Productivity in Banking and Effects from Deregulation*

David B. Humphrey

I. INTRODUCTION

There has been a marked decrease in the rate of productivity growth in the United States and other countries since the early 1970s. The likely reasons for this slowdown have been surveyed recently in Cullison (1989). The slowdown shows up in measures of single factor (labor) productivity as well as in the more comprehensive multifactor measure, which includes the productive effects of labor and capital together. For example, productivity in the U.S. nonfarm business sector only rose at a 0.22 percent annual average rate over 1973-87. But for the 25 years prior to 1973, productivity growth was over seven times larger (at 1.68 percent a year). The slowdown was even more striking for some U.S. service sectors. In particular, the Finance, Insurance, and Real Estate (FIRE) service sector experienced an average labor productivity growth rate that was negative, at −0.41 percent a year over 1973-87. In the 25 years before 1973, however, this growth averaged 1.41 a year (Baily and Gordon, 1988, pp. 355, 395).

Banking makes up 20 percent of the FIRE service sector (net of owner-occupied housing) and thus contributes importantly to this sector's behavior. The purpose of this paper is to provide estimates of total factor productivity for the banking service sector over the past decade (1977-87) and to investigate the cause of the low productivity growth found. Productivity results are reported from two growth accounting models: one based on a production function and another based on a cost function. Both approaches indicate a similarly low rate of productivity advance for the banking industry, ranging between −0.07 (production approach) to 0.6 percent (cost approach) a year.

It is argued that low productivity growth in banking is largely due to the effects of bank deregulation initiated in the early 1980s. Deregulation permitted the establishment of new interest-bearing consumer checking accounts and eliminated ceilings on time and savings deposit interest rates. Deregulation during the 1980s, preceded by the intensive use of cash management techniques by corporations in the 1970s, effectively removed banks' virtual monopoly control over zero-interest checking accounts and low-interest small consumer time and savings deposits. Core deposit interest costs rose but were not offset by either reduced costs elsewhere or with an expansion in measured bank output. Apparently, market share considerations limited the desire by banks to reduce operating costs enough to fully offset the rise in interest expenses.

While banks may have experienced very low (to negative) productivity growth, users of banking services have benefited. But the benefits, which are similar to an increase in the "quality" of banking output, are not captured in any measure of banking output. Thus, although measured bank productivity growth is low or negative, it would be inappropriate to conclude that society as a whole has not benefited. Rather, there has been a redistribution of productivity benefits in which users of banking services have gained at the expense of banks.

II. PRODUCTIVITY IS "OUTPUT PER UNIT OF INPUT," BUT WHAT IS BANK OUTPUT AND WHAT ARE THE INPUTS?

What Do Banks Produce?

In many industries, physical measures of output and inputs are readily available and, importantly, a consensus also exists on how best to measure them. In the electric power industry, for example, the obvious measure of output is kilowatt-hours of electricity produced. Inputs used to produce electric power include the number of workers, the real value

* The opinions expressed are those of the author alone. Comments by William Cullison, Tony Kuprianov, and David Mengle are appreciated. Alex Wolman contributed outstanding research assistance.

of electric generators and transmission facilities, and the tons of fuel inputs used. In contrast, in the banking sector physical measures of output are not readily available (although they exist for some banks); indeed no strong consensus exists regarding what it is that banks produce. As a result, measures of banking productivity can use different definitions of outputs and inputs.

Banks produce a variety of payment, safekeeping, intermediation, and accounting services for deposit and loan customers (Benston and Smith, 1976; Mamalakis, 1987). Some have argued, however, that banks primarily produce loans. With this (asset) approach, the production of deposit services is viewed as merely payment in kind for the use of funds from which to make loans (Sealey and Lindley, 1977). In effect, this is a "reduced form" model of the banking firm: the production of deposit services is treated as an intermediate output to depositors who provide loanable funds, so deposit services are netted out.

But there is no reason to focus on only a single banking output such as loans, especially because the production of deposit services accounts for half of all physical capital and labor input expenditures. Because deposit services are such a large component of bank value added, explicit modeling of their productive structure, along with that of loans, will yield a more accurate description of this structure for the bank as a whole. This objective can be achieved using a structural model of a multiproduct banking firm. In such a model, the production of deposit services would not be netted out; instead, it would be one of a set of bank outputs.

For purposes of analysis, banks are considered to produce payment and safekeeping outputs (associated with demand deposits and savings and small denomination time deposits) as well as intermediation and loan outputs (associated with real estate loans, consumer installment and credit card loans, and commercial, industrial, and agricultural loans). Over the last decade, these five deposit and loan output categories accounted for 75 to 80 percent of value added in banking (Berger and Humphrey, forthcoming, see table). Such a categorization of bank output, with one exception (time deposits), is consistent with that identified in the user cost approach to determine bank inputs from outputs (Hancock, 1986; Fixler and Zieschang, forthcoming).

Measures of Bank Output

Based on data availability, there are at least three different measures of banking output that could be used in productivity analyses: (1) the number of transactions processed in deposit and loan accounts (a flow measure); (2) the real or constant dollar value of funds in the deposit and loan accounts (a stock measure); or (3) the numbers of deposit and loan accounts serviced by banks (a stock measure).[1] Because output is typically a flow, not a stock, the preferred measure is seemingly an output flow. Stock measures would only be used if a flow measure were unavailable or because the stock measure might be proportional (on average) to a flow measure.

A time-series transactions flow measure of aggregate banking output is compiled by the Bureau of Labor Statistics (BLS, 1989). However, this measure exists only for the aggregate of all banks and has a limited number of observations. Thus for most purposes, researchers have been forced to rely on stock measures of bank output and to assume that there is a proportionality between stocks and flows, so use of stocks succeeds in approximating flows. Because one possible stock measure—number of deposit and loan accounts—is essentially unavailable for time-series analysis,[2] researchers have relied on the stock

[1] A fourth measure, concerning bank debits and deposit turnover (published monthly in the *Federal Reserve Bulletin*), should not be used. These data are in value terms and include both check and wire transfer debits. As a result, the virtually exponential growth in the value of wire transfers will grossly dominate this series, even though wire transfer expenses are a minute portion of total bank costs. While it is possible to remove the value of wire transfer debits, the end result would be a measure of the value of check and ACH debits, which is inferior to the quantity measure of aggregate check and ACH transactions captured in the transaction flow measure discussed immediately below.

[2] See the Appendix for more detail on data availability.

Summary of Bank Total Factor Productivity Estimates
(annual average growth rates; 1977-87)

	QT	QD
Growth Accounting Method:		
Production Function	−0.00%	−0.07%
Cost Function	0.60	0.50
Econometric Estimation Method:[1]		
Cost Function:		
Hunter & Timme (1991)	—	1.05
Humphrey (1991)	—	−1.01

[1] Both of these studies used multiproduct indicators of bank output rather than the single aggregate index QD. Transactions flow data (QT) are not available to be used in pooled times-series, cross-section econometric analyses.

of the real value of deposits and loans. These data are available over time and for each bank in the United States. As a result, cross-section information can be pooled over time, allowing the estimation of more sophisticated econometric models than is possible with any of the other measures of bank output. It is assumed, but has never been tested, that the transaction flow of bank output over time is proportional to the stock of real deposit and loan balances (Box 1).[3] That these two alternative measures of bank output have had a somewhat similar variation over the last decade is documented below. While this does not strongly support the assumption of strict proportionality between bank output flow and stock, it does suggest that somewhat similar estimates of productivity may be obtained using either output measure for this period. This point is demonstrated below.

[3] The same assumption is made in cross-section studies in banking where scale economies are the focus of modeling and estimation.

Box 1

When Will Stock and Flow Measures of Bank Output Be Proportional to Each Other?

Stock and flow measures of banking output will be proportional to one another when only the two following influences determine the growth in nominal deposit and loan balances over time. First, nominal deposit and loan balances grow because of population growth. An expanding population leads to a larger demand for bank transaction services as more deposit accounts are opened, more checks are written, and more savings deposits and withdrawals occur. Thus, over time, increased transaction flows will be associated with larger stocks of deposit balances. Population growth and economic expansion also leads to loan growth. The nominal value of the stock of bank loans will rise as new loan transactions occur and expand at a greater rate than outstanding loans are retired. The second influence is inflation, which raises the average size of loans made and the average idle deposit balances held by users of bank services. If only these two influences determine the variation in nominal deposit and loan balances, then deflation by some appropriate price index will give the real value of deposit and loan balances and also reflect the underlying flow of bank transactions.

Inputs Needed to Produce Output

There is less controversy on measuring bank inputs. Labor (number of workers or total hours worked) and the real or constant dollar value of physical capital (usually the book value of premises, furniture, and equipment deflated by some price index) clearly represent inputs needed to produce bank output.[4] However, there is less agreement about also treating the real or constant dollar value of loanable funds—core deposits plus purchased funds—as an input.

If labor and capital were the only inputs, then measured productivity would refer to bank operating costs. Since operating costs are less than one-third of total banking costs, however, an operating cost productivity measure by itself would not indicate the degree to which productivity improvements may affect user costs or bank profits. More importantly, since capital and labor operating expenses which support a branch network are substitutes for the interest costs of purchased funds (federal funds, CDs, Eurodollars, etc.), operating expenses are not a stable proportion of total costs either over time or (especially) across different-sized banks.[5] This instability can bias productivity estimates derived solely from operating expenses, just as it has been shown to bias the determination of bank scale economies (Humphrey, 1990). Hence the appropriate cost concept from which to estimate bank productivity is total costs, which includes operating plus interest expenses. From this it follows that the five appropriate inputs are labor, capital, demand deposits, small time and savings deposits, and purchased funds. Thus a total factor measure of productivity is preferred.

Unlike other industries, total costs for an aggregate bank cannot be determined by simply summing all costs at all banks. Some costs, such as the cost of funds purchased from other banks in the interbank

[4] Researchers familiar with the many problems associated with measuring real capital stock will find the measurement method employed in this paper to be overly simple and potentially misleading. Fortunately, these capital measurement problems will have only a relatively small effect on the banking productivity results because the share of capital expenditures in total cost is itself small, around 15 percent.

[5] Purchased funds permit a bank to grow faster and attain a larger size than if it relied solely on a base of branch-generated deposits.

funds market (e.g., federal funds), are costs only to individual banks but need to be excluded when aggregate data are used. This exclusion is necessary because if there were only one aggregate bank, which is the implicit assumption in using aggregate data in the type of models specified, interbank costs would not exist and total costs need to be reduced by this amount. The cost of funds purchased outside of the U.S. banking system, such as virtually all large CDs, Eurodollars, and other liabilities for borrowed money, however, would remain.

To sum up, both input (cost) and output (service flow or stock) characteristics of core deposits are specified (following Wykoff, 1991), rather than only one or the other as is usually done in the literature. In contrast, purchased funds have only input characteristics. Overall, five categories of bank output and five areas of input costs are specified.

III. Growth Accounting Estimates Of Banking Productivity

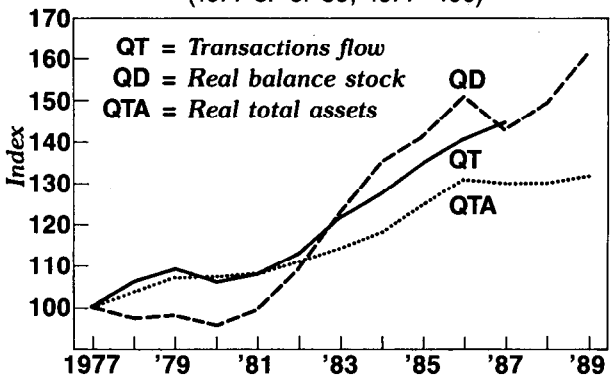
There are essentially two ways to measure bank productivity. The growth accounting approach (Box 2) uses raw data on input and output growth rates plus information on input cost shares while an econometric approach specifies a cost or production function relating outputs to inputs and estimates this relationship statistically. While the focus in this paper is on the growth accounting approach, results of existing econometric studies of bank technical change and productivity are also noted.

The data necessary to determine banking productivity from growth accounting models based first on a production function and second on a cost function (both shown in Box 2) are different with the exception of the measure of bank output. In what follows, the time-series variation of two bank output measures are compared, after which productivity results based on these output measures in both production and cost-growth accounting models are then contrasted.

Transactions Flow and Real Balance Stock Measures of Bank Output

The transaction measure of bank output used here is the BLS index of deposit and loan transactions (QT). In contrast, the stock measure is an index of the real value of deposit and loan account balances

Figure 1
A Comparison of Flow and Stock Measures of Banking Output
(1977-87 or 89; 1977=100)



(QD).[6] Both are shown in Figure 1. For comparison purposes, the real value of total bank assets (QTA) is also shown.[7] Over 1977-87, the annual average rate of growth of QT was 3.8 percent while that for QD was almost identical at 3.7 percent. But the average figures can be misleading since QD was very flat in the early 1980s but grew more rapidly than QT at the middle of the decade. Thus the assumed proportionality between bank transactions flows (QT) and the stock of real balances (QD) is only approximate over this period even though the R^2 between QT and QD is relatively high (.82). In comparison, QTA grew by only 2.7 percent on an annual average basis and, if used as a measure of banking output here (as some have argued), would understate the expansion of bank output compared with the other two measures.[8] Such understatement holds even though the R^2 between QT and QTA is higher (.97) than that between QT and QD.

A Production-Based Measure of Banking Productivity

The Bureau of Labor Statistics computes annually an aggregate measure of labor productivity in

---

[6] The construction of both of these indexes are described in the Appendix. The BLS data are available only through 1987 (BLS, 1989).

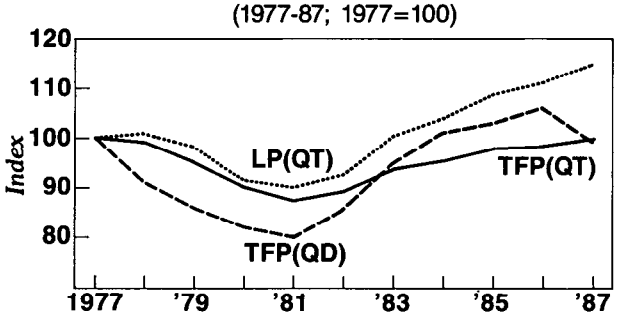
[7] Real total assets were obtained by deflating the nominal value of total banking assets by the GNP deflator.

[8] Since interbank sales of funds (e.g., federal funds sold) have grown over time and show up in total assets, the aggregate value of these assets will be overstated by this amount compared to a situation where there is only one aggregate bank and interbank sales no longer appear on the balance sheet. Thus the understatement possible when using total assets as an indicator of aggregate bank output is even greater than that shown in the figure since these total asset values have not been corrected for this double counting.

Figure 2

Production Approach: Single-Factor (Labor) and Total Factor Productivity

(1977-87; 1977=100)



TFP(QT) = *Total factor productivity with transactions flow output*

TFP(QD) = *Total factor productivity with real balance stock output*

LP(QT) = *Labor productivity with transactions flow output*

banking using transactions (QT) as its measure of output. This series, LP(QT), is shown in Figure 2. Cyclical behavior of labor productivity is due to cycles in bank output transactions flows, specifically cycles in new loans being made as deposit transaction growth was always positive.[9]

Over the 1977-87 period, the average annual increase in numbers of workers was 2.4 percent[10] while banking output (QT) rose by an average 3.8 percent. Because output grew faster than the labor input, labor productivity is positive (at 1.4 percent a year). But labor productivity is not representative of overall banking productivity if other inputs grew more rapidly or slowly than labor.[11]

Our (rough) estimate of the growth of the real value of bank physical capital is 1.8 percent annually with the real value of demand deposits falling by 3.5 percent, time and savings deposits growing by 5.9 percent, and purchased funds growing by 3.1 percent.[12] The net result is that the cumulative level of total factor productivity (TFP), using the QT transactions flow output measure, is below that for labor productivity. A similar result occurs when TFP is derived using the QD real balance stock output measure. Overall, neither measure of total factor productivity in a production-based growth accounting model shows any growth[13] while the BLS labor productivity measure grows by 1.4 percent a year.[14]

A Cost-Based Measure of Banking Productivity

In a cost-based growth accounting approach (see Box 2), input prices are used in place of input quantities and costs are attached to producing bank output. The productivity results using both output measures in a cost model are shown in Figure 3. While the time pattern of the productivity indexes differ over 1977-87, they start and end at almost the same points so their annual average growth rates are again quite similar, only this time they are slightly positive—a 0.6 percent growth rate for QT and 0.5 percent for QD.[15]

The differences in productivity estimates between the production and cost approaches can be seen in Figure 4. Total factor productivity estimates

[9] This result is seen in unpublished data on the six separate components of QT (described in the Appendix) from the BLS.

[10] Real labor input is from the BLS series on number of workers in banking. The number of full-time equivalent workers from the *Call Report* grew by only 1.6 percent a year over the same period.

[11] The bank labor productivity series derived in Baily and Gordon (1988), p. 395, cannot be used for comparison here. This is because their measure of bank output growth, derived from National Income and Product Account data, is itself based on the growth of the labor input. Thus labor productivity growth will be zero by definition as the growth in bank output equals that of the labor input.

[12] The real value of these three funds categories is the nominal value divided by the GNP deflator. The real value of bank capital is described in the Appendix.

[13] More specifically, TFP using QT (QD) in the production-based growth accounting model has a growth rate of −0.0 (−0.07) percent. The difference in TFP using QD versus QT is directly related to QD being flat in the late 1970s but experiencing more rapid growth than QT in the mid-1980s (see Figure 1).

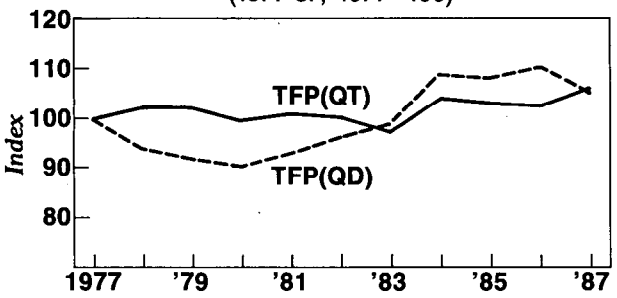
[14] Two alternative deflators for the replacement price of bank physical capital were used for illustration. These were the GNP deflator and the ratio of current capital expenditures (historical depreciation) to the book value of physical capital. For the QT output measure, average annual TFP was −0.28 percent and −0.58 percent, respectively (rather than −0.0 percent as reported above). For the QD output measure, these rates were −0.35 percent and −0.64 percent (rather than −0.07 percent as reported). All of these results use the BLS series on the number of banking workers rather than the (slower growing) number of full-time equivalent workers from the *Call Report*. Use of the *Call Report* labor data would change the QT productivity growth rate from −0.0 percent to 0.06 percent and the QD measure from −0.07 percent to 0.13 percent.

[15] As in Figure 2, the divergence between the two TFP estimates in Figure 3 is due to QD being flat in the late 1970s but having a higher growth rate than QT in the mid 1980s. Also, use of alternative deflators for the value of bank physical capital resulted in slightly lower productivity growth rates (a result similar to that obtained for the production-based measure of banking productivity—see previous footnote).

Figure 3

Cost Approach: Total Factor Productivity

(1977-87; 1977=100)



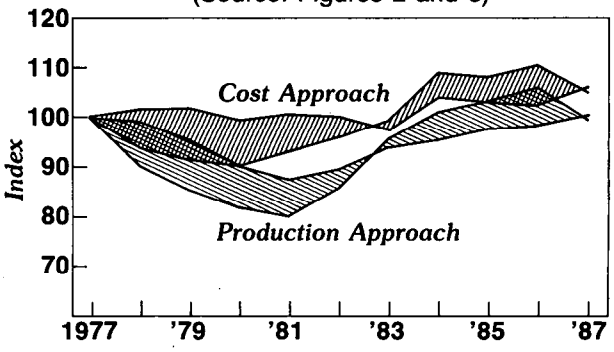
TFP(QT) = *Total factor productivity with transactions flow output*

TFP(QD) = *Total factor productivity with real balance stock output*

Figure 4

Comparison of Productivity Estimates Based on Production and Cost Growth Accounting Models

(Source: Figures 2 and 3)



derived from output and input quantities in Figure 2 are contrasted with those based on output cost and input prices in Figure 3. Results from the production approach suggest that productivity was mostly negative or zero over the period and therefore slightly lower than the cost approach, which yielded results showing zero to slightly positive productivity growth. In either case, the results show very low productivity growth, much lower than the annual 1.4 percent advance suggested in the BLS labor productivity series (Figure 2).

IV. Econometric Estimates Of Banking Productivity

No studies, to our knowledge, have attempted to econometrically estimate TFP for U.S. banks.[16] Those U.S. studies that do exist have, instead, estimated only the effect of technical change. In a standard (translog) cost function context, ln C = f(ln Q, ln P_i, t), technical advance—indexed by time t—is expressed as $-\partial \ln C/\partial t$ while scale economies are $\partial \ln C/\partial \ln Q$. Total factor productivity is the combined effect of these two measures, adjusted for the change in output (dlnQ), or:

$$(5)\quad TFP = -\partial \ln C/\partial t + (1 - \partial \ln C/\partial \ln Q)\, d\ln Q.$$

Estimates of technical change in banking have ranged from 0.96 percent a year over 1980-86 for a panel of 219 large banks (Hunter and Timme, forthcoming) to −0.90 percent over 1977-88 for a panel of 683 banks accounting for two-thirds of all bank assets (Humphrey, forthcoming).[17] In both of these studies, the scale economy estimate was so close to 1.00 that the scale adjustment to TFP in (5) has only a small effect (altering the annual values above to 1.05 and −1.01 percent, respectively). As seen in the table, the econometric estimates of banking TFP lie on either side of those from the growth accounting approach. Even so, all the estimates are relatively small, much less than one might have expected *a priori*.[18]

V. Why Was Measured Banking Productivity So Low Over The Last Decade?

Cash Management and Deregulation: The Loss of Low-Cost Deposits

In the late 1970s, historically high interest rates greatly increased the use of cash management techniques by corporations. This meant large reductions

[16] Two studies do exist for other countries; one for Canada (Parsons, Gotlieb, and Denny, 1990) and another for Israel (Kim and Weiss, 1989).

[17] The −0.90 percent figure is from one of the preferred models estimated where bank physical capital is treated as a quasi-fixed input and a time-specific dummy variable is used (instead of a simple time trend) to reflect technical change. Two other studies of U.S. bank technical change exist (Hunter and Timme, 1986; Evanoff, Israilevich, and Merris, 1989) but these were concerned with only operating costs—not total costs—and are therefore not comparable with the analysis here.

[18] Indeed, the positive productivity growth rate from the Hunter and Timme (forthcoming) study can be turned into a small negative value when two deposit interest rates are specified in their model—one for core deposits, the other for purchased funds—rather than using the purchased funds rate for both as they did (see Humphrey, forthcoming, for details).

Box 2

Growth Accounting Measures of Banking Productivity[a]

Production Approach: Total Factor Productivity

Bank output (Q) is produced by combining the real value of capital (K), labor (L), demand deposits (D), small time and savings deposits (S), and purchased funds (F) inputs according to some production relation that changes in efficiency (A) over time: $Q = A\, f(K, L, D, S, F)$. Expressed in terms of growth rates, the growth in total factor productivity ($\dot{A}/A$) is defined to be the difference between output growth and the expenditure share (w_i, i = K, L, D, S, F) weighted average of the growth in inputs:

Total Factor Productivity

$$(1)\quad \dot{A}/A = \dot{Q}/Q - w_K\dot{K}/K - w_L\dot{L}/L - w_D\dot{D}/D - w_S\dot{S}/S - w_F\dot{F}/F$$

where for X_i = Q, K, L, D, S, F:

$\dot{X}_i/X_i$ = an annual growth rate expressed as the index X_{it}/X_{it-1}, where t is time.

The use of expenditure share weights (w_i) presumes that the observed input prices—the rental price of capital, the wage rate, and the user cost of demand deposits, time and savings deposits, and purchased funds—equal the value marginal product of each input to the bank. When the w_i sum to 1.00, there is constant returns to scale.[b] The productivity measure (1) reflects total factor productivity (TFP) because the productivity effects of all inputs to the bank are being accounted for, along with returns to scale. While TFP is the most comprehensive measure of productivity, it is also the most difficult to compute because of the data required.

Multifactor and Single-Factor (Labor) Productivity

When more aggregative productivity measures are derived, such as for all manufacturing or all services, intermediate inputs are assumed to net out so only capital and labor inputs are used. The resulting measure is called multifactor productivity:

[a] This discussion is drawn from Hulten (1986).

[b] In the econometric approach to measuring productivity, the w_i are estimated statistically and need not sum to 1.00. In the growth accounting approach used here, the observed expenditure shares will sum to 1.00 by definition, imposing constant returns to scale. This restriction should only have a small effect on the results since numerous cross-section banking studies either support constant costs at the mean of all banks or are within 5 percentage points of it (so the cost elasticity of output ranges from slight economies of .95 to slight diseconomies of 1.05). See the surveys of Mester (1987), Clark (1988), and Humphrey (1990).

in idle demand deposit balances which did not pay explicit interest. The process is described and documented in Porter, Simpson, and Mauskopf (1979) and can be seen in Figure 5. Increased use of cash management techniques has emerged as the dominant explanation for the unexpectedly slow growth in the monetary aggregates during the 1970s. To compensate for the loss of demand deposits, banks came to rely more heavily on higher-cost purchased funds. Such a shift would have raised the real average cost per dollar of bank assets even if all input prices had remained constant. Since real average cost (corrected for input price changes) is the inverse of productivity, measured TFP would have fallen for this reason alone.

The negative cost effects from corporate cash management were continued with the banking deregulation of the early 1980s. Deregulation permitted noncorporate bank customers to switch from demand deposits to interest-earning Negotiable Order of Withdrawal (NOW) and Money Market Deposit Accounts (MMDAs). These new instruments inhibited the growth of demand deposits, shifting the

Multifactor Productivity

$$(2)\quad \dot{A}^*/A^* = \dot{Q}/Q - w_K\dot{K}/K - w_L\dot{L}/L$$

where $w_K + w_L = 1.00$.

The least comprehensive measure of productivity involves only the productivity of labor (LP) or output per unit of labor input: LP = Q/L. The growth in labor productivity is expressed as a reduced version of (1) or (2):

Labor Productivity

$$(3)\quad \dot{LP}/LP = \dot{Q}/Q - w_L\dot{L}/L.$$

Clearly, the growth of labor productivity in (3) will only equal the growth in TFP in (1) when labor is the only input (i.e., $w_L = 1.00$) or when the growth of other inputs are equal to that for labor (i.e., $\dot{L}/L = \dot{K}/K = \dot{D}/D = \dot{S}/S = \dot{F}/F$).

Cost Approach: Total Factor Productivity

All of the above equations showing productivity growth in terms of a production function have a corresponding cost function representation. That is, productivity can alternatively be expressed as the residual growth in average cost not accounted for by the growth in input prices over time. In simple terms, total factor productivity in a cost function context ($\dot{B}/B$) represents shifts in the average cost curve after controlling for changes in input prices:

Total Factor Productivity

$$(4)\quad \dot{B}/B = (\dot{C}/C - \dot{Q}/Q) - w_K\dot{PK}/PK - w_L\dot{PL}/PL - w_D\dot{PD}/PD - w_S\dot{PS}/PS - w_F\dot{PF}/PF$$

where:

$\dot{C}/C - \dot{Q}/Q$ = the growth rate of average cost, expressed as the growth in total cost less the growth in output; and

$\dot{PX}/PX$ = the growth rates of factor input prices and the user cost of funds, X = K, L, D, S, F.[c]

Under constant returns to scale, productivity growth using the production relationship in (1) equals minus one times the productivity growth from the cost relationship in (4) or $\dot{A}/A = -\dot{B}/B$.[d]

[c] The measurement of these variables is discussed in the Appendix.

[d] $\dot{A}/A$ is positive because increases in productivity in (1) increases output while $\dot{B}/B$ is negative as increases in productivity in (4) reduces cost.

deposit expansion which did occur into interest-earning time and savings deposits (see Figure 5).[19]

Prior to deregulation, banks had substituted convenient branch offices, service personnel, and nonpriced services (e.g., free checking) for their inability to pay something close to a market rate on demand, savings, and small time deposits (Evanoff, 1988). Once deregulation removed interest rate ceilings and permitted consumer interest checking, banks quickly paid higher rates for the same funds. From a cost standpoint, banks subsequently found themselves to be "overbranched." The profitability of their deposit base fell from $61 billion in 1980, in constant 1988 dollars, to $4 billion in 1988 (Berger and Humphrey, forthcoming).

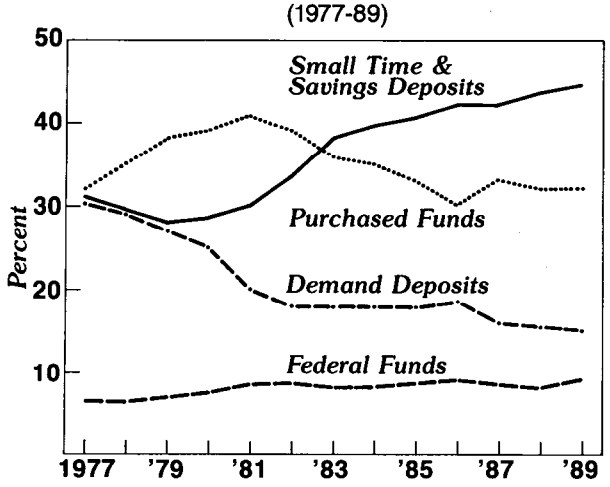
In effect, corporate cash management and deregulation removed banks' virtual monopoly control over zero-interest checking accounts and low-interest small

[19] While checks can be written on NOW and MMDA balances, they are not (legally speaking) available on demand and so have been classified with time and savings deposits in the data collected by regulatory authorities.

Figure 5

Percent Composition of Aggregate Bank Liabilities

(1977-89)



consumer time and savings deposits (as rate ceilings on these deposits were also removed).[20] Subsequent competition induced banks to shift from low-to higher-interest cost funds inputs without a fully offsetting reduction in factor inputs used to provide branch convenience and other low-priced deposit services. In addition, since the deposit services provided were largely unchanged as corporations conserved on idle balances and consumers shifted from one type of checking account to another, either measure of bank output used here would have been stable. With costs rising but output stable, costs per unit of measured output should rise, even when corrected for input price changes, lowering TFP.

In addition to cash management and deregulation, the inflation of the late 1970s and early 1980s also contributed to the rise in bank costs. During this inflationary period, some idle demand balances and low-cost time and savings deposits would have continued to shift to Money Market Mutual Funds (MMMFs) and been replaced by higher cost CDs sold by banks to the MMMFs. But in order to control operating costs, MMMFs restricted the number of checks written per month and specified high minimum amounts. Such limitations would likely have prevented any substantial disintermediation of demand deposits and thereby helped keep bank costs relatively low. Since over 80 percent of the deregulated bank balances were NOW and MMDA deposits (which experienced the largest rate increases following deregulation), it is clear that the great majority of the negative effects for banks seen during this period are due to deregulation, not inflation.

This analysis, we believe, explains why researchers have failed to observe much positive net technical change or productivity growth in banking during the last decade. Going beyond this explanation, part of the problem is also related to our inability to accurately capture all potentially important aspects of bank output. If branch convenience and the continued provision of underpriced deposit services are valued by users, then certainly some of the (now extra) costs incurred by banks in providing "unnecessarily" high levels of these services after deregulation have served to increase the quality of bank output. If one adopts this view, then what appears to be a productivity decrease may instead be the result of understating output growth as benefits received by bank depositors rose relative to their pre-deregulation level.

An analogous situation occurred in the electric utility industry during the 1970s. Expensive pollution control restrictions were mandated for electric utilities and, although these costs were largely made up by rate increases, measured output of this industry—kilowatt-hours—did not rise commensurately. As a result, measured total factor productivity was seen to fall (e.g., Gallop and Roberts, 1983). But if cleaner air resulted, then the quality of this industry's output actually rose but will not be captured in the output measure used. It is argued here that the same sort of thing occurred in banking.

Market-Share Reasons for Not Reducing Branch Convenience as Interest Costs Rose

It is easy to argue that the cost effect of deregulation could have been minimized if all banks had pared their branch operations more rapidly and to a greater degree. As it was, the real deposit/branch ratio was still falling until 1982, when it reached a minimum of around $28 million in core deposits per branch office. This meant that banks were still effectively building branches more rapidly than its customer base was expanding, increasing convenience (and operating costs) in the process. While the employee/branch ratio was more or less falling continually over this period, only after 1982 did the real deposit/branch ratio start to rise, reaching around $36 million in 1988.

Seemingly, market share considerations inhibited a more rapid and comprehensive reduction in bank

[20] Another aspect of deregulation was that thrift institutions obtained the ability to offer checkable deposits. This increased competition and contributed to the reduction in banks' monopoly power over this low-cost product.

operating costs as interest expenses from deregulation rose. Since choice of a bank by a depositor is largely based on convenience (according to industry surveys), a dramatic and profitable reduction in one bank's branching network would serve also to expand market share and profits at competing banks that retained their branch networks. In the end, both sets of banks would have experienced higher profit *rates* in the short run, but market shares and profit *levels* would have been redistributed away from those banks that cut their branch networks the most. Thus most banks seemingly chose to sacrifice short-term profits in order to maintain market share and hoped that long-term profit would follow as deposit growth continued to exceed the establishment of new branches.

Outlook for the Future

The outlook is not very bright. First, the wave of interstate mergers that have occurred already, along with those expected during the 1990s (when many states will eliminate their existing out-of-state merger barriers), bring with them costly "one-time" expenditures to integrate back office operations and standardize the banking products offered. While these expenditures will permit some cost reductions to be realized, they will also add considerable software and equipment expenses.

Second, the problem of excess banking capacity, as evidenced by too many branches, cannot easily be solved as long as failed or failing banks and thrifts continue to be purchased by institutions with the bulk of their own branch network typically outside of the purchased bank's deposit market area. Rarely do regulators simply close a failed bank's branches, and rarely do banks in the same market area purchase branches simply to close them. Instead, a failed bank's branch network is typically sold to an institution outside the market area and the buyer typically keeps most of the branches open, perpetuating the oversupply problem.

If the antitrust market concentration restrictions on bank mergers were considerably relaxed, then costs associated with overlapping branch networks would fall. Such cost reductions result when large competitors in the same deposit market area are encouraged to acquire each other and close excess branch offices (e.g., as occurred with Crocker and Wells Fargo in California). While market concentration would rise, it is not clear that increased concentration would or has led to much uncompetitive behavior in the form of reduced price competition and increased profits. Indeed, recent research indicates that low costs are the dominant explanation for higher bank profits in concentrated markets (Timme and Yang, 1990), not concentration itself as has long been asserted. Overall, given the two problems just outlined, it is hard to be optimistic about the future of productivity in banking. The most likely outcome is continued slow growth until the industry is able to shrink itself sufficiently through greater reductions in operating costs per dollar of deposits or assets. Thus future productivity growth will more likely stem from reducing current excess costs than from further technological progress.

VI. SUMMARY

Measured productivity in banking over the last decade has been growing at a very low rate. Using aggregate data over 1977-87, it is estimated that total factor productivity growth has only been between −0.07 to 0.60 percent a year.[21] These estimates are based on a nonparametric growth accounting approach using first a production function and second a cost function. These results were robust to a number of influences (three different deflators for deriving the real value of bank physical capital and two different labor employment series). Importantly, these results are also robust to using two different indicators of banking output: one a flow measure of deposit and loan transactions and the other a stock measure of the real value of deposits and loan balances.

The primary explanation for the low productivity growth experienced has been the shift in zero-interest cost corporate and some consumer demand deposits to purchased funds in the 1970s (a result of improved corporate cash management techniques, higher interest rates, and the rise of Money Market Mutual Funds), plus a later shift of consumer demand deposits to interest-earning and checkable time and savings deposits in the 1980s (a result of banking deregulation which removed interest rate ceilings on time and savings and established new interest-earning checking accounts at both banks and thrifts). These developments significantly raised the cost of bank loanable funds. However, banks did not fully offset these higher costs by lowering operating expenses, reducing branch and service convenience, to compensate for the higher interest being paid. It is argued that market share considerations limited this response.

---

[21] Similarly low positive to low negative annual rates of productivity growth have also been found over a longer period, 1967-87 (Humphrey, 1991).

The outlook for the future is not bright. What is necessary is a substantial reduction in operating costs, since banking no longer has a virtual monopoly over zero-interest checking accounts and low-interest small consumer time and savings deposits. Future bank mergers, while reducing costs in some instances, will also lead to expensive "one-time" expenditures to integrate back office operations and standardize banking products. And bank failures, rather than removing excess branch office capacity as would occur in other industries, have tended to perpetuate the overcapacity conditions that have led to higher costs. Increases in banking productivity, when they come, are more likely to result from reductions in current operating costs and a rationalization of overlapping branch networks than from further technological progress.

APPENDIX

Availability of Data and Measurement of Banking Output and Price Indexes

Data Availability

Aggregate data on the number of deposit accounts from the FDIC are only available for two years over the past ten, while no aggregate data are available on the number of (new plus outstanding) loan accounts. While numbers of deposit and loan accounts are reported in the Federal Reserve's annual *Functional Cost Analysis* survey, the data cannot be used in a time-series analysis. First, the sampled banks change by upwards to 15 to 20 percent each year so that a consistent time series covering the same set of banks is not available. Second, the very largest banks, those that service the largest number of such accounts and experience the greatest rate of growth, are not included in the survey.

Indexes of Bank Output

The transactions flow index of banking output (QT) was developed by the Bureau of Labor Statistics (BLS, 1989). This index measures demand deposit output by the number of checks and electronic funds transfers processed, which reflects the debiting and crediting of demand deposit accounts as well as the payment processing and accounting activities associated with these activities. Similarly, savings and small denomination time deposit output is captured by measuring deposit and withdrawal activity in these accounts. Loan output is represented by the number of new real estate loans, consumer installment and credit card loans, and commercial, industrial, and agricultural loans made during the year. Lastly, trust and fiduciary activities are assumed to be proportional to the number of trust accounts serviced. Investment activities are treated as an intermediate good and netted out, since their variation has historically been associated with secondary reserves (where securities are sold to fund higher-than-expected loan demand or deposit withdrawal activity and vice versa). In any event, investment activities, plus the provision of safe deposit boxes, investment advice, and insurance, account for only a little more than 4 percent of bank employment, and their omission is not believed (by the BLS) to have a significant effect on the variation in measured output. Employment shares were used to weight these separate transaction flows into a single index of banking output.

The alternative index of the real value of deposit and loan account balances (QD) was developed by the author. It represents a cost-share weighted average of the dollar value of five deposit (demand deposits, small time and savings deposits) and loan categories (real estate loans, consumer installment and credit card loans, and commercial, industrial, and agricultural loans) from aggregate *Call Report* data. The cost-share weights are from the annual *Functional Cost Analysis* surveys for banks with more than $200 million in deposits. Nominal values of these five output categories were deflated by the GNP deflator to approximate real values.

Total Cost of Output and Input Prices

Total cost is from the *Call Report* and excludes double counting at the aggregate level by deleting the cost of purchased federal funds (see text). The price of capital is a bank-weighted average of the new contract cost per square foot of bank and office building space for nine regions of the United States reported in F.W. Dodge, *Construction Potentials Bulletin* (various years). Other capital price deflators were also used and their effects are noted in the text (footnote 14). The real value of bank physical capital used is book value deflated by the capital price index. The price of labor is total expenditure on labor divided by the number of full-time equivalent workers (both from the *Call Report*). The prices per dollar of each of the three funds categories are in terms of user costs, composed of the interest rate paid (i), the per dollar reserve requirement (RR), and the per dollar service charge income (SC). Following Hancock (1986), but neglecting FDIC deposit insurance costs, user costs (UC) are in general $UC = (i + r_{FF} RR - SC)/(1 + r_{FF})$, where r_{FF} is the rate on federal funds, a market rate. The denominator adjusts for the fact that the numerator costs are only fully realized at the end of a one-year period, rather than at the beginning. RR and SC are small for time and savings deposits and are difficult to separate out from those on demand deposits, for which i is zero. With these considerations in mind, our user costs are: $UC_D = (r_{FF} RR - SC)/(1 + r_{FF})$; $UC_S = i_S/(1 + r_{FF})$; and $UC_F = i_F/(1 + r_{FF})$. In implementation, total costs and the two factor input prices were deflated by the GNP deflator to reflect real values. User costs are already in real terms (see Hancock, 1986).

REFERENCES

Baily, Martin N., and Robert J. Gordon, "The Productivity Slowdown, Measurement Issues, and the Explosion of Computer Power," in William Brainard and George Perry (Editors), *Brookings Papers on Economic Activity* 2 (1988), The Brookings Institution, Washington, DC: 347-420.

Benston, George, and Clifford Smith, Jr., "A Transactions Cost Approach to the Theory of Financial Intermediation," *Journal of Finance* 31 (May 1976): 215-31.

Berger, Allen N., and David B. Humphrey, "Measurement and Efficiency Issues in Commercial Banking," in Zvi Griliches (Editor), *Output Measurement in the Services Sector*, University of Chicago Press (forthcoming).

Board of Governors of the Federal Reserve System, *Functional Cost Analysis*, National Average Report, Commercial Banks, Washington, DC (various years).

———, *Consolidated Report of Condition and Income*, Washington, DC (various years).

Clark, Jeffery, "Economies of Scale and Scope at Depository Financial Institutions: A Review of the Literature," Federal Reserve Bank of Kansas City *Economic Review* 73 (September/October 1988): 16-33.

Cullison, William E., "The U.S. Productivity Slowdown: What the Experts Say," Federal Reserve Bank of Richmond *Economic Review* 75 (July/August 1989): 10-21.

Evanoff, Douglas D., "Branch Banking and Service Accessibility," *Journal of Money, Credit and Banking* 20 (May 1988): 191-202.

Evanoff, Douglas D., Philip R. Israilevich, and Randall C. Merris, "Technical Change, Regulation, and Economies of Scale for Large Commercial Banks: An Application of a Modified Version of Shephard's Lemma," Working Paper, Federal Reserve Bank of Chicago, Chicago, IL (June 1989).

F.W. Dodge Division, *Dodge Construction Potentials Bulletin*, Summary of Construction Contracts for New Addition and Major Alteration Projects, McGraw Hill, New York (various years).

Fixler, Dennis J., and Kimberly D. Zieschang, "User Costs, Shadow Prices, and the Real Output of Banks," in Zvi Griliches (Editor), *Output Measurement in the Services Sector*, University of Chicago Press (forthcoming).

Gallop, Frank M., and Mark J. Roberts, "Environmental Regulations and Productivity Growth: The Case of Fossil-Fueled Electric Power Generation," *Journal of Political Economy* 91 (August 1983): 654-74.

Hancock, Diana, "A Model of the Financial Firm with Imperfect Asset and Deposit Elasticities," *Journal of Banking and Finance* 10 (March 1986): 37-54.

Hulten, Charles R., "Productivity Change, Capacity Utilization, and the Sources of Efficiency Growth," *Journal of Econometrics* 33 (October/November 1986): 31-50.

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

Hunter, William C., and Stephen G. Timme, "Technological Change and Production Economies in Large U.S. Commercial Banking," *Journal of Business* (forthcoming).

Humphrey, David B., "Cost and Technical Change: Effects of Bank Deregulation," *Journal of Productivity Analysis* (forthcoming).

———, "Flow Versus Stock Indicators of Banking Output: Effects on Productivity and Scale Economy Measurement," Working Paper, Federal Reserve Bank of Richmond, Richmond, VA (May 1991).

———, "Why Do Estimates of Bank Scale Economies Differ?," Federal Reserve Bank of Richmond *Economic Review* 76 (September/October 1990): 38-50.

Kim, Moshe, and Jacob Weiss, "Total Factor Productivity Growth in Banking: The Israeli Banking Sector 1979-1982," *Journal of Productivity Analysis* 1 (1989): 139-53.

Mamalakis, Markos J., "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 1987): 169-92.

Mester, Loretta, J., "Efficient Production of Financial Services: Scale and Scope Economies," Federal Reserve Bank of Philadelphia *Economic Review* 73 (January/February 1987): 15-25.

Parsons, Darrell, Calvin Gotlieb, and Michael Denny, "Productivity and Computers in Canadian Banking," Working Paper, Department of Economics, University of Toronto, Canada (June 1990).

Porter, Richard, Thomas Simpson, and Eileen Mauskopf, "Financial Innovation and the Monetary Aggregates," *Brookings Papers on Economic Activity* 1 (1979), The Brookings Institution, Washington, DC: 213-29.

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

Timme, Stephen G., and Won K. Yang, "On the Use of a Direct Measure of Efficiency in Testing Structure-Performance Relationships," Working Paper, Department of Finance, Georgia State University, Atlanta, GA (September 1990).

U.S. Department of Labor, Bureau of Labor Statistics, *Productivity Measures for Selected Industries and Government Services.* Bulletin 2322 (February, 1989): 170.

Wykoff, Frank C., "Commercial Banking Productivity Growth: Evidence from Large Bank Balance Sheets," Working Paper, Department of Economics, Pomona College, Claremont, CA (January 1991).