n=1}^{2} \gamma_{1n}^1 \ln w_n + \sum_{i=1}^{5} \rho_{1i}^1 \ln Y_i + \varepsilon^2; \tag{3}$$

$$\begin{aligned}\ln ID = {} & \alpha^3 + \sum_{i=1}^{2} \beta_i^3 \ln Y_i + \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} \beta_{ij}^3 \ln Y_i \ln Y_j + \lambda_B^3 \ln B \\ & + \frac{1}{2}\lambda_{BB}^3 \ln B \ln B + \sum_{i=1}^{2} \lambda_{Bi}^3 \ln B \ln Y_i + \varepsilon^3; \end{aligned} \tag{4}$$

$$\ln IPF_{jk} = \alpha_{jk}^4 + \beta_{jk}^4 \ln[(\sum_{i=3}^{5} Y_{ijk}) + OA_{jk}]; \tag{5}$$

where $OC$ = operating costs owing to (1) labor and (2) physical capital and other expenses; $SOC_1$ = share of operating costs paid to input $l$ (labor); $ID$ = interest on deposits (demand and retail time and savings deposits); $IPF$ = interest on purchased funds (federal funds purchased, large CDs, foreign de-

19. For a more extensive discussion of the differences among the frontier approaches, see Berger and Humphrey (1991).

posits, and other liabilities for borrowed money); $Y_i$ = real dollar amount of output $i$, (1) demand deposits, (2) time and savings deposits, (3) real estate loans, (4) commercial and industrial loans, and (5) installment loans; OA = other assets; $B$ = number of banking offices; and $w_m$ = price of input $m$, (1) labor and (2) capital. Coefficients are indicated by $\alpha$, $\beta$, $\lambda$, $\tau$, $\gamma$, and $\rho$; error terms are indicated by $\varepsilon$. The superscripts on the coefficients and error terms signify the equation numbers and the *jk* subscripts on equation (5) refer to size class and quartile.[20] All dollar-value data are in real terms (using the GNP deflator), as are all the cross-year comparisons shown below. Note that the characteristics of deposits as both inputs (interest costs, *ID*) and outputs (real dollar values $Y_1$, $Y_2$, reflecting transactions and account maintenance service flows) are included in the model.

### 7.2.4 Bank Inefficiency Measures and Empirical Results

*Decomposition of differences between the highest- and lowest-cost quartiles.*

The model shown in equations (2)–(5) was estimated by iterative seemingly unrelated regression (ITSUR) for banks in the lowest-cost quartile and for banks in the highest-cost quartile (performed separately for banks in branching and unit banking states). For each size class, the proportionate difference in unit costs between high-cost and low-cost banks to be decomposed is

$$Diff = [A\hat{C}^{Q4} - A\hat{C}^{Q1}]/A\hat{C}^{Q1}, \tag{6}$$

where $A\hat{C}^{Qi} \equiv \hat{C}^{Qi}(X^{Qi})/TA^{Qi}$, $\hat{C}^{Qi}$ is the predicted cost function using the parameter estimates of equations (2)–(5) obtained using the $Qi$ data, $X^{Qi}$ is the vector of mean outputs and other regressors for the size class for the $i$th quartile, and $TA^{Qi}$ is the mean total assets for the size class for the $i$th quartile (size class scripts are suppressed for expositional ease). Thus, *Diff* is the proportional increase in predicted unit costs of $Q4$ data relative to the $Q1$ data, evaluated at the mean of each size class.

Differences in output levels and mix, branch offices, other assets, input prices, and purchased funds levels are not necessarily the result of inefficiencies. These are attributed to exogenous differences in the local markets in which banks operate. Therefore, the part of *Diff* owing to these data differences is referred to as the market component, or

$$Market = [A\hat{C}^{Q4*} - A\hat{C}^{Q1}]/A\hat{C}^{Q1}, \tag{7}$$

20. The purchased funds equation (5) is restricted so that there are no scale or product mix effects within a size class–quartile pairing. This corresponds to a national market for these funds in which every bank has virtually the same opportunities. However, banks in different quartiles may pay different average rates and have different efficiencies because they take different positions in the market with respect to maturity structure or funds type or because they respond differently to changes in market conditions. This restriction improved estimation performance considerably. Additional details of the model are in Berger and Humphrey (1991).

where $A\hat{C}^{Q4*} \equiv \hat{C}^{Q1}(X^{Q4})/TA^{Q4}$. Equation (7) differs from (6) in that the predicted cost for $Q4$ data is evaluated using the efficient technology (estimated from the $Q1$ thick-frontier cost function), rather than the inefficient technology (estimated from $Q4$ data). Embedded in the computation of $\hat{C}^{Q1}(X^{Q4})$ is the assumption that an efficient firm would pay the average interest rate on purchased funds actually paid by $Q1$ firms. Thus, *Market* captures the effects on costs of differences in the levels of the data ($X^{Q4}$ versus $X^{Q1}$), but not in the cost function, because costs are evaluated using only the parameters from the efficient cost function ($\hat{C}^{Q1}$).

The remaining differences in average costs that cannot be attributed to output levels and mix, branch offices, other assets, input prices, and purchased funds levels are assumed to be owing to inefficiencies:

$$(8) \qquad Ineff = [A\hat{C}^{Q4} - A\hat{C}^{Q4*}]/A\hat{C}^{Q1} \equiv Diff - Market.$$

*Ineff* captures only the difference in the estimated cost functions that are taken to represent inefficiency, holding the data constant at $Q4$. Included in *Ineff* are financial inefficiencies in the payment of produced deposit and purchased funds interest, as well as operating inefficiencies in the use of physical labor and capital.[21]

Inefficiencies can be decomposed into several sources by examining the differences in predicted costs attributable to each cost equation separately. For example, the proportion of *Ineff* owing to operating cost inefficiencies is given by

$$(9) \qquad Ineff_{OC} = [A\hat{C}^{Q4}_{OC} - A\hat{C}^{Q4*}_{OC}]/[A\hat{C}^{Q4} - A\hat{C}^{Q4*}],$$

where $A\hat{C}^{Q4}_{OC}$ and $A\hat{C}^{Q4*}_{OC}$ in the numerator indicate the same predicted average costs as in the denominator, except that only operating costs are included. The inefficiencies owing to interest on deposits and purchased funds are computed in similar fashion.[22]

*Market factors and bank inefficiency.* Table 7.3 shows the value of *Diff* and its decomposition for banks in both branching and unit banking states for 1980, 1984, and 1988. The table includes computations for the overall mean of the data and for the mean exclusive of banks in the largest size class (over $10 billion in assets), which are not well matched in size and can be distorting. The differences in predicted costs range from 19 percent to 44 percent for

21. The inefficiency measure also reflects cost differences among banks not specified as market factors—quality differences, left-out variables, and measurement errors. These additional effects are believed to be small. Banking output is quite homogenous across banks within a size class, so quality differentials are negligible. The problem of having a limited number of regional prices for the capital input is of greater concern, but this is mitigated by the fact that capital has only about a 15 percent share in total costs.

22. Data and variable definitions are given in more detail in appendix table 7A.1. The results are virtually unchanged when the large banks (assets over $1 billion) are dropped, but the robustness of the large bank results cannot be verified because there are too few of them per quartile to estimate a separate cost function with confidence.

**Table 7.3** **Decomposition of Costs between Highest- and Lowest-Cost Quartiles 1980–1988 (%)**

| Year | Difference in Predicted Average Costs (*Diff*) (1) | Total Market Factors (*Market*) (2) | Total Inefficiencies (*Ineff*) (3) |
|---|---|---|---|
| | *Branch Banking States* | | |
| 1980: | | | |
| Overall mean | 26.1 | 9.0 | 17.1 |
| Mean < $10 billion | 30.3 | 5.7 | 24.7 |
| 1984: | | | |
| Overall mean | 26.4 | 2.9 | 23.6 |
| Mean < $10 billion | 28.4 | 3.6 | 24.8 |
| 1988: | | | |
| Overall mean | 43.8 | 2.2 | 41.7 |
| Mean < $10 billion | 35.1 | 4.2 | 30.9 |
| | *Unit Banking States* | | |
| 1980: | | | |
| Overall mean | 31.9 | 7.0 | 24.8 |
| Mean < $10 billion | 30.9 | 4.6 | 26.3 |
| 1984: | | | |
| Overall mean | 19.2 | 0.1 | 19.1 |
| Mean < $10 billion | 21.7 | 1.5 | 20.2 |
| 1988: | | | |
| Overall mean | 19.5 | −6.2 | 25.7 |
| Mean < $10 billion | 27.0 | 0.8 | 26.2 |

*Note:* Columns (2) and (3) sum to column (1). See Berger and Humphrey (1990, appendix tables A2A, A2B, and A2C) for size class detail.

all banks over all time periods (col. 1), similar to the raw data in figure 7.1. When this difference is decomposed into market factors (col. 2) and a residual reflecting inefficiencies (col. 3), the inefficiencies clearly dominate. Also, when the results are disaggregated by size class (not shown), the smallest firms show the greatest inefficiencies, consistent with figure 7.1.[23]

*Decomposing inefficiency into operating and financial components.* Table 7.4 shows the decomposition of inefficiencies (*Ineff*) for the same three years. For banks other than those in the largest size classes, operating cost inefficiencies (col. 1) are generally substantially greater than either of the financial (interest cost) inefficiencies (cols. 2 and 3). For the largest banks, purchased funds inefficiencies generally are largest and significantly affect the figures shown for the overall mean. This is because purchased funds are intensively used by large banks (so their weight is higher) and because the rates on these funds are quite volatile and banks differ in their speeds of adjustment to relative rate changes among purchased funds categories.

During this period, there were a number of significant regulatory changes that (*a*) removed interest rate ceilings on savings and small time deposits

**Table 7.4** **Decomposition of Inefficiencies between Highest- and Lowest-Cost Quartiles, 1980–1988 (%)**

| Year | Operating Cost ($Ineff_{OC}$) (1) | Produced Deposit Interest ($Ineff_{ID}$) (2) | Purchased Funds Interest ($Ineff_{PF}$) (3) |
|---|---|---|---|
| | | *Branch Banking States* | |
| 1980: | | | |
| Overall mean | 9.9 | 3.0 | 4.2 |
| Mean < $10 billion | 13.9 | 5.0 | 5.8 |
| 1984: | | | |
| Overall mean | 12.7 | 2.8 | 8.1 |
| Mean < $10 billion | 16.4 | 3.9 | 4.5 |
| 1988: | | | |
| Overall mean | 22.6 | 1.1 | 18.0 |
| Mean < $10 billion | 25.9 | 1.2 | 3.8 |
| | | *Unit Banking States* | |
| 1980: | | | |
| Overall mean | 14.6 | 2.6 | 7.6 |
| Mean < $10 billion | 17.7 | 3.6 | 5.0 |
| 1984: | | | |
| Overall mean | 14.0 | −0.2 | 5.3 |
| Mean < $10 billion | 15.7 | −0.1 | 4.6 |
| 1988: | | | |
| Overall mean | 21.3 | −0.3 | 4.7 |
| Mean < $10 billion | 25.4 | −0.1 | 0.9 |

*Notes:* Percentages add up to total inefficiencies. Columns (1), (2), and (3) sum to column (3) in table 7.3. See Berger and Humphrey (1990, appendix tables A3A, A3B, and A3C) for size class detail.

(starting in 1981 and completed in 1986); and (*b*) permitted banks to offer checkable consumer accounts that paid an uncontrolled interest rate (starting in 1981 and expanded in 1982). From this perspective, 1980 may be viewed as a predereguation period, 1984 as a mid-deregulation period, and 1988 as a postderegulation period by which time the adjustments to deregulation may or may not have been completed.

The usual expectation is that deregulation reduces inefficiency in the long run, but there is some question as to how long that process takes. As seen in table 7.4, operating cost inefficiencies, the main source of inefficiencies and the one expected to be most affected by deregulation, remained approximately constant from 1980 to 1984 and then increased significantly from 1984 to 1988. This pattern was particularly pronounced for larger banks. Thus, operating cost dispersion has increased, and it appears that the adjustment process to the new less regulated equilibrium may not yet be completed. Prior to the 1980s, banks substituted operating expenses (more convenient offices and free deposit services) for their inability to pay market rates on all deposits. After interest ceilings were raised and many zero-interest consumer demand bal-

ances shifted to interest-earning checking accounts, the substitution of operating costs for interest expenses was reversed. The optimal mix between the provision of banking services and the payment of interest to depositors shifted in favor of the latter, but movement to the new equilibrium mix took time because it required closing branches and other capital changes, staff reorganizations, and so on.

Some additional data on the changes in real deposits per branch office tend to support this explanation of the time pattern of inefficiencies. From 1980 to 1984, real deposits per branch office grew 3.8 percent for banks in branching states and then increased by 13.2 percent from 1984 to 1988. This difference is even more pronounced for the larger size classes. Apparently, it took several years for banks to arrange branch closings and mergers to reduce the service/interest ratio toward its new equilibrium and large banks on average moved at a faster pace. One reason for the delay is that many banks likely had new branches in the planning and building pipeline before determining the full effects of deregulation. Consistent with this explanation, the total number of branches nationwide continued to grow, but at a decreasing rate over the 1980s, even as many large banks were closing branches.[24]

*Technical versus allocative inefficiency.* Operating inefficiencies may be further decomposed into their technical and allocative components, which derive from proportionate overuse and incorrect mix, respectively, of the physical labor and capital inputs. Using the methodology of Kopp and Diewert (1982) and Zieschang (1983), the main result (not shown) is that almost all of the operating cost inefficiencies are in the technical category, with less than 10 percent owing to allocative inefficiencies for all three years analyzed.

### 7.2.5 The Relationship between Cost Dispersion and Bank Failure

The importance of cost dispersion or inefficiency in banking ultimately depends on whether banks identified as high cost have difficulty competing. This issue is examined by determining the relationship between costs and bank failures, the premiere measure of competitiveness. Over the nine years from 1981 to 1989, 1,074 banks failed, a substantial increase over previous postwar decades when typically fewer than ten banks per year failed. Without question, the deregulation of deposits in the early 1980s played a part in raising failure rates, raising costs directly through the removal of interest ceilings on deposits and increasing the competition among banks. Although it is not possible to say that high costs by themselves caused any banks to fail, the analy-

23. See our working paper (Berger and Humphrey 1990, appendix tables A2A, A2B, A2C) for this disaggregation.

24. Total banking offices grew 3.2 percent from 1980 to 1981, 1.8 percent from 1984 to 1985, and 1.6 percent from 1987 to 1988. At the same time, many large banks cut branch operations and staff severely. As examples, Bank of America cut branches by 27 percent (about 350 offices) and staff by 34 percent; Manufacturers Hanover reduced staff by 24 percent. Despite these and many similar cuts, a study for the American Bankers Association (Booz-Allen and Hamilton 1987) reported that as of 1986 about half of all branches remained unprofitable and required further cost cutting.

sis below suggests that having relatively high costs is a consistent associated factor.

For each of the three years 1980, 1984, and 1988, all banks have been ranked into four average cost quartiles (as noted above). From these rankings, the cost quartile position for each bank that failed in a subsequent year was determined. The summary results are presented in table 7.5. Of the 768 banks that failed over the nine years, 1981–89, and had been started up by 1980 and had complete call report data for that year, 41 percent were in the highest-cost quartile ($Q4$) during 1980. This is more than three times as many as the 13 percent that were in the lowest-cost quartile ($Q1$). Of the 748 that failed after 1984 but existed and had complete data for 1984, 57 percent were from $Q4$, more than eight times as many as the 7 percent that were in $Q1$. Finally, of the 178 banks that failed in 1989 and had complete data in 1988, 66 percent were ranked as having the highest costs, almost 15 times as many as were ranked in the lowest-cost quartile. As these results indicate, high-cost banks incur an appreciably greater probability of failure, and this probability increases as the time of failure nears.

There are several possible reasons for this positive relationship between high costs and bank failure: First, the high leverage–low spread nature of banking means that a relatively small increase in costs can wipe out earnings and financial capital relatively quickly. Second, high costs tend to be symptomatic of poor management in general. Firms that control costs poorly also tend to have poorly conceived loan policies that contribute to a high failure probability. Third, high costs reduce expected rates of return on equity, ceteris paribus, which may induce a high-cost bank to increase expected return by undertaking more risky activities (i.e., shift to a point further out on its risk-expected return possibilities frontier), increasing the probability of failure. This may be accentuated by the moral hazard aspects of FDIC insurance. When either of the latter two explanations hold, costs contribute to the failure, but the reported reason may be fraud or a high-risk loan portfolio.

### 7.2.6 The Stability and Relationship to Profits of Low-Cost and High-Cost Banks

With the exception of the small effect of the market factors, the differences in costs between the high- and low-cost banks have largely been attributed here to an inefficiency residual. However, it is also possible that these cost differences may reflect short-term differences in luck or omitted variables such as product quality or risk. In this section, we investigate these alternatives by examining the stability of the cost quartiles and their relationship with long-term profits.[25]

25. Another potential explanation of the cost differences between quartiles, differences in monopsony power, may be ruled out as unimportant. The component of the market factors owing to differences in input prices for capital and labor is trivial. Also, although some previous research suggests that banks exercise monopsony power in setting deposit interest rates (see Berger and

**Table 7.5** **Cost Quartile Ranking of Banks That Failed over 1981–1989 (%)**

| Cost Quartile | Failure Percentage by Quartile | | |
|---|---|---|---|
| | 1980 | 1984 | 1988 |
| Q4 (highest cost) | 41.4 | 57.4 | 65.7 |
| Q3 | 24.7 | 24.5 | 21.3 |
| Q2 | 20.6 | 11.2 | 8.4 |
| Q1 (lowest cost) | 13.3 | 7.0 | 4.5 |
| No. of failed banks operating in year of quartile ranking | 768 | 748 | 178 |

**Table 7.6** **Stability and Relation to Profits of Cost Quartiles: Correspondence of Low-Cost and High-Cost Banks for three Single Years with Cost and Profit Quartiles Formed Using Data from 1980, 1984, 1988 Combined**

| | Banks in Cost Q1 for a Single Year (%) | | | | Banks in Cost Q4 for a Single Year (%) | | | | No. of Banks |
|---|---|---|---|---|---|---|---|---|---|
| | Q1 | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4 | |
| *Long-term Cost Quartiles* | | | | | | | | | |
| Branching states | 76 | 20 | 4 | 1 | 0 | 3 | 21 | 76 | 5,403 |
| Unit states | 75 | 20 | 4 | 1 | 0 | 4 | 21 | 74 | 1,625 |
| *Long-term Profit Quartiles* | | | | | | | | | |
| Branching states | 49 | 27 | 15 | 9 | 11 | 18 | 26 | 45 | 5,403 |
| Unit states | 49 | 30 | 14 | 7 | 14 | 14 | 24 | 48 | 1,625 |

The upper half of table 7.6 examines the stability or consistency over time of the lowest- and highest-cost quartiles ($Q1$ and $Q4$, respectively) by showing the correspondence between the quartiles for the three individual years separately and quartiles formed by combining the cost data from all three years together. To avoid the problems of entry, exit, merger, and altered branching laws, we focus on only those banks that (1) existed in all three years and (2) were in states that did not switch between unit and branching status. As shown, 76 percent of banks we rank in $Q1$ using a single year's data also had costs in the lowest quartile for the three years combined. That is, 76 percent of the 1980, 1984, and 1988 $Q1$ banks were also in $Q1$ for the combined data set. Further, 96 percent of $Q1$ banks had costs over time below the median and only 1 percent had costs in the highest quartile using all three years' data. Similarly, 76 percent of the banks we rank in $Q4$ had costs in the highest quartile for the three years combined, 97 percent remained above the median,

Hannan 1989), such an effect on total costs would be small, because differences in predicted deposit interest costs constitute only about 12 percent of the total predicted cost difference.

and less than half of 1 percent were in the lowest quartile for the three years combined. These stability results suggest that the differences in costs between quartiles do in fact represent long-term differences in firm-specific efficiencies, rather than short-term differences in luck or measurement error, because the latter explanations would imply little stability in the quartiles over time. Moreover, our efficiency comparisons between banks in the lowest- and highest-cost quartiles would only in very rare circumstances involve a misordering of efficiencies.[26]

The lower half of table 7.6 examines the relationship between the cost quartiles for individual years and long-term profits by computing profit quartiles that average net income per dollar of assets for the three years combined. The data show that the costs are strongly negatively related to long-term profits—being in the lowest-cost quartile ($Q1$) in a single year makes it from five to seven times as likely that a bank will be in the highest long-term profit quartile ($Q1$) than in the lowest profit quartile ($Q4$). Similarly, being in the highest-cost quartile for a single year makes it from three to four times as likely that a bank will be in the lowest rather than highest profit quartile.[27] These data are consistent with inefficiency being the dominating explanation of the difference in costs between quartiles but are not consistent with the omitted effects of product quality differences or bank risk being dominant. If high-cost banks simply spent more on service quality and were reimbursed on the revenue side, then the cost quartiles would not be highly related to profits. Similarly, if high-cost banks simply chose high risk–high expected return financial strategies and had high costs because they paid high-risk premia for their funds, then costs would be positively, rather than negatively, related to profits on average.[28]

## 7.3 Shifts in the Thick Frontier over Time

Overall unit costs in an industry can change over time because of (1) technical innovation (reflected primarily by shifts in the minimum cost frontier);

26. An alternative method of examining stability, which yielded similar results (not shown), was also employed. It was found that being in either $Q1$ or $Q4$ in 1980 made it three to five times more likely that the bank would again be in that quartile in 1988, rather than migrating to the other extreme. Our use of cost stability over time to identify the presence of inefficiencies from luck is analogous to Gordon's (1965) use of cost stability over product lines to identify managerial inefficiency in airlines.

27. There is also evidence that the strong, negative relationship between costs and profits held in the 1970s as well as in the 1980s. See Kwast and Rose (1983).

28. One caveat to this analysis is that average total costs, the basis of the quartile rankings, is functionally related to the dependent variables in the cost regressions and this could bias the slope coefficients. That is, the quartiles are based on $(OC + ID + IPF)/TA$, while $\ln OC$, $\ln ID$, and $\ln IPF$ are the dependent variables. However, for several reasons discussed in Berger and Humphrey (1991), this does not appear to present a serious problem. The most important reason is that the quartiles were formed separately by 13 size classes, which removes the great majority of the relationship between the cost variables and the quartiles, because the overwhelming determinant of $\ln OC$, $\ln ID$, and $\ln IPF$ is bank size. E.g., the smallest bank in the largest size class (over $10 billion in assets) is more than 1,000 times larger than the largest bank in the smallest size class (less than $10 million in assets).

(2) changes in average efficiency or the variability of market conditions (reflected primarily by changes in the dispersion of costs from the frontier); and (3) the effects of regulatory changes and other disequilibrium phenomena (reflected in both shifts in the frontier and changes in the dispersion from it). The dispersion analysis of the previous section showed (*a*) that banking costs were substantially dispersed from the frontier; (*b*) that this dispersion was largely dominated by efficiency differences; and (*c*) that the dispersion increased considerably between 1984 and 1988. Results (*a*) and (*b*) suggest that efficiency was important and result (*c*) suggests that there may have been some disequilibrium effects from deregulation remaining in 1988.

This section focuses on frontier shifts over time in order to determine the effects of technical change and deregulation as well as the associated disequilibrium. By confining attention to the thick frontier, the effects of technical change and regulation can be determined without the confounding influence of changes in the dispersion of costs from the frontier.

### 7.3.1 The Effect of Deregulation on Deposit Costs and Revenues

The deregulation of bank interest rates and the development of interest-earning consumer checking accounts in the early 1980s had a direct effect of transferring moneys from banks to consumers by increasing deposit interest rates.[29] In effect, banks had some legally enforced monopsony power over retail depositors, primarily on their checking accounts, which was eliminated. The direct effect on interest expenses of banks is independent of bank efficiency, because insured banks in the same market must pay approximately the same (service-adjusted) deposit rates whether they are efficient or not.

Table 7.7 shows how deposit rate deregulation has decreased bank monopsony power over depositors. The first row of the table shows total (implicit and explicit) deposit revenues for 1980, 1984, and 1988, calculated as in equation 1 above. The implicit revenues on deposits declined over time as deposit rates moved closer to market rates, but explicit revenues were not raised sufficiently to cover the implicit revenue reductions. Although deposit revenues were cut almost in half in real terms from 1980 to 1988, deposit operating costs increased, virtually eliminating deposits as a profit center (as estimated profits fell from $61.2 billion to $3.9 billion). Thus, deregulation caused banks to pay out substantially more interest, and this was not offset either by increases in explicit fees or by reductions in operating costs.

### 7.3.2 Shifts in the Thick Frontier and Their Decomposition

The shifts in the thick frontier over time are computed and decomposed in much the same way as the difference between the thick frontier and the highest-cost quartile were computed and decomposed in section 7.2. The

29. The deregulation of deposit rates was largely sparked by two events: (1) the unexpected inflation of the late 1970s and early 1980s, and (2) the growth of competition for bank deposits from money market mutual funds, which have no interest rate restrictions.

**Table 7.7** **The Decline in Bank Deposits as a Profit Center (billions of 1988 dollars)**

| | 1980 | 1984 | 1988 |
|---|---|---|---|
| Implicit & explicit revenues | 98.9 | 65.9 | 51.3 |
| Allocated operating costs | 37.7 | 42.5 | 47.4 |
| Net contribution to profits | 61.2 | 23.4 | 3.9 |

*Note:* All figures in 1988 dollars. See Berger and Humphrey (1990, appendix tables A1A, A1B, A1C) for more details.

unadjusted shift in the frontier from year $t$ to $t + 4$ (analogous to the difference in average costs between quartiles *Diff*) is given by

$$UShift = [A\hat{C}^{t+4} - A\hat{C}^{t}] / A\hat{C}^{t}, \tag{10}$$

where $A\hat{C}^{s} \equiv \hat{C}^{s}(X^{s})/TA^{s}$, $\hat{C}^{s}$ is the predicted cost function using $Q1$ data for year s, $X^{s}$ is the mean argument vector for $Q1$ for year $s$, and $TA^{s}$ is the mean total assets for $Q1$ for year $s$ ($Q1$ scripts are suppressed for convenience in this section). Next, the frontier shift is corrected for exogenous differences in the data that may be owing to changes in market factors (i.e., input prices, output levels and mix, branch offices, and purchased funds levels). Holding these factors in $X$ constant at their time $t$-values gives the adjusted frontier shift:

$$AShift = [A\hat{C}^{t*} - A\hat{C}^{t}] / A\hat{C}^{t}, \tag{11}$$

where $A\hat{C}^{t*} \equiv \hat{C}^{t+4}(X^{t})/TA^{t}$.

A further adjustment must be made for an additional market factor not explicitly included in the cross-section analyses: aggregate interest rates. The large swings in aggregate rates during the 1980s undoubtedly had large effects on bank interest costs. The cost function parameters in $\hat{C}^{t+4}$ in equation (11) implicitly include the aggregate interest rate prevailing at time $t + 4$ and apply it to data vector $X^{t}$. To subtract out this effect, it is assumed that in the absence of technical change and regulation, purchased funds interest rates would move in lockstep with the 90-day Treasury Bill rate, $i_{TB}$. This corresponds with the national nature of the purchased funds (PF) market and the fact that $i_{TB}$ is an appropriate opportunity cost of funds for most participants. For time and savings deposits (TS), rates normally do not move as freely with market rates, because they have a service component and appear to provide depositors with some implicit insurance against swings in market rates.[30] It is assumed that in the absence of technical change and deregulation, the annual average TS interest rate ($i_{TS}$) would move proportionately with $i_{TB}$. Making these two adjustments to *AShift* gives the final shift in the frontier to be decomposed:

$$\begin{aligned} FShift = AShift - \{&[(i_{\mathrm{TB}}^{t+4} - i_{\mathrm{TB}}^{t}) \cdot (PF^{t}/TA^{t}) \\ &+ ((i_{\mathrm{TB}}^{t+4}/i_{\mathrm{TB}}^{t}) \cdot i_{\mathrm{TS}}^{t} - i_{\mathrm{TS}}^{t}) \cdot (TS^{t}/TA^{t})]/AC^{t}\}. \end{aligned} \tag{12}$$

30. See Hannan and Berger (1990) for an examination of the rigidity of deposit interest rates.

The shifts in the thick frontier over 1980–84 and over 1984–88 are shown by bank size class in table 7.8. Columns (1) and (2) show a striking contrast—the unadjusted frontier shifts (*UShift*) indicate increases in costs for all but the very largest size class of efficient banks in the 1980–84 interval, followed by

**Table 7.8** **Shifts in the Thick Cost Frontier: 1980–1984 and 1984–1988 (total percentage change over 4 years, in 1988 dollars)**

| | Unadjusted Frontier Shift *UShift* (%) | | Shift Adjusted for Market Factors and Interest Rates *FShift* (%) | |
|---|---|---|---|---|
| Asset Size Class (millions of dollars) | 1980–84 (1) | 1984–88 (2) | 1980–84 (3) | 1984–88 (4) |
| | *Efficient Banks in Branch Banking States* | | | |
| 0–10 | 32.0 | −14.6 | 33.5 | 6.3 |
| 10–25 | 25.2 | −18.0 | 30.2 | 4.7 |
| 25–50 | 22.3 | −18.7 | 26.7 | 3.1 |
| 50–75 | 19.8 | −18.5 | 22.9 | 2.1 |
| 75–100 | 18.9 | −18.4 | 21.1 | 2.4 |
| 100–200 | 17.4 | −18.0 | 18.4 | 0.7 |
| 200–300 | 13.5 | −19.6 | 15.8 | 1.0 |
| 300–500 | 15.5 | −16.9 | 15.6 | 0.4 |
| 500–1000 | 11.5 | −16.2 | 11.3 | −0.9 |
| 1000–2000 | 9.7 | −19.1 | 11.1 | −3.7 |
| 2000–5000 | 5.1 | −19.3 | 5.6 | −1.1 |
| 5000–10,000 | 4.7 | −18.5 | 2.4 | 4.6 |
| Over 10,000 | −3.8 | −22.6 | 3.3 | 0.6 |
| Overall mean* | 7.3 | −19.6 | 10.0 | 0.7 |
| Mean < 10,000* | 12.2 | −18.4 | 13.0 | 0.8 |
| | *Efficient Banks in Unit Banking States* | | | |
| 0–10 | 33.9 | −15.0 | 32.6 | 4.1 |
| 10–25 | 26.1 | −20.0 | 26.8 | 3.7 |
| 25–50 | 22.5 | −20.0 | 22.6 | 2.1 |
| 50–75 | 18.0 | −21.4 | 19.7 | 0.3 |
| 75–100 | 18.5 | −21.5 | 18.9 | −0.6 |
| 100–200 | 14.0 | −22.2 | 15.1 | −1.7 |
| 200–300 | 12.7 | −22.1 | 11.1 | −3.2 |
| 300–500 | 11.9 | −21.6 | 12.2 | −4.2 |
| 500–1000 | 11.9 | −28.1 | 9.6 | −7.1 |
| 1000–2000 | 3.5 | −20.9 | −5.6 | −9.5 |
| 2000–5000 | −2.8 | −19.6 | 0.3 | −5.9 |
| 5000–10,000 | 16.7 | ‡ | 9.5 | ‡ |
| Over 10,000 | −12.4 | ‡ | −5.1 | ‡ |
| Overall mean* | 12.4 | −21.4 | 12.6 | −1.7 |
| Mean < 10,000* | 15.2 | ‡ | 14.6 | ‡ |

*The mean values reported are asset share weighted sums of the size classes indicated.

‡There were too few large unit banks to form quartiles in 1988 for the top 2 size classes. Thus the mean values reflect the first 11 size classes.

decreases in costs for all sizes from 1984 to 1988. Over 1980–84, small banks experienced more severe cost increases than large banks. This is largely because small banks tend to fund more with produced deposits, for which the average rate paid increased from 4.7 percent to 7.0 percent (because of deregulation); large banks more often use purchased funds, for which the average rate paid decreased from 11.6 percent to 8.7 percent. This suggests that deregulation hurt smaller banks more than large banks over this time period. In contrast, the unadjusted frontier shift over 1984–88 affected banks more equally because interest rates on produced deposits and purchased funds both fell over this time period.

Columns (3) and (4) of table 7.8 show the same frontier shifts after adjusting for market factors and aggregate interest rates (*FShift*). The net effect of these two adjustments is to offset one another for 1980–84, so *FShift* and *UShift* are about the same. In contrast, for 1984–88, the removal of local market factors was swamped by the continuing fall in market rates, so that the fall in costs is essentially eliminated once all adjustments are made.

The shifts in the adjusted frontier (*FShift*) can be further decomposed by cost type into operating costs, interest on produced deposits, and interest on purchased funds, each of which has a different economic interpretation. The change in operating costs over time among the efficient firms is interpreted as representing net technical change, which includes pure technical change, any changes in the level of service produced per dollar of deposits and loans as a result of deregulation, and any disequilibrium effects on the operating costs of the best practice banks. The former two are expected to reduce costs because technical changes generally reduce costs and because banks would be expected to provide less service per dollar of deposits after deregulation (e.g., by increasing the level of real deposits per branch office) in an optimal tradeoff with the higher deposit interest rates. The disequilibrium effects are expected to raise costs as banks incur short-term costs in adjusting to their new long-term equilibrium by shutting branch offices, and so on. The change in interest on deposits net of aggregate interest rate changes is interpreted as the direct effect of the deregulation of deposit interest rates. As mentioned above, these rates went up over 1980–84, even while market rates were decreasing. The change in interest on purchased funds net of aggregate interest rate changes is a residual reflecting maturity, liquidity, or credit risk changes in purchased funds rates not captured by changes in the Treasury Bill rate.[31]

These separate effects are shown in table 7.9 for all sizes of banks. The percentage shift in each type of cost is weighted by its share of total costs, so that the figures sum to *FShift* in table 7.8. Column (1) shows that operating costs per dollar of assets increased as deregulation began in the 1980–84 interval, particularly for banks in the smallest size classes. Column (2) indicates a slightly better performance in the 1984–88 interval, with the smaller banks

31. This does not rule out the possibility that technology can reduce interest costs by improving the monitoring of market conditions, etc. However, such effects are swamped by aggregate interest rate and deregulation effects on deposits and purchased funds.

**Table 7.9** **Net Technical Change and Deregulation Effects: 1980–1984 and 1984–1988 (total percentage change over 4 years, in 1988 dollars)**

| Asset Size Class (millions of dollars) | Operating Cost Net Technical Change (%) 1980–84 (1) | 1984–88 (2) | Interest on Deposits Deregulation Effect (%) 1980–84 (3) | 1984–88 (4) | Interest on Purchased Funds Residual Effect (%) 1980–84 (5) | 1984–88 (6) |
|---|---|---|---|---|---|---|
| | | | *Efficient Banks in Branching States* | | | |
| 0–10 | 8.0 | 2.6 | 24.2 | 3.5 | 1.2 | 0.2 |
| 10–25 | 7.2 | 2.9 | 22.9 | 1.2 | 0.1 | 0.5 |
| 25–50 | 5.2 | 2.0 | 20.5 | 0.7 | 1.1 | 0.4 |
| 50–75 | 3.7 | 1.4 | 18.1 | 0.3 | 1.1 | 0.4 |
| 75–100 | 2.7 | 0.8 | 16.9 | 0.4 | 1.5 | 1.2 |
| 100–200 | 1.9 | 0.1 | 15.1 | 0.1 | 1.4 | 0.4 |
| 200–300 | 1.3 | −0.3 | 13.4 | 0.1 | 1.0 | 1.2 |
| 300–500 | 0.6 | −1.2 | 12.8 | −0.3 | 2.3 | 2.0 |
| 500–1000 | 0.2 | −2.2 | 11.4 | −0.4 | −0.2 | 1.7 |
| 1000–2000 | 0.3 | −4.2 | 10.4 | −0.9 | −0.7 | 1.3 |
| 2000–5000 | −0.2 | −4.3 | 6.9 | −1.0 | −1.1 | 4.2 |
| 5000–10,000 | 2.8 | −3.3 | 8.7 | 0.0 | −9.1 | 7.8 |
| Over 10,000 | 3.6 | −1.9 | 2.8 | −1.3 | −3.1 | 3.7 |
| Overall mean | 2.3 | −1.9 | 9.4 | −0.5 | −1.6 | 3.1 |
| Mean < 10,000 | 1.7 | −1.9 | 12.3 | −0.2 | −1.0 | 2.9 |
| | | | *Efficient Banks in Unit Banking States* | | | |
| 0–10 | 10.5 | 7.4 | 20.6 | −3.4 | 1.5 | 0.1 |
| 10–25 | 7.8 | 5.8 | 18.1 | −2.4 | 0.9 | 0.4 |
| 25–50 | 5.5 | 4.0 | 15.5 | −2.3 | 1.7 | 0.4 |
| 50–75 | 3.7 | 2.3 | 13.8 | −2.1 | 2.1 | 0.2 |
| 75–100 | 3.3 | 1.1 | 12.5 | −1.9 | 3.1 | 0.3 |
| 100–200 | 2.3 | 0.2 | 10.9 | −2.0 | 2.0 | 0.1 |
| 200–300 | 0.4 | −1.4 | 7.9 | −2.1 | 2.8 | 0.3 |
| 300–500 | 0.1 | −3.9 | 7.2 | −1.8 | 4.9 | 1.5 |
| 500–1000 | −0.3 | −4.2 | 6.6 | −1.8 | 3.2 | −1.1 |
| 1000–2000 | −2.6 | −11.3 | 0.2 | 0.1 | −3.2 | 1.6 |
| 2000–5000 | −0.2 | −12.2 | 0.0 | −0.9 | 0.6 | 7.1 |
| 5000–10,000 | −0.9 | * | −0.6 | * | 10.9 | * |
| Over 10,000 | −1.6 | * | −1.1 | * | −2.3 | * |
| Overall mean | 2.0 | −0.9 | 8.5 | −1.9 | 2.1 | 1.0 |
| Mean < 10,000 | 2.4 | * | 9.6 | * | 2.6 | * |

*Note:* Columns (1), (3), and (5) sum to column (3) in table 7.8 and columns (2), (4), and (6) sum column (4) in table 7.8 by construction.

*see ‡ in table 7.8.

showing a lesser cost increase; the larger banks had cost decreases that offset the increases in the earlier interval. Under normal circumstances (i.e., equilibrium growth), these results, particularly those for smaller banks, would be quite unusual, because technical progress is rarely negative for such a long interval of time. However, the 1980s were not normal circumstances and the

results may be interpreted as representing substantial disequilibrium brought about by deregulation. As mentioned above, deregulation altered the balance previously obtained where capital (mainly extra branches) and labor partially compensated depositors for artificially low rates paid on deposits. When rates were deregulated, higher deposit rates were established faster than capital and labor were reduced. When the new equilibrium is reached, average costs are expected to be lower because of pure technical innovation and a lower level of service per dollar of deposits as banks eliminate the extra service that was substituted for legally prohibited interest payments.[32] The evidence cited above on the large increase in real deposits per branch office in the 1984–88 interval suggests that banks were still overbranched and overstaffed and were bearing the transition costs of closing some branches. In addition, to the extent that banks had excess capacity as a result of the temporary disequilibrium, the increase in average operating costs may be overstated as a measure of negative technical change (see Berndt and Fuss 1986).[33]

Columns (3) and (4) of table 7.9 show that deregulation increased deposit interest costs from 1980–84, when most of the initial rate increases occurred, but had little effect from 1984–88. As mentioned above, the interest cost increases from deregulation are much greater for smaller banks, who tend to secure a higher proportion of their funds from produced deposits. Columns (5) and (6) show the residual purchased funds effects, which are small as expected, except for the largest banks which use these funds intensively.

Table 7.10 expresses these effects in annual rates of change for all sizes of $Q1$ banks together and allows for examination of the net effects over the entire eight-year interval. The first two columns simply reexpress the effects in table 7.9 in annual terms; the final column represents a new computation of the frontier shift between 1980 and 1988 using 1980 as a base year. These results suggest that slight technical progress over the last four years offsets the disequilibrium cost increases of the first four years. For banks in branching states, there is a small net annualized reduction in operating costs over the entire eight-year period. For banks in unit banking states, where there was less latitude for cost savings through branch office closings, there was no net measured progress. When this information is combined with the findings given earlier that (1) measured inefficiencies were still relatively high in 1988; (2) smaller banks still had higher operating costs in 1988 than in 1984; and (3) the growth of banking offices was still decelerating in 1988, it suggests that the banking industry had still not reached its new post-deregulation equilibrium by 1988. The effect of deregulation on produced deposit interest costs is

32. An exception to this would be a technical innovation that both increased costs and increased the quality of service. The advent of the automated teller machine (ATM) may fit this category. Undoubtedly, ATMs increased the quality of bank service. There is also some evidence (see Berger 1985) that ATMs may have raised bank costs, especially during the 1980–84 interval when many machines were not at mature volume.

33. Formally, the parameters of the operating cost portion of the cost function $\hat{C}^{t+4}$ in (11) estimated for 1984 may overstate costs owing to excess capacity at that time, yielding an increased operating cost component of *FShift*.

**Table 7.10** **Cost Effects of Technical Change and Deregulation: Annual Growth Rates (at the mean for the low-cost banks) (%)**

| | 1980–84 | 1984–88 | 1980–88 |
|---|---|---|---|
| *Branch banking states:* | | | |
| Net technical change (OC) | 0.6 | −0.5 | −0.5 |
| Deregulation (ID) | 2.3 | −0.1 | 0.6 |
| Residual (IPF) | −0.4 | 0.8 | 0.2 |
| *Unit banking states:* | | | |
| Net technical change (OC) | 0.5 | −0.2* | 0.1* |
| Deregulation (ID) | 2.1 | −0.5* | 0.7* |
| Residual (IPF) | 0.5 | 0.2* | 0.4* |
| *BLS bank labor productivity* | 3.0 | 3.8‡ | 3.3‡ |

*Source:* Computed from table 7.9 plus a similar run comparing 1980 and 1988 data.

*See ‡ in table 7.8.

‡BLS index available only through 1987.

on net an increase, but it virtually all occurred in the first time interval, and its cumulative effect is dying out. Technical change and deregulation seem to have had little or no effect on purchased funds costs, as expected. Examining the final 1980–88 column, it appears that on net, adjusted real banking costs have increased slightly for banks in branching states, as operating cost improvements have not quite offset additional interest payments. For unit states, the lesser possibilities for branch office closings to save on operating costs have left them with a greater increase in real costs over the 1980–88 time interval.

### 7.3.3 Comparison with the BLS Measure of Bank Labor Productivity

It is instructive to compare our results with the BLS measure of bank labor productivity shown in the bottom row of table 7.10. The BLS measure relates bank output as a physical flow (numbers of transactions processed and new loans made) to a single input (labor). It would be expected that the BLS measure would be approximately equal and of opposite sign to our net technical change figures. What is observed, however, is quite different. The BLS finds productivity growth of 3 percent or better per annum for both the 1980–84 and 1984–88 time intervals; we observe a slight increase followed by a slight decrease in costs per unit of output.

One important reason for these seemingly incongruent results is that the BLS measure is based only on labor input; our measure is based on real operating costs, which implicitly include all physical inputs and adjustments in their proportions.[34] When markets are in equilibrium, the change in labor in-

34. Another potential reason for these differing results is that the BLS measure is inclusive of all banks, not just those on the thick cost frontier as used here. Although it is possible that the banks in the three highest-cost quartiles reduced operating costs by more than enough to offset the increased operating costs of the lowest-cost quartile, the data suggest that, if anything, the higher-cost banks fared even worse than the low-cost banks over the 1980s. Thus, this possibility can be discounted as an important explanation of the differing results.

put is a good proxy for the proportional change in all inputs or total operating costs. During the disequilibrium of the 1980s, however, such relationships may not hold. Costs likely increased relative to employment because the expenses of liquidating branches and of installing automated teller machines have often been paid to nonbank capital and labor sources, rather than to employees of the bank (the measured labor input). In addition, to the extent that employment grew more slowly or decreased as a result of branch closings and other pressures from deregulation, any reductions have likely been mostly confined to low-cost employees serving in branch offices, which would decrease measured employment more than proportionate with true value-weighted labor or total operating costs.[35]

An analysis of some raw data tends to confirm these hypotheses. Over the 1980–84 interval, the ratio of bank employment to real operating costs for all banks together fell at a 3.3 percent annual rate. This alone is more than enough to explain the 3.0 percent growth in the BLS productivity index when in fact, productivity in terms of all factors may have fallen. It also can account for nearly the entire difference between the BLS 3.0 percent growth and our 0.5 percent to 0.6 percent increase in operating costs per unit of output. Over the 1984–88 interval, the employment-cost ratio fell at a 2.1 percent annual rate, large enough to explain most of the 3.8 percent measured BLS productivity gains and most of the deviation from our measures. As expected, the fall in labor cost share, 1.0 percent per annum for both time intervals, was smaller than the fall in the employment-cost ratio, consistent with the hypothesis that the proportion of lower-paid employees has decreased over time. This evidence suggests that the BLS measure may be misleading during periods of significant disequilibrium.

## 7.4 Conclusions

Commercial banking is one of the most difficult service industries in which to measure output, technical change, or productivity growth. The problem of choosing which banking functions constitute the important outputs is difficult because many banking revenues are implicit and commingled, so that the flow of explicit revenues is an unreliable guide to the flow of banking services. However, the value-added approach, in which the flows of physical labor and capital inputs are matched to banking functions, identifies the important bank outputs as being the major deposit and loan categories.

It is also difficult to measure technical change and productivity growth because of the confounding effects of changes in inefficiency over time and the deregulation of the deposit side of banking. If inefficiency is not taken into

35. Our use of operating costs implicitly values labor (and capital) at their appropriate marginal value product weights to the extent that different prices paid to different workers accurately reflect their productive values.

account, then measures of technical change or productivity growth may confuse shifts in the minimum-cost technology with changes in the deviations from that technology. In addition, higher deposit interest rates were quickly adopted as a result of deregulation, but the offsetting reductions in depositor services (such as reducing branching convenience) have been relatively slow. These different factors are accounted for here by estimating multiple-equation thick-frontier cost functions for each of three years, 1980, 1984, and 1988, which roughly correspond to pre-, mid-, and postderegulation periods. Cost dispersion and inefficiency are analyzed for each of these years and the shifts between years are decomposed into operating and financial cost categories.

The major findings are as follows: Most of the dispersion in bank costs appears to represent inefficiencies, rather than market factors, such as differences in input prices, scale of operations, or product mix. Except for the very largest banks, the inefficiencies are mainly operational in nature, involving overuse of physical labor and capital inputs, rather than financial, involving excessive interest costs. As well, the set of low-cost banks is seen to remain quite stable over time and to have the highest profits and lowest probabilities of failure during the 1980s, indicating that cost differences are not simply owing to luck and that they are important to bank performance. In addition, operating cost dispersion and inefficiency rose substantially over the period, particularly from 1984 to 1988, suggesting a less than complete adjustment to the new, less regulated equilibrium.

The shift over time in the thick-frontier cost function, after adjustment for changes in market factors and aggregate interest rates, shows important changes in both operating and interest costs resulting from deregulation. First, operating costs for the low-cost banks rose over the 1980–84 interval and then fell over 1984–88. However, the process was uneven, with larger banks able to close and restructure branch operations and otherwise reduce costs; smaller banks continued to have increasing operating costs over 1984–88. Had the progress to the new post-deregulation equilibrium been substantially complete by 1988, one would have expected both technical progress and a shift toward supplying fewer services per dollar of deposits to have resulted in considerable net technical change. However, the overall change is quite small and uneven. Combining these findings with the increase in cost dispersion and the increased real deposits per branch office suggests that progress toward the postderegulation equilibrium remained incomplete by 1988, especially for smaller and less efficient banks.

Second, deregulation removed a substantial source of monopsony power over depositors for banks, raising interest costs significantly and virtually eliminating deposits as an independent profit center. Even by 1988, several years after deregulation, this increase in deposit interest costs generally was not offset by decreases in operating costs, except for relatively large and relatively efficient banks. Given the strong empirical association between high

costs and bank failures, it is likely that this loss of monopsony power contributed to the dramatic increase in bank failures in the 1980s.

Finally, our results contrast sharply with those of the BLS labor productivity index for banking, which shows productivity rising at a 3 percent or more per annum through the 1980s. The major reason for this difference appears to be the use of bank employment as the single factor by BLS.

Overall, deregulation appears to have resulted in little, if any, net technical change or productivity growth in banking in the 1980s. However, offsetting this lack of progress are the benefits of deregulation to consumers, which are not reflected in measured bank output. Consumers obtained a higher return on deposits without a fully offsetting reduction in branch office convenience or higher service fees. Thus, part of the cost increases from deregulation could alternatively be interpreted as increases in output quality, suggesting that the true combined effect of technical progress and deregulation is more favorable than that measured here. In any event, we have identified why measured technical change has been so slow in the 1980s—the reason is banking deregulation and the less-than-cost-minimizing response to it by the banking industry.

# Appendix

**Table 7A.1 Summary of Data (all banks, 1988)**

| | | Branch Banking States (10,961 banks) | | Unit Banking States (1,844 banks) | |
|---|---|---|---|---|---|
| | | Mean | Standard Deviation | Mean | Standard Deviation |
| Cost variables: | | | | | |
| *OC* | Operating costs (% of assets)* | 3.49 | 1.75 | 3.58 | 1.51 |
| $SOC_1$ | Labor share of operating costs (%) | 48.93 | 8.54 | 47.81 | 9.29 |
| *ID* | Interest on produced deposits (% of assets)* | 4.03 | 0.87 | 4.01 | 0.81 |
| *IPF* | Interest on purchased funds (% of assets)* | 0.89 | 0.86 | 0.68 | 0.69 |
| *TC* | Total operating plus interest costs (% of assets)* | 8.41 | 1.76 | 8.27 | 1.44 |
| Output variables: | | | | | |
| *DD* | Demand deposits (% of assets)* | 13.57 | 6.68 | 14.51 | 6.73 |
| *TS* | Retail time & savings deposits (% of assets)* | 63.88 | 13.00 | 65.03 | 11.61 |
| *RE* | Real estate loans (% of assets)* | 24.15 | 12.82 | 21.68 | 12.57 |
| *CI* | Commercial & industrial loans (% of assets)* | 17.58 | 10.38 | 17.79 | 10.60 |
| *IN* | Installment loans (% of assets)* | 11.28 | 9.36 | 10.10 | 7.13 |
| Other variables: | | | | | |
| *B* | Number of banking offices | 5.05 | 20.33 | 1.59 | 1.06 |
| *OA* | Other (nonloan) assets (% of assets)* | 46.98 | 15.12 | 50.43 | 14.09 |
| *TA* | Total assets $000,000, 1988 dollars (not used in regressions) | 223.04 | 20.61 | 104.39 | 23.80 |
| $w_1$ | Price of labor, $000 per year, 1988 dollars | 25.65 | 7.21 | 25.48 | 5.24 |
| $w_2$ | Price of physical capital, 1988 dollars (assumed to be proportionate to the replacement cost of office space in the region, taken from F. W. Dodge) | 81.87 | 9.98 | 72.80 | 1.93 |

*Source:* Reports of condition and income (call reports), except as noted. The flow figures are the annual totals from the December 1988 call; the stock figures are averages from the December 1987, June 1988, and December 1988 calls (to avoid biases from growth or decline over the year).

*Numbers are expressed relative to assets for exposition only. Regressions are based on raw data in $000.

# References

Aly, Hassan Y., Richard Grabowski, Carl Pasurka, and Nanda Rangan. 1990. Technical, Scale, and Allocative Efficiencies in U.S. Banking: An Empirical Investigation. *Review of Economics and Statistics* 72 (May): 211–19.

Barnett, William A. 1980. Economic Monetary Aggregates: An Application of Index Number and Aggregation Theory. *Journal of Econometrics* 14 (September): 11–48.

Berger, Allen N. 1985. The Economics of Electronic Funds Transfers. Outline. Board of Governors of the Federal Reserve System, Washington, D.C., October.

Berger, Allen N., and Timothy H. Hannan. 1989. The Price-Concentration Relationship in Banking. *Review of Economics and Statistics* 71 (May): 291–99.

Berger, Allen N., Gerald A. Hanweck, and David B. Humphrey. 1987. Competitive Viability in Banking: Scale, Scope, and Product Mix Economies. *Journal of Monetary Economics* 20 (December): 501–20.

Berger, Allen N., and David B. Humphrey. 1990. Measurement and Efficiency Issues in Commercial Banking. Finance and Economic Discussion Series, Working paper no. 151. Board of Governors of the Federal Reserve System, Washington, D.C., December.

———. 1991. The Dominance of Inefficiencies over Scale and Product Mix Economies in Banking. *Journal of Monetary Economics* 28 (August): 117–48.

Berndt, Ernst, and Melvyn A. Fuss. 1986. Productivity Measurement with Adjustments for Variations in Capacity Utilization and Other Forms of Temporary Equilibrium. *Journal of Econometrics* 33 (October/November): 7–29.

Board of Governors of the Federal Reserve System. Various years. *Reports of Condition and Income* (call reports), and *Functional Cost Analysis*. National Average Report for Commercial Banks, Washington, D.C.

Booz-Allen and Hamilton. 1987. Managing Delivery System Economics. Bank Branch Profitability Study for the American Bankers Association, October.

Bureau of Labor Statistics. 1989. *Productivity Measures for Selected Industries and Government Services*. U.S. Department of Labor, Bulletin no. 2322, Washington, D.C., February.

Donovan, Donal J. 1978. Modeling the Demand for Liquid Assets: An Application to Canada. *International Monetary Fund Staff Papers* 25 (December): 676–704.

F. W. Dodge Division. 1980–88. *Dodge Construction Potentials Bulletin*. Summary of Construction Contracts for New Addition and Major Alteration Projects. New York: McGraw-Hill.

Ferrier, Gary D., and C. A. Knox Lovell. 1990. Measuring Cost Efficiency in Banking: Econometric and Linear Programming Evidence. *Journal of Econometrics* 46 (October/November): 229–45.

Fixler, Dennis J., and Kimberly D. Zieschang. 1990. Output and Price Measurement in Commercial Banking. Unpublished manuscript. Bureau of Labor Statistics, Washington, D.C., February 7.

Gordon, Robert J. 1965. Airline Costs and Managerial Efficiency. In *Transportation Economics*, 61–94. New York: Columbia Univ. Press.

Greene, William H. 1990. A Gamma Distributed Stochastic Frontier Model. *Journal of Econometrics* 46 (October/November): 141–63.

Hancock, Diana. 1985a. Bank Profitability, Interest Rates, and Monetary Policy. *Journal of Money, Credit, and Banking* 14 (May): 179–92.

———. 1985b. The Financial Firm: Production with Monetary and Nonmonetary Goods. *Journal of Political Economy* 93 (October): 859–80.

Hannan, Timothy H., and Allen N. Berger. 1991. The Rigidity of Prices: Evidence from the Banking Industry. *American Economic Review* 81 (September): 938–45.

Humphrey, David B. 1992. Cost and Technical Change: Effects of Bank Deregulation. *Journal of Productivity Analysis*. Forthcoming.

Hunter, William C., and Stephen G. Timme. 1986. Technical Change, Organizational Form, and the Structure of Bank Production. *Journal of Money, Credit, and Banking* 18 (May): 152–66.

Klein, Michael A. 1971. A Theory of the Banking Firm. *Journal of Money, Credit, and Banking* 3 (May): 261–75.

Kopp, Raymond J., and W. Erwin Diewert. 1982. The Decomposition of Frontier Cost

Function Deviations into Measures of Technical and Allocative Inefficiency. *Journal of Econometrics* 19 (August): 319–31.

Kwast, Myron L., and John T. Rose. 1983. Profitability Differences among Large Commercial Banks During the 1970s. *Magazine of Bank Administration* 59 (September): 54–62.

Mamalakis, Markos J. 1987. The Treatment of Interest and Financial Intermediaries in the National Account: The Old "Bundle" versus the New "Unbundle" Approach. *Review of Income and Wealth* 33 (June): 169–92.

Sealey, Calvin, and James Lindley. 1977. Inputs, Outputs, and a Theory of Production and Cost at Depository Financial Institutions. *Journal of Finance* 32 (September): 1251–66.

Stevenson, Rodney E. 1980. Likelihood Functions for Generalized Stochastic Frontier Estimation. *Journal of Econometrics* 13 (May): 58–66.

Zieschang, Kimberly D. 1983. A Note on the Decomposition of Cost Inefficiency into Technical and Allocative Components. *Journal of Econometrics* 23 (December): 401–5.

## Comment

Frank C. Wykoff

Allen N. Berger and David B. Humphrey (BH) report two empirical regularities in commercial banking activity during the 1980s: (1) the variance in average costs among banks in the United States was large and persistent—average costs of the highest-cost quartile of banks exceeded average costs of the lowest-cost quartile by 30 percent to 50 percent; and (2) these large unit cost differences are not related to bank size, branching, or other observed causal variables but seem to be associated with profitability and failure rates.

BH measure and interpret these unit cost differences using two data bases—(1) functional cost analysis (FCA), consisting of a large but varying nonrandom sample of banks who voluntarily report to the FED on the allocation of costs to different activities, such as deposits and loans, and (2) call reports, in which virtually all American banks, as required by law, report to bank regulators book values of capital, costs of funds, rates of return, and other financial statistics.

The FCA data attributes value added to sources—two categories of deposits, three categories of assets, and other sources. BH find that 48 percent of value added comes from deposit accounts, 30 percent from loans and 22 percent from other activities. Largely on the basis of this evidence, BH define output to consist of the two deposit categories plus the three loan categories.

BH then use call report data to estimate translog cost functions for bank output, with suitable normalizations, from input prices, costs, and levels of outputs for the quartile of banks with the lowest average costs—BH call this quartile, the most efficient banks.

Frank C. Wykoff is Eldon Smith Professor of Economics at Pomona College and the Claremont Graduate School and editor in chief of *Economic Inquiry*.

Figure 7C.1 illustrates their econometric methodology for estimating the cost function. By estimating unit costs as a function of variations in input prices and levels of outputs, they trace out the unit isoquant, $q = 1$, of the average efficient firm by rotating isocost curves like II. If, given input prices, the average efficient bank were a cost minimizer, then it would produce at a point, such as points *a* and *e*, on the isoquant tangent to the isocost. Thus, both the mix of inputs and the level of costs would be optimal.

BH compare the costs incurred by the most costly quartile, the least efficient banks, to the estimated unit isoquant of the efficient banks. The "least efficient" quartile of banks are operating beyond the frontier isoquant, $q = 1$, producing at points like *b* and *c*. The distances from *b* to *a* and from *c* to *e*, along rays from the origin, constitute measures of inefficiency. BH also trace growth of productivity of the average efficient banks over the 1980s and decompose their measure between technical change and efficiency causes. Thus, although much of the paper focuses on efficiency issues, the authors also study productivity growth.

How important are the BH results, how can they be explained, and what do their results have to do with output measurement and productivity growth per se? In my judgment, this paper presents very important empirical evidence

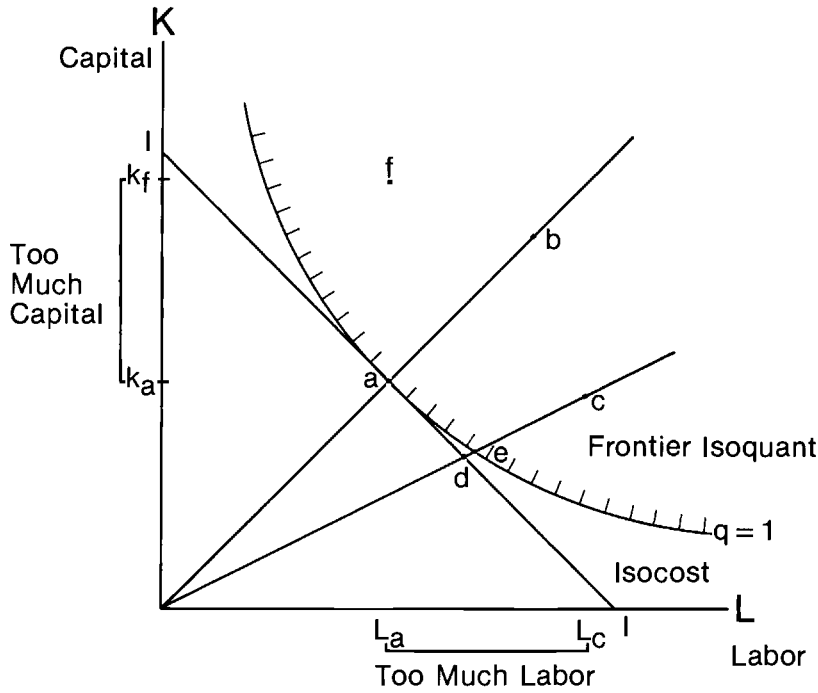


**Fig. 7C.1 Cost differences among banks**

based on comprehensive data on the nation's banking system. The key result, that many banks have been operating in the 1980s at the margin—with excessive costs and low profits, does not auger well for the banking industry should the economy slip into recession. Many major banks could face serious financial stresses, and some could fail. On the brighter side, this same result implies plenty of room for improvement. Many banks could trim fat and tighten their belts. Such a tightening would produce a one-shot jump in productivity growth for the banking system as a whole. Perhaps the BH paper will spur banks and their regulators to move forward with this belt tightening before a recession forces them to do so in a crisis context.

To correct a flawed system, though, it is helpful to know the root causes of the flaws. In this context, we may wonder why do average cost differences, unrelated to observed explanatory variables, persist in the banking system for so long? I have been around Chicago economists long enough to see flares when told of persistent inefficiencies, and so I must ask whether these inefficiencies represent unexploited rent-seeking opportunities?

What do BH mean by the words "efficient" and "inefficient"? BH define efficient firms as those with relatively low average costs and inefficient firms as those with relatively high average costs. This is not necessarily *X* inefficiency nor inefficiency in the broad sense of an economy operating inside its production possibilities frontier.

BH's finding of persistent inefficiencies is a puzzle calling for some explanation. I suggest four possible explanations of their results:

*1. Rising marginal costs.* BH's observation, that unit costs vary among banks, may not be very interesting per se. Supply curves slope upward precisely because each increment in output pulls in marginal resources, including labor and management, that are less well suited to the purpose than previous increments. Persistent cost differences may simply reflect the reality that different qualities of resources are needed to satisfy the entire market—the industry is operating on the upward sloping portion of the marginal cost curve. (This results in producer surpluses for superior firms.)

Were inefficiencies to exist, in the sense that resources outside banking could earn more by leaving their present pursuits and produce banking services, then these resources would, unless restrained by regulation, move in and capture the unexploited rents. But BH do not show that inefficiencies exist in this sense. That is, they do not show that unexploited rent-seeking opportunities exist in banking. They show only that unit costs, and profitability, differ among banks.

*2. Regulatory barriers to entry.* Regulation can protect inefficient firms by creating barriers to entry. An Averch-Johnson regulator, for instance, would cause firms to overcapitalize.[1] Like airlines, who before deregulation flew too

1. See Averch and Johnson (1962) and Diewert (1981) for an econometric model justifying the use of BH-type methods for estimating a variable cost function under Averch-Johnson-type regulation.

many planes on too many routes, BH banks might have too much capital so that their input mix is suboptimal. This would place them at a point like *f* in figure 7C.1—employing too much capital. BH empirically test for and reject the possibility of the wrong input mix, but they might be wrong about this. Bank regulation, however, is probably not Averch-Johnson, taking instead some form of branch and geographical restriction. This could result in many banks operating at high average costs. If this were true, however, then BH have an econometric problem. The assumption of cost minimization needed for their econometric technique to trace out the unit isoquant would be false. Furthermore, BH do not model regulation in order to explain the average cost variance nor to justify estimation of a cost function. As far as I know, the only econometric model developed for regulated firms that justifies BH methods for estimating a variable average cost curve is by Diewert (1981).

*3. Capitalization of land values.* Consider two banks that are the same size, have the same output mix, use the same technology and input mix, yet bank A's costs, except interest costs, are all larger than bank B's. Why? A is a New York bank, and B is a South Dakota bank. Higher land values in N.Y. have become capitalized into all input prices. Tellers cost more; paper products cost more; deliveries cost more. Even lunch costs more. Only money, traded in a world market, costs the same. Banks A and B are both efficient but average costs at bank A, point *b* in figure 7C.1, exceeds average costs at bank B, point a in figure 7C.1, despite the same input mix. As far as I can tell, BH's data are consistent with this story. Because BH do control in their regressions for both rental price and labor cost variations among banks, they may already have largely captured this effect. Furthermore, they also found substantial cost and profit differences among banks in the same large cities.

*4. Different product mix.* Is it possible that bank customers and the services they demand differ across banks? Is it further possible that these service differences cannot be detected in data on financial instruments? Perhaps. Deposits are only representations of the actual underlying flow of services provided by banks to deposit customers.

Consider for example, a $2,000 checking account of two different customers in two different banks. The bank A customer may require more labor and capital services than the bank B customer, and this higher level of services may impose higher unit costs on bank A than on B. The bank A niche may be its appeal to a clientele who differ from bank B clientele. The flow of services that accrue to deposit customers are varied, complex, and subtle—visits to the teller, withdrawals, time saved in a complex variety of check and credit card transactions, access to one's funds at various locations and various periods during the week. Larger or smaller variances of holdings may differ across customers. All those services associated with deposit accounts accrue to different customers, who with apparently identical deposit accounts impose different cost of services on different banks. Security costs, for example, may

well differ significantly within a given metropolitan region. This would imply cost differences but not necessarily differences in deposit accounts.

Loans, similarly, are only representations of the actual underlying flow of services provided by banks. How much analysis must go into assessing oil exploration loans in Texas as opposed to housing loans in Maine or shipping loans in Long Beach? Providing credit to Latin American nations may have different costs than a line of credit to IBM. Pooling risks may be easier in diverse California than in homogeneous Nebraska. Some of these service differences are very difficult, if not impossible, to detect from data on the various pieces of paper produced by banks—financial instruments. BH have only data on financial instruments, assets, and liabilities, not data on the ultimate services that flow to customers from these instruments.

This fourth potential explanation brings me to the core of my comment on the state of knowledge about banking output and productivity growth measurement. Section 7.1 of the BH paper contains a discussion of the appropriate treatment of bank liabilities. Should deposit accounts be treated as output, input, or what? This question is not resolved by BH nor by anyone in the literature. Fixler and Zieschang (1991, and chap. 6, this vol.), for instance, have decided to treat deposits as inputs when net financial flows accrue to depositors and outputs when net financial flows accrue to banks. As noted BH treat deposits as outputs whereas others have treated them as inputs. Some components of liabilities are viewed by BH as inputs, such as federal funds purchased and large CDs. Interest paid on core deposits are treated as a cost. I believe the BH discussion in which they explain differences in approaches, like many others, skims over the fundamental questions, and, unless one focuses the debate on these fundamental questions, disagreements over how to treat deposit accounts and over the key question concerning the output of banking will remain unresolved. It may turn out that BH are absolutely correct in their empirical choices, but in my view we need better explicit conceptual reasons for our choices.

To focus the debate, consider table 7C.1, which shows five possible assumptions about the roles of deposits in the productivity framework, examples of economists who have assumed each role, and questions that must be answered by those making each assumption. Are deposit accounts (liabilities) only inputs that banks use to produce output on the asset side of the balance sheet? Economics has a long history, traceable to both Karl Marx and Adam Smith, doubting the productivity of banks and bankers, so the view that banks do not provide productive services, especially to depositors, is widespread. See, for example, Fixler and Zieschang's (1991) critique of the United Nations system of national accounts. Advocates of the view that banks provide no output to depositors must explain then why people open bank accounts, store money in the banking system, write checks, deposit money, withdraw cash, carry bank cards, check guarantee cards, and so forth. This is

**Table 7C1. What Are Deposits? Five Possible Assumptions with References and Questions for Each**

| Assumption | Reference | Question |
|---|---|---|
| Inputs | Sealey and Lindley 1977 | If deposit customers do not receive outputs from the banks, then why do they spend time and effort to travel to banks to give them these free inputs? |
| Outputs | Berger and Humphrey (chap. 7) | Why are these bank outputs so cheap and why have their nominal prices been comparatively stable, thus falling even in real terms, over the years? |
| Both | Arndt 1984; Triplett (comment on chap. 6 and 7); Wykoff 1991 | How do you measure the price of the outputs and inputs for purposes of partitioning prices and quantities of output growth? |
| Either | Fixler and Zieschang (chap. 6, this vol.); Hancock 1985; Barnett 1980 | Advocates must answer questions under both "outputs" and "inputs," above, because one or the other is assumed for each liability at each moment in time. |
| Neither | Wykoff (this comment) | If deposits are not output, then what is output and what are deposits if they are neither? |

a substantial amount of activity to undertake without compensation. Do they do so in order to voluntarily provide input for bankers without receiving compensating value? If so, how do capitalist bankers force customers to provide them with these inputs without charge? Does nothing but trivial yield accrue to the depositor?

BH and many others treat deposits as output. This position, too, requires its supporters to answer a difficult question, If deposits are outputs only, then why have the explicit nominal prices of these products been so low and so unchanging? Through the volatile period of price level instability from 1965 to 1981 bank deposit fees and charges were flat. Even throughout deregulation under the Decontrol Act of 1981, fees, charges, and rates on deposit accounts have been remarkably low and inflexible. Even though financial institutions are going through difficult times during the 1980s and early 1990s—the Third World loan crisis, bank failures, the oil and real estate collapse, and the savings and loan scandals—explicit charges to customers have not changed very much. If deposits are major outputs, then banks are giving away their products at very low and very stable nominal prices. Rate changes have not even accompanied large fluctuations in the inflation rate! Even the staunchest defender of capitalists do not suggest that they give away their products without compensation. BH, in table 7.7 provide a potential answer to this question by pointing out that profitability of deposit accounts has declined. This implies that banks have been forced by competitiveness to limit charges on deposit accounts.

Arndt (1984), Triplett (comment on chap. 6 and 7), and I elsewhere (1991) have argued that deposits are inputs and outputs simultaneously.[2] Banks receive valuable inputs, cash, that they use as loans. They are willing to pay for their inputs. Depositors, simultaneously, receive outputs from banks that accrue on the accounts. They would willingly pay for these accounts. However, because both the bank and the depositor receive benefits, the gains *largely* offset one another and no flow of payments from either party to the other occurs.

Thus we have a very odd kind of barter transaction involving money in which one type of money is trade for another. The bank receives an input, money, and pays for it with an output, a deposit account. The customer receives an output, the deposit account, and pays for it by providing an input, money.

These barter exchanges are not always of exactly equal value to each party in the trade. When the input value exceeds the output value, then added compensation is demanded by the deposit customer. Thus bank rates on deposits exceed any fees and a tiny *net* cash flow accrues to depositors. When output value exceeds input value, the net cash flow in the form of fees and charges accrues to banks. Advocates of this approach, too, must resolve a tricky issue—how does one measure the gross price of the trade-in-kind exchange? All we have data on is those tiny net flows of explicit fees and explicit interest charges.

Fixler and Zieschang, following Hancock (1985) and Barnett (1980) treat deposits as either inputs or outputs depending on the minuscule net flow of deposit rates and fees. It seems to me that this view is wrong, even nonsensical. They have to answer both sets of questions raised for the input only and output only crowds, When the deposits are outputs, why are they so cheap? When they are inputs, why do people provide them to banks?

My position on this debate requires that I explain what the outputs are and, if deposits are not output, then what are they? These are tough questions that begin to bring us even deeper toward the heart of the matter. In my view, deposits are neither outputs nor inputs. Deposits, in my view, are financial instruments associated with a flow of a wide variety of complex and subtle services received by deposit customers. Deposits are also intermediate goods created in the bank production process partly to provide these services and partly to generate other financial instruments that, in turn, generate final product services.[3]

To determine exactly what these service flows are gets to the very heart of an issue not mentioned so far in discussions of bank productivity growth—

2. Triplett (1990) provides a cogent explanation and several examples of other barter transactions that involve trades in kind.

3. Mamalakis (1987) presents a heuristic discussion of services provided by financial institutions that focus on, among other things, the time dimension of loans. This is a valuable point of departure toward identifying the different social services provided by banks.

namely, what do banks do? As a service-sector firm, a bank must provide services. I do not claim to have fully resolved exactly what banks do nor exactly what the services are, but it does seem clear that financial instruments, per se, are not services. But what the services are, and how one measures the quantities and prices of these services is not fully resolved in the economics literature itself.[4] Nonetheless, I would like to suggest the direction we should look for answers, because I do believe we are closer to the answers than suggested by the present treatment of banking in the productivity literature.

In recent years, the comparatively new *transactions-cost* approach to analyzing markets has challenged the standard neoclassical approach that underlies virtually all productivity growth analysis. The central model of neoclassical theory views competitive market demand and supply schedules as sums of separable decisions of producers and consumers who maximize constrained objective functions. The transactions-cost approach, based on ideas of Ronald Coase in the 1930s and 1940s, focuses on differences among and unique features of various markets that cause these markets to be organized in different ways.

The stock market resembles a Walrasian auction market, but one that must be set up by a stock exchange. The New York Stock Exchange, for example, reaps a return and other benefits for providing the service of creating the market. Sunday flea markets held in drive-in movie theaters resemble Arrow-Debreu barter exchange markets, but here also market organizers receive a return for creating the market. Other markets are organized differently. Steel workers negotiate compensation packages. As discussed by Walter Oi (chap. 4, this vol.), retail store arrangements are very complex. The market for doctors is set somehow in college chemistry departments, and as Coase (1988) points out those who set up shopping malls bring buyers and sellers together and receive significant compensation for this service.

I would argue that the services provided by banks are better understood in the context of a transactions cost model than a neoclassical model.[5] Whereas in the neoclassical model firms and markets exist in which trades occur, the transactions-costs approach argues that markets must be made. One essential function of banks is that they make markets in money. This means they quote a price, absorb imbalances during trading, assure immediacy, insure traders against minor stochastic fluctuations in available supplies and demands, and banks do very much more. They operate the payments system. They transform maturities so as to reconcile the market for loans. They assess risks and label customers as worthy of various levels of credit. They provide investment advice; manage portfolios; provide safekeeping for funds; insure against theft; make trades more convenient, and provide payment services.

The essence of this wide variety of service activities is inherent characteristics of uncertainty, spatial separation, costliness of private information and

4. See Santomero (1984) for a summary of modeling efforts of banks as firms.

5. Goodhart (1989) contains an excellent discussion of the role of banks and banking as viewed in the transactions cost literature as opposed to the Arrow-Debreu approach.

the costs of time. Until and unless we can model exactly what services banks provide and how banks provide these services to facilitate the exchange process in various markets, we are not going to know how to measure the services even if we had unlimited access to data. Until we know what the services are, we cannot tell statistical agencies what data is missing and what to collect. In short, we are, in my judgment, a long way from having viable measures of output in banking.

## References

Arndt, H. W. 1984. Measuring Trade in Financial Services. *Banca Nazionale del Lavorro Quarterly Review* 149:197–213.

Averch, H., and L. Johnson. 1962. Behavior of the Firm under Regulatory Constraint. *American Economic Review* December, 1053–69.

Barnett, William A. 1980. Economic Monetary Aggregates. *Journal of Econometrics* 14:11–48.

Coase, Ronald H. 1988. *The Firm, the Market, and the Law.* Chicago: Univ. of Chicago Press.

Diewert, W. Erwin. 1981. The Theory of Total Factor Productivity Measurement in Regulated Industries. In Thomas G. Cowing and Rodney E. Stevenson. *Productivity Measurement in Regulated Industries*, ed. New York: Academic Press.

Fixler, Dennis J., and Kimberly D. Zieschang. 1991. Measuring the Nominal Value of Financial Services in the National Income Accounts. *Economic Inquiry* 29 (January): 153–68.

Goodhart, C. A. E. 1989. *Money, Information and Uncertainty.* Cambridge, Mass.: MIT Press.

Hancock, Diana. 1985. The Financial Firm: Production with Monetary and Nonmonetary Goods. *Journal of Political Economy* 93, no. 5: 859–80.

Mamalakis, Markos J. 1987. The Treatment of Interest and Financial Intermediaries in the National Accounts: the Old "Bundle" Versus the New "Unbundle" Approach. *Review of Income and Wealth* 33 (June): 169–92.

Santomero, Anthony. 1984. Modeling the Banking Firm: A Survey. *Journal of Money, Credit, and Banking* 16, no. 4, pt. 3:576–602.

Sealey, C. W., and James Lindley. 1977. Inputs, Outputs, and a Theory of Production and Cost at Depository Financial Institutions. *Journal of Finance* 32, no. 4:1251–66.

Wykoff, Frank C. 1991. Commercial Banking Productivity Growth: Evidence from Large Bank Balance Sheets. Working paper, Claremont Graduate School.

## Comment Jack E. Triplett

I have elsewhere remarked that progress in the measurement of banking has been inhibited by two major unresolved questions: (1) What are the outputs?

Jack E. Triplett is chief economist of the Bureau of Economic Analysis, U.S. Department of Commerce.

and (2) What are the inputs? Because these questions correspond exactly to the issues that are displayed in the two papers and discussion in part IIIA of this volume, it may be useful first to summarize approaches to banking output that are found in the literature.[1] Comments on the two papers appear in sections 7C.3 and 7C.4.

## 7C.1 The Traditional National Accounts Approach

The oldest measure of banking output is the one contained in the national accounts of most countries. In national accounts, the banking output measure is determined largely as a consequence of the treatment of interest flows. Production originating in a firm (value added) is defined to include net interest payments (interest paid minus interest received), so that the value added of financial firms' borrowing and lending activities is

$$\text{VA} \equiv \Sigma i_j D_j - \Sigma r_i L_i, \tag{1}$$

where the first term records the firm's deposits (or other financial liabilities) and interest rates paid and the second loans (or other financial assets) and interest rates received. The result is, obviously, normally negative.

Because interest earnings enter negatively into equation (1), the major source of bank revenue (income from lending activity) is excluded definitionally from the measure of banking output. Gorman (1969) colorfully remarks that the national accounts treatment of interest flows—unless adjusted—leaves the "commercial bank . . . portrayed as a leech on the income stream."

To avoid a clearly nonsensical output measure, banks are assumed in national accounts to provide unpriced or free services to depositors (such as check cashing for which no explicit charges are made) that are equal in value to *the entire net proceeds from banks' lending operations*. In some formulations, borrowers are also deemed to receive free services (bookkeeping, credit ratings, and the like). In either case, an imputation for banking output takes the form:

$$\Sigma f_u S_u \equiv -(\Sigma i_j D_j - \Sigma r_i L_i), \tag{2}$$

where $f_u$ and $S_u$ are the implicit fee and (unobserved) quantity of unpriced service $u$, and the other symbols are defined as in equation (1). The total output of the banking industry includes the imputed value of unpriced services, as defined in equation (2), plus the value of services for which an explicit charge is levied (not only certified checks and so forth—a very small part of bank revenue—but also in principle the panoply of financial and fiduciary services that characterize a modern bank). In the United Nations' (but not in the American) implementation, an additional step assures that most of banking output is excluded from GDP and from international transactions.

1. This material is condensed from Triplett (1991).

The national accounts approach to banking was introduced by Yntema (1947); see also United Nations (1968).

Criticisms of the national accounts approach to banking output are quite old. Equation (2) implies that banks act as agents for their depositors (or perhaps for both depositors and borrowers); there is little evidence confirming such a model of bank behavior, or the idea that banks convert their entire earnings into unpriced services. More fundamental is criticism of equation (1) and its exclusion of loan revenue from bank output. Warburton (1958) asserted that a bank's sources of revenue (interest earnings from loans) are as good an indicator of what banks produce and sell as are the revenues of a coal mine or a laundry, and proposed an alternative services approach that would recognize lending activity as the primary bank output. The services approach has been advocated recently by Sunga (1984), Ruggles (1983), and others. The exclusion of banks' provision of finance to borrowers from the national accounts measure of banking output is a serious defect for any analytic purpose.

### 7C.2 The View from the Finance Literature

Another approach that emphasizes bank deposits occurs in the macroeconomic literature of money and banking, and finance. In this literature, the major concern is the bank's role as a portfolio manager, so the banking firm is usually modeled as a seller of deposits (Fama 1980; Pesek 1970; Saving 1977; and Towey 1974)—which is equivalent, of course, to depicting banks as suppliers of money. The traditional money and banking view of banks even has some remote connection to the banking measurement used in national accounts.

Baltensperger (1980) and Niehans and Hewson (1976) point out that the traditional finance-macro approach, because it concentrates on portfolio management, neglects the real side of the economy and also neglects the fact that banks function as distributors of funds. To model banks as distributors of funds, it is necessary to think of them as purchasing funds from depositors and offering interest and bartered depositor services as payment for the use of depositors' funds. The traditional money and banking paradigm—banks selling liquid securities to depositors—is not inappropriate for its own purposes, but it is unenlightening as a paradigm for analyzing bank production and productivity.

### 7C.3 Bank Production Function Approaches

Models of real banking activity and measures of bank output have been developed in the bank regulation literature. To determine whether economies of scale or economies of scope exist in banking, researchers have estimated explicit multioutput production or cost functions, where various bank finan-

cial outputs and inputs and the usual capital, labor, and materials inputs are specified. Hancock (1991) provides comprehensive references to bank production and cost function studies.

Though obtaining a valid measure of output is crucial for modeling bank production and costs, a variety of approaches have been followed, and a consensus on conceptual questions has not yet emerged.

One approach—inexplicably known as the production, or sometimes the value-added, approach (but better termed the activity approach)—takes any bank activity that absorbs real resources as a bank output. Benston, Hanweck and Humphrey (1982) remark, "Output should be measured in terms of what banks do that cause operating expenses to be incurred." In their paper, Allen N. Berger and David B. Humphrey follow a modified activity approach. They define bank outputs "as those activities which have . . . large expenditures on labor and physical capital . . ."; however, they also acknowledge "input characteristics" of deposits and set up their empirical work to incorporate aspects of deposits as both outputs of banking and as banking inputs. U.S. measures of banking labor productivity (Dean and Kunze, chap. 2. this vol.) adopt the activity approach—bank output includes counts of loan and deposit activities (such as loan applications processed and checks cleared).

Critics contend that the cost criterion followed in the activity approach does not adequately serve to distinguish financial inputs from financial outputs. Obtaining any financial input incurs some labor and capital costs (processing certificates of deposit, e.g.). In the empirical work, however, the bank deposits that are usually identified as outputs under the activity approach are precisely the ones (demand deposits) where depositor compensation contains large elements of bartered services; those bartered services are clearly produced by the bank and should be included in any comprehensive measure of bank output.

In a second approach, the researcher distinguishes a priori between those banking activities that are properly considered the outputs of a bank and others that are deemed financial inputs. For example, Mester (1987) assumes, of savings and loan institutions, that "output is best measured by the dollar value of earning assets of the firm, with inputs being labor, capital, and deposits." Three outputs (two types of loans, plus other assets) and three deposit inputs (passbook, NOW accounts, and certificates) were specified. Because only bank assets, and not bank liabilities, are specified as outputs, this approach is usually termed the asset approach to defining bank output (though sometimes it is also referred to as the intermediation approach). Bank deposits are regarded as financial inputs to banks, a necessary source of finance that permits them to sell finance to others.

The asset approach implies that banks buy funds and sell funds, much the same as any other specialized merchant. It is equivalent to the services approach in the national accounts literature (see sec. 7C.1).

A criticism of the assets approach is that its grouping of inputs and outputs

is arbitrary. The choices made by some researchers are disputed by others, and the approach admits no mechanisms for resolving such debates. As it has usually been implemented, the asset approach fails to acknowledge the substantial bank production of services that are bartered to depositors as part of the compensation for the use of their funds, a flaw it shares with the parallel services approach in the national accounts literature.

A third approach resolves the issues empirically. Appealing to Barnett's (1980) notion of the "user cost of money," Hancock (1985, 1991) permits any particular banking activity to be an input or an output according to the sign of its derivative in a bank profit function, which she estimates empirically. In Hancock's findings, loans are bank outputs (which is consistent with both activity and asset approaches—and, of course, inconsistent with the national-accounts approach); time deposits are inputs, but demand deposits are outputs. Fixler and Zieschang follow Hancock's approach in their paper and obtain similar empirical results, including the finding that demand deposits are bank outputs.

A major advantage of the user-cost approach is that it permits statistical tests of the hypotheses maintained in other approaches. Note, however, a potential bias to the empirical results for deposits. Time deposits are typically paid for in strictly monetary terms, so the user cost measure is adequately represented when the nominal cost of deposits is employed in the estimating equation. Demand depositors, on the other hand, receive a large portion of their return in unpriced services. Banks' user costs of demand deposits are accordingly understated when the value of these bartered services is omitted, which biases the estimated sign of demand deposits in the profit function.[2]

The bias can readily be seen in Hancock's (1991, 31–32) expression for the real user cost of a particular deposit type, which (slightly simplified) is

$$(3) \qquad U_i = -1 + (1 + r_i + d_i + Rk_i - s_i)/(1 + R),$$

where the variables are defined as follows: $U_i$ = real user cost per dollar of type $i$ deposits; $r_i$ = interest rate paid to depositors; $d_i$ = deposit insurance rate for the type $i$ deposits; $R$ = discount rate; $k_i$ = reserve requirement for type $i$ deposits; and $s_i$ = actual service charges earned on type $i$ deposits. Equation (3) implicitly takes the bank's acquisition cost for funds to consist only of direct interest payments, $r_i$. For demand deposits, NOW accounts, and similar sources of funds, nominal interest payments account for only a portion of acquisition cost. On conventional checking accounts, for example, $r_i = 0$, and the entire bank acquisition cost is made up of services for which no explicit charge is made. The value of these services, or the cost of producing them, is omitted from equation (3); if the value of free checks and the like were added in to the numerator of equation (3), the effect must obviously

2. I am indebted to Diana Hancock for helpful comments on the analysis in the following paragraphs.

increase the estimated value of $U_i$, which would make it more likely that demand deposits would be classified as financial inputs (for a financial input, $U_i > 0$).

Nominal interest rates are complete measures of compensation for purchased funds and are nearly complete for certificates of deposit and other simple time deposits. In these cases, the estimate of real user cost is positive, because $r_i$ is appropriately measured. These deposits are accordingly classified as financial inputs by Hancock (1985, 1991; and also by Fixler and Zieschang, whose approach is similar).

If banks adjust service schedules and interest rates on the various accounts they offer so as to equalize the cost of funds at the margin, this implies that the user cost of funds from all sources would be equal; this is, of course, a testable hypothesis, but the hypothesis cannot be tested with data that fail to incorporate a major portion (unpriced services) of banks' acquisition cost of certain funds. Adding an imputation for the value of unpriced depositor services to the nominal cost of demand deposits would correct the bias, and, one expects, move the estimates in the direction of making demand deposits financial inputs to the bank.[3]

The omission of unpriced depositor services from the bank deposit user cost measure could also account, in part, for the puzzling sign reversals in Fixler and Zieschang's findings for deposits. The greater is the proportion of direct interest in total depositor compensation, the more likely are demand deposits to emerge as bank financial inputs. Presumably, deregulation of deposits increased the proportion of explicit payments in total depositor compensation.

## 7C.4 Conclusions and Research Directions

In the three literatures on measuring banking activity summarized in sections 7C.1–7C.3, the fundamental difficulty arises in the treatment of demand deposits. The underlying cause of the difficulty is the fact that banks compensate depositors at least in part with bartered services, and data on prices and quantities of those bartered services are not available.

When deposits are treated as bank output (activity and user-cost approaches, in part), the logic *must* be that a count of the volume of deposits serves as a proxy for unpriced services produced by the bank and provided to depositors as compensation for the use of their funds. But by thus obtaining an imperfect proxy for the unobserved portion of bank output, the researcher understates a major part of the bank's cost of funds (though not necessarily

3. The omission of unpriced services from Barnett's (1980) formulation of user costs was noted by Offenbacher (1980, 55), who wrote: "Barnett follows the vast majority of money demand studies by assuming that it is useful to treat regulated own rates of return [to deposit holders] as the true rates. . . . It may be more useful to assume that [interest rates ceilings on bank deposits] are almost totally ineffective . . . [and] banks completely evade the ceilings and pay a competitive rate of return on deposits." Evasion of interest rate ceilings (then set at zero for demand deposits) took the form of provision of varying quantities of depositor services.

understating total bank costs) and distorts cost of funds comparisons between banks that use purchased funds, compared with those that obtain funds from traditional deposits.

When deposits are treated solely as financial inputs, on the other hand (the asset approach), the substantial part of bank output made up of unpriced services produced by the bank is omitted. The cost of financial inputs is likewise understated by the portion of depositor compensation that takes the form of unpriced services. The same problem arises with respect to the services approach in the national accounts literature: it would correct the conceptual incongruity in the national accounts definition of bank output (its omission of loan activity from the output measure) at the cost of excluding unpriced services that are imputed (if inadequately) in the present measure.

The national accounts measure of banking contains, of course, an estimate of the value of unpriced depositor services, but not a defensible one. The national accounts estimate of depositor services is clearly too large, because it, in effect, assigns the loan rate as the opportunity cost forgone by depositors.[4]

All approaches to banking thus suffer from the absence of data on bartered banking transactions. No approach satisfactorily deals with demand deposits in the absence of such data, and no approach gets around the basic data deficiency.

Once the barter nature of banks' transactions with depositors is recognized, then it becomes clear that one must separate conceptually depositor services (the bank output) from the deposits themselves, which function as purchased financial inputs to the bank. The value of free checks, automatic-teller-machine usage, and so forth must be added to banks' output. Simultaneously, the same values must be added to the cost of banks' purchased financial inputs. From the depositor's perspective, the value of unpriced services is simultaneously income and outlay on banking services.

Obtaining values for unpriced depositor services is a formidable problem. It seems natural to view depositor compensation as consisting of a bundle of interest and unpriced services, much as labor compensation is made up of direct wages plus benefits. One method, applicable in regulated and unregulated environments alike, is to assume that the full value of the bundle is equal for all types of accounts—that banks equalize at the margin the cost of funds

4. Fixler and Zieschang (1991) maintain that, under certain circumstances, the user-cost approach they follow can rationalize the idea that banks pay out their earnings to depositors, and this seems to offer support for the traditional national accounts treatment of banking. Their demonstration is indeed helpful in assessing the plausibility of the agency model of bank behavior (that is, the assumption embodied in eq. [2], sec. 7C.1). However, the essential part of the national accounts approach to banking is its treatment of loans as negative contributions to bank output, in equation (1). This treatment of loans is shared by no other approach to banking (including that of Fixler and Zieschang, this vol.). The negative contribution of loans to equation (1) gives rise to the corresponding necessity for inserting a negative sign before the bracketed quantity in equation (2). That negative sign in equation (2)—and not the sensible mathematics that eliminates it—is essential to the logic of the national accounts approach to the output of financial firms.

from different sources (this implies that deposits are indeed financial inputs to the bank), or that depositors value equally at the margin a dollar's worth of interest and a dollar's worth of unpriced services.[5] This assumption implies that

$$\Sigma f_{uj} S_{uj} = (i_k - i_j)D_j, \tag{4}$$

where $i_k$ is the interest rate paid on some account with minimal services (a certificate of deposit, perhaps, or purchased funds), $i_j$ is the explicit interest (if any) paid on the $j$th type of account, and $(\Sigma f_{uj} S_{uj})$ designates the quantity of unpriced services earned on the $j$th account.

If alternative mixes of interest and services are observed on various accounts, which is true under deregulation, a hedonic function (Griliches 1971) might be used to estimate the unpriced components of depositor compensation (Triplett 1991). This approach is a generalization of equation (4). It requires both schedules of direct interest payments and of uncharged services (the quantity vector $S_{uj}$ in equation (4), which would be used in combination to estimate the implicit price vector $f_{uj}$). Data for implementing a hedonic approach have yet to be assembled, but it is in principle little more difficult than any other hedonic investigation.

Beyond this, the heterogeneity of bank loans has not been addressed satisfactorily in empirical estimates. Irrespective of their approach to banking output, banking production function studies frequently consider whether bank output activity is best specified by the count of the numbers of loans (or deposits) of different types, or by their respective monetary volumes. The issue arises, of course, because loans are not a homogeneous commodity: They differ in size and also in other characteristics (riskiness, e.g., or compensating balance requirements). Compensating balance requirements imply that the nominal quantity of loans overstates, and the nominal interest rate understates, the true magnitudes of the loan transaction. Moreover, because banks have extended their financial activities beyond the traditional deposit-taking and lending roles, banking output measures must incorporate these nontraditional activities; some of them (brokerage, selling insurance, executing hedging arrangements) are areas where defining or measuring the output of the activity, or its price, pose conceptual problems comparable in difficulty to the ones confronted in traditional banking.

A perhaps more fundamental question also remains. When banks sell finance (or rent loanable funds) to borrowers, what is the nature of the services that finance provides? The ultimate test for the empirical validity of a measure of bank output is to find some effect on, say, the production process and pro-

5. Presumably it is after tax returns that are equated by depositors. Interest income is taxable; implicit unpriced services income is not. In Triplett (1991) I argued that the relevant bank marginal cost might differ from the direct cost of producing services if the method of depositor compensation affects the costs of reserves. Neither of these complications needs to be considered here.

ductivity of business borrowers, for whom banking output is an intermediate input.

## References

Baltensperger, Ernst. 1980. Alternative Approaches to the Theory of the Banking Firm. *Journal of Monetary Economics* 6:1–37.

Barnett, William A. 1980. Economic Monetary Aggregates: An Application of Index Number and Aggregation Theory. *Journal of Econometrics* 14 (September): 11–59 (Annals of Applied Econometrics 1980–83. A Supplement to the *Journal of Econometrics*).

Benston, George J., Gerald A. Hanweck, and David B. Humphrey. 1982. Scale Economies in Banking: A Restructuring and Reassessment. *Journal of Money, Credit, and Banking* 14, pt. 1 (November): 435–50.

Fama, Eugene F. 1980. Banking in the Theory of Finance. *Journal of Monetary Economics* 6:39–57.

Fixler, Dennis J., and Kimberly D. Zieschang. 1991. Measuring the Nominal Value of Financial Services in the National Income Accounts. *Economic Inquiry* 29 (January): 53–68.

Gorman, John A. 1969. Alternative Measures of the Real Output and Productivity of Commercial Banks." In *Production and Productivity in the Service Industries,* ed., Victor R. Fuchs, 155–89. NBER Studies in Income and Wealth, vol. 34. Irvington-on-Hudson, N.Y.: Columbia Univ. Press.

Griliches, Zvi, ed. 1971. *Price Indexes and Quality Change: Studies in New Methods of Measurement.* Cambridge, Mass.: Harvard Univ. Press.

Hancock, Diana. 1985. The Financial Firm: Production with Monetary and Nonmonetary Goods. *Journal of Political Economy* 93:859–80.

———. 1991. *A Theory of Production for the Financial Firm.* Boston: Kluwer Academic.

Mester, Loretta J. 1987. A Multiproduct Cost Study of Savings and Loans. *The Journal of Finance* 42 (June): 423–45.

Niehans, Jurg, and John Hewson. 1976. The Eurodollar Market and Monetary Theory. *Journal of Money, Credit, and Banking* 8 (February): 1–27.

Offenbacher, Edward K. 1980. Economic Monetary Aggregates—Comment. *Journal of Econometrics* 14 (September): 11–59 (Annals of Applied Econometrics 1980–83: A Supplement to the *Journal of Econometrics*).

Pesek, Boris P. 1970. Bank's Supply Function and the Equilibrium Quantity of Money. *Canadian Journal of Economics* 3 (August): 357–85.

Ruggles, Richard. 1983. The United States National Income Accounts, 1947–1977: Their Conceptual Basis and Evolution. In *The U.S. National Income and Product Accounts: Selected Topics,* ed., Murray F. Foss, 15–96. NBER Studies in Income and Wealth, vol. 47. Chicago: Univ. of Chicago Press.

Saving, Thomas R. 1977. A Theory of the Money Supply with Competitive Banking. *Journal of Monetary Economics* 3:289–303.

Sunga, Preetom S. 1984. An Alternative to the Current Treatment of Interest as Transfer in the United Nations and Canadian Systems of National Accounts. *Review of Income and Wealth* 30:385–402.

Towey, Richard E. 1974. Money Creation and the Theory of the Banking Firm. *The Journal of Finance* 29 (March): 57–72.

Triplett, Jack E. 1991. Measuring the Output of Banks: What Do Banks Do? BEA discussion paper no. 53. Washington, D.C.: Department of Commerce.
United Nations. 1968. *A System of National Accounts.* Studies in Methods, series F, no. 2. New York: United Nations.
Warburton, Clark. 1958. Financial Intermediaries. In *A Critique of the United States Income and Product Accounts,* 509–16. NBER Studies in Income and Wealth, vol. 22. Princeton, N.J.: Princeton Univ. Press.
Yntema, Dwight B. 1947. National Income Originating in Financial Intermediaries. In NBER Studies in Income and Wealth, vol. 10. New York: NBER.

## Comment Diana Hancock

There are several issues that must be resolved before the existing literature on output aggregation can be applied to banking. Of primary importance is a methodology for classifying and measuring financial services. Although it is agreed that banking firms produce heterogeneous services, there has been little consensus on the measurement of their outputs and inputs. The outputs used by various researchers include total assets, earning assets, loans, total deposits, produced deposits, demand deposits in dollar terms, the number of deposit and loan accounts, gross operating income, and combinations of these measures.

The central questions in what can be termed "the classification problem" are (1) Which balance sheet items produce services that are *net* outputs, and which ones are *net* inputs? In particular, are demand deposit services net outputs, or are these services intermediate inputs? and (2) How does one measure the outputs and inputs, or put prices on them? The measurement of price is dual to the question, What units is output measured in? One can be obtained from the other if the necessary conditions for producer equilibrium are satisfied. Another way of posing the problem is whether stock or flow variables measure the relevant concept of bank output and input.

Even with appropriate prices and quantities for financial services determined, the following topics need to be addressed before exact aggregate output indexes for banking can be constructed. First, tests for whether the necessary separability restrictions hold to construct each output subaggregate need to be performed.[1] Second, if all prices move proportionately, then a Hicksian aggregation scheme can be used to aggregate over the firm's joint output supplies. This proportionality assumption is unlikely to hold for financial service prices, due in part to regulation, and hence aggregation over out-

Diana Hancock is an economist in the Division of Monetary Affairs, Board of Governors of the Federal Reserve System, Washington, D.C.

1. A subaggregate refers to an index containing fewer than all the prices or quantities used in production.

puts is possible only if outputs are separable from inputs in the financial firm's technology.[2] Third, the functional form for the aggregator function, and whether it is linear homogeneous, determines the appropriate nonparametric approximation, or index number, which corresponds to the economic output quantity index in banking.

The papers by Dennis J. Fixler and Kimberly D. Zieschang and Allen N. Berger and David B. Humphrey both use nonparametric approaches to classify financial services as inputs or outputs. Fixler and Zieschang employ the user-cost approach to determine whether a balance sheet item is a net output of a bank. In contrast, Berger and Humphrey use a value-added approach.

The user-cost approach tackles the classification problem by deriving complete rental prices for each balance-sheet item. The user cost of a financial service, or price, is the net effective cost per dollar of holding the asset or liability on the balance sheet over period $t$. These prices depend on the opportunity cost of capital as well as interest rates, capital gains, reserve requirements, and insurance premiums. In continuous time, the user cost for each asset is the difference between the bank's opportunity cost of capital and its holding revenue rate. If the holding revenue is not sufficient to cover the opportunity cost of capital, then the balance-sheet item contributes to the financial institution's costs, and the financial product is a net input. If, however, holding revenues are greater than the opportunity cost of capital, then the firm's production of this service contributes to revenue, and the service is a financial output. The user cost for each liability incorporates the implicit revenue from deposit balances and takes into account reserve requirements.[3] If holding costs are greater than the opportunity cost of capital, then holding the liability on the balance sheet contributes to costs, and the liability is classified as an input.

Estimation of the opportunity cost of capital is important because it influences the prices, the classification of inputs and outputs, and hence the revenues and cost earned from the production of financial services. The paper by Fixler and Zieschang obtains an estimator for the opportunity cost of capital that comes from the specification of the technology producing intermediation

2. See William Barnett, The Microeconomic Theory of Monetary Aggregation, in *New Approaches to Monetary Economics*, ed. William Barnett and Kenneth Singleton (Cambridge: Cambridge Univ. Press, 1987), 124, for a discussion of this problem in the context of money aggregation.

3. Berger and Humphrey calculate the implicit revenues from deposit balance $j$ as implicit revenue $= (1 - r_j/r_{ff})\,[DB_j\,(1 - k_j)]r_{TB}$; where $r_j$ is the average interest rate paid on deposit $j$, $r_{ff}$ is a market rate such as the federal funds rate, $DB_j$ is the dollar balance of deposit $j$, $k_j$ is the reserve requirement rate, and $r_{TB}$ is the 90-day Treasury Bill rate. Rearranging terms the implicit revenue per dollar of deposit balance $j$ is $(r_{TB} - k_j r_{TB} - r_j(r_{TB}/r_{ff}) + k_j r_j(r_{TB}/r_{ff}))$. These implicit revenues are included in the user cost calculation with the assumption that the appropriate opportunity cost of capital is both the 90-day Treasury Bill rate and a proxy for the market rate. See Diana Hancock, The Financial Firm: Production with Monetary and Nonmonetary Goods, *Journal of Political Economy* 93(1985): 859–80 for a derivation of user cost formulas for asset and liability items for banks.

services, and the assumption of profit maximizing behavior. The authors assume that the value of the opportunity cost of capital for each bank is equal to some proportion of its return on assets, and this proportion is the same for all banks. By adopting this approach, when the industry representative opportunity cost of capital is estimated, a distribution of bank-specific opportunity costs is also obtained because each bank has a different return on assets.[4] By developing a theoretically appropriate opportunity cost of capital for banking, the authors have helped to answer the question of how inputs and outputs for financial services are classified using the user-cost approach.[5]

Fixler and Zieschang estimate a linear homogeneous conditional distance function to obtain an estimator for the opportunity cost of capital. A Malmquist economic output index is the ratio of this conditional distance function for two time periods. The assumption of the linear homogeneity is important because otherwise the output quantity aggregate depends on the choice of the reference quantity used to condition the distance function.[6] The translog specification for the distance function used in their paper can produce a second-order approximation to any distance function. An appropriate nonparametric approximation to the Malmquist quantity index is the Törnqvist Divisia index.[7] This index is chained, and measures changes relative to the previous period rather than a base period. It remains suitable even when the technology is changing over time, and the aggregator function is shifting. This feature is crucial in the measurement of banking output in the 1980s. Berger and Humphrey state, "The shift over time in the thick frontier cost function, after adjustment for changes in market factors and aggregate interest rates, shows important changes in both operating and interest costs resulting from deregulation." Hence, a useful measure of aggregate bank output needs to be flexible enough to allow financial institutions to respond to their external environment and technological changes over time. An extension of this approach is to test whether separability conditions for output subaggregates hold. It may be possible to construct aggregates which only use financial service data.

The value-added approach assumes that the firm's technology can be written,

$$Q_t = g(\mathbf{x}_{1_t}, \mathbf{x}_{2_t}), \tag{1}$$

4. The different returns on assets reflect differences in risk, liquidity, and duration across institutions that affect their opportunity cost of capital.

5. Market proxies, such as the 90-day Treasury Bill rate, are provided for the opportunity cost of capital.

6. Fixler and Zieschang use the level of deposits as the reference quantity in their estimation of the technology for banking firms. Quality variables are incorporated in the conditional distance function, too.

7. W. E. Diewert, Exact and Superlative Index Numbers, *Journal of Econometrics* 4(1976): 115–45, has shown that the discrete Divisia index is exact for the Malmquist quantity index even if the distance function is a nonhomogeneous translog if the reference level is chosen appropriately. Fixler and Zieschang calculate a Törnqvist index of real bank output.

where $Q_t$ is an exact output aggregate for period $t$, and the firm's input vector has been partitioned such that $\mathbf{x}_{1_t}$ is the vector of quantities of primary input (such as labor and capital) and $\mathbf{x}_{2_t}$ is a vector of intermediate inputs. The factor price vector is partitioned in a corresponding manner with $\mathbf{w}_t = (\mathbf{w}_{1_t}, \mathbf{w}_{2_t})$. The firm's variable profit function conditional on $\mathbf{x}_{1_t}$ is $\pi_t = \pi(\mathbf{x}_{1_t}, \mathbf{w}_{2_t}, \mathbf{p}_t)$, where $\mathbf{p}_t$ is the output price vector. The true index of real value added is

$$\pi_{t_0 t_1} = \frac{\pi(\mathbf{x}_{1_{t_0}}, \mathbf{w}_2^*, \mathbf{p}^*)}{\pi(\mathbf{x}_{1_{t_1}}, \mathbf{w}_2^*, \mathbf{p}^*)}, \tag{2}$$

which depends on the reference prices $\mathbf{w}_2^*, \mathbf{p}^*$.

The need to select the reference prices become unnecessary if and only if $g$ is separable so that

$$Q_t = G(\psi(\mathbf{x}_{1_t}), \mathbf{x}_{2_t}). \tag{3}$$

In this case $\pi_t = \pi_1(\mathbf{x}_{1_t})\pi_2(\mathbf{w}_2, \mathbf{p})$ and $\pi_{t_0 t_1} = \pi_1(\mathbf{x}_{1_{t_0}})/\pi_1(\mathbf{x}_{1_{t_1}})$. If $\psi$ has a translog functional form, then a discrete Divisia index is exact for $\pi_{t_0 t_1}$. In continuous time, the Divisia index is always exact for $\psi(\mathbf{x}_{1_t})$, which is value added.

Berger and Humphrey extend the value-added approach to classify financial services as inputs or outputs. The primary input costs, salaries and fringe benefits, occupancy, furniture and equipment expenses are allocated ex ante to specific balance-sheet items such as real estate loans, and demand deposits using an external source of operating cost allocations.[8] Outputs are defined as those services that are responsible for the largest amount of operating costs.

This approach assumes that the firm's technology can be written

$$Q_t = h(\mathbf{x}_{11_t}, \ldots, \mathbf{x}_{1n_t}, \mathbf{x}_{2_t}), \tag{4}$$

where the primary input vector $\mathbf{x}_{1_t}$ has been partitioned into $n$ separate banking functions. Value added is calculated for each financial service, and the index of real value added is

$$Q_t = H[\gamma(\mathbf{x}_{11_t}), \ldots, \gamma(\mathbf{x}_{1n_t}), \mathbf{x}_{2_t}]. \tag{5}$$

8. Berger and Humphrey use functional cost analysis data. This is a cost accounting system, developed by the Federal Reserve, that assigns direct and joint costs to specific banking functions, such as demand deposits. This system is based on expert information, participant surveys, and accounting rules of thumb.

Underlying this representation of the technology is the assumption that the transformation function is nonjoint in inputs, or that there exist individual subproduction functions for each financial service.[9]

Tests need to be performed using banking data to determine whether the necessary separability conditions hold to construct either a general true index of value added, or an index of product specific value added for banking. The latter index would require additional testing on the structure of the technology for the banking firm. The resulting economic quantity aggregates can be approximated using index number theory.

Berger and Humphrey investigate technical change, or shifts in the production technology for banking output over time. Shifts in the minimum-cost technology are distinguished from changes in the dispersion of bank costs away from the minimum technology. The dispersion is decomposed into inefficiency components and market factor components. They find that, if inefficiency is not taken account of, then measures of technical change may be biased in periods of disequilibrium.[10] This result is important because it indicates that measurement of technical progress requires estimation of the firm's technology. The rate of technological change may not be able to be measured exactly in banking using input and output indexes.

In conclusion, the literature on aggregation and index number theory can be used to construct economic measures of banking output. Examination of the production technology is essential to test whether the necessary separability conditions hold, if jointness in production is statistically important, and to study technical change and productivity. This examination may also help determine whether deposit services are intermediate inputs or outputs.

9. Berger and Humphrey do not impose this structure on their estimating system once outputs and inputs have been classified using the value-added approach. Z. T. Adar, T. Agmon, and Y. E. Orgler, Output Mix and Jointness in Production in the Banking Firm, *Journal of Money, Credit, and Banking* 7(1975): 235–43, argue that interdependence may arise from the joint use of certain inputs by many banking products. Jointness in production is evident in the joint use of information by different departments. An example is the use of depositor information when evaluating a loan application. Jointness in production, also called economies of scope, has been found to be statistically significant in some but not all studies of financial service production.

10. Berger and Humphrey argue that much of the disequilibrium in the 1980s was caused by deregulation, and the less than cost-minimizing response to it by banks.