Why Do Estimates of Bank Scale Economies Differ?*

David B. Humphrey

I. INTRODUCTION

A number of policy issues turn on whether or not large commercial banks, merely because of their size, are more efficient than small banks. Such scale economies, where average cost declines as bank output rises, would result from spreading fixed costs over a greater volume of output. Scale economies are an important policy consideration in interstate bank branching.

Interstate branching was long prohibited on the grounds that (1) industry concentration and monopoly power would result, and (2) local areas may be less well served by giant banks having little interest in these localities, as more profitable uses for funds would likely be found elsewhere. Cost savings associated with large scale economies, however, might overcome these negatives. As well, interstate branching would allow banks to diversify their portfolios geographically, strengthening the industry. Consumer and business bank customers would likely benefit from lower prices and reduced banking risks which could follow.

In contrast, if scale economies were small, fears of concentration might outweigh any perceived benefits of expansion. It would then be more politically tenable to limit the size and geographical distribution of banks. While there still could be loan risk diversification, this benefit by itself might not justify the concentration of economic power in truly giant banking organizations.

The level of bank scale economies is an empirical question, but one where widely differing results have made it difficult to form a clear and unambiguous conclusion. Fortunately, there are now enough studies to attempt to sort out *why* past results have differed. Such a sorting out is useful in its own right and for the implications it has for policy decisions that depend on scale economies in banking. It also illustrates the benefits a detailed analysis could have for other areas of economics where empirical findings differ and can cloud proper policy formation (such as in the appropriate definition of the money supply).

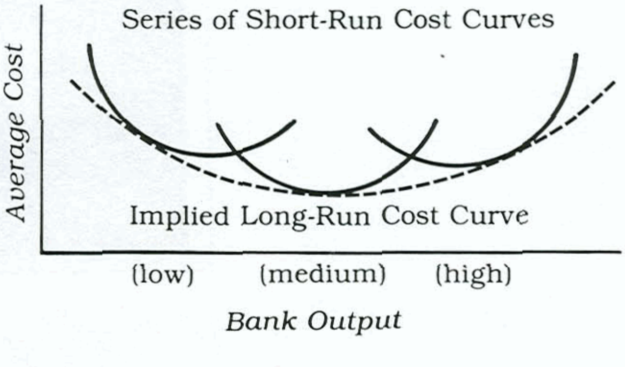
In many sciences, researchers use the same experimental technique to generate new and independent data and then look for consistency in the results. In contrast, economists generally use similar data but vary the experimental technique—that is, the particular specification and definition of variables, functional form, and time period used. Thus robust results are less frequent. If enough studies are performed, however, a pattern to the results may emerge suggesting why they differ. Then we can compare the relative advantages of different experimental techniques. Instead of a single scale economy conclusion that applies in all cases, we obtain a set of different results that illustrate how sensitive our measures are to the research design chosen. From this and from some additional thought on how we best measure scale economies, we develop a general conclusion on the size and significance of scale economies in banking.

II. COMMON DIFFERENCES AMONG STUDIES

Graphically, bank scale economies appear as the slope of an average cost curve indicating how costs vary with output. An example is shown in Figure 1. A series of short-run average cost curves (solid lines) for three different-sized banks, each producing different levels of bank output, trace out an implied long-run average cost curve (dotted line). A downward-sloping long-run average cost curve reflects scale economies. An upward slope reflects diseconomies, since higher average costs are incurred when more output is produced. The assumption is that a cross-

Figure 1

Illustrative Bank Average Cost Curves



[*] Comments by Mike Dotsey, Bob Graboyes, Tom Humphrey, and Dave Mengle are appreciated, although the opinions expressed are those of the author alone. Able research assistance was provided by Bill Whelpley.

section of different-sized banks at a point in time will reveal the appropriate long-run curve; from this is derived a measure of scale economies. Thus as smaller banks expand their output in the future, their costs are likely to "look like" the costs of larger banks today.

The cost curve itself (and the implied scale economies reflected in it) is actually derived from an equation similar to (1), below, where costs (C) are regressed on the level of bank output (Q) and other variables which affect costs but need to be held constant in the cross-section data set:

(1) $C = f(Q, \text{other variables})$.

Other variables, such as the prices of labor and capital factor inputs, need to be held constant in a cross-section in order to statistically separate movements *along* the cost curve (due to changes in output) from *shifts* in the cost curve (due to influences on bank costs which are essentially unrelated to output).

With this background, we now outline the most common differences observed in bank scale economy studies and assess how these differences have affected the results derived from them. More specifically, our purpose is to critically review the literature on bank scale economies, to select a preferred method for estimating these economies, and thereby to determine which empirical result is the most appropriate for policy purposes, as well as defensible on theoretical grounds. The most common research design differences among studies of bank scale economies concern the following:

(1) Cost definition (operating cost versus total cost);

(2) Bank output definition (numbers of accounts versus dollars in these accounts);

(3) Functional form used (linear versus quadratic);

(4) Scale economy evaluation (single office versus banking firm);

(5) Time period used (high versus low interest rate period);

(6) Commingling scale with scope (single versus multiple output); and

(7) Bank efficiency differences (assume all observations are efficient versus only those on the efficient frontier).

In the following sections, each of these differences is discussed in conjunction with one or more published studies. Some other differences occur and, when appropriate, they too are noted.

III.
Operating Versus Total Bank Costs

This section concerns how the dependent variable—cost (C)—is defined in equation (1). Many studies relate only *operating* costs to bank output levels in estimating scale economies (Langer, 1980; Nelson, 1985; Hunter and Timme, 1986; Evanoff, Israilevich, and Merris, 1989). Operating costs include wages, fringe benefits, physical capital, occupancy, and materials cost, along with management fees and data processing expenses paid to the holding company and other entities. On average, operating costs only comprise slightly over 25 percent of total costs. Most other studies have used total costs, which are obtained by adding interest expenses on purchased funds and core deposits to operating costs.[1] The two interest cost categories are large and each exceed operating costs since they comprise around 35 and 40 percent, respectively, of total costs. Clearly, it makes a difference which definition of cost is used to derive an estimate of scale economies.

The difference in cost definitions—operating versus total costs—would not be an issue if all banks had the *same* percentage composition of interest and operating expenses regardless of their size. This is because interest expenses typically have little or no economies associated with them. Therefore, adding these roughly constant cost expenses to operating costs (giving total costs) means that any scale economies or diseconomies found using operating costs alone would only be attenuated, rather than reversed, if the ratio of interest to operating costs were the same across banks. But this ratio is not even close to being stable across banks. The proportion of assets funded with purchased funds rises substantially as banks get larger so that the proportion of purchased funds interest expense in total cost rises while the proportion of core deposit interest expense and operating cost falls.

For example, at small branching banks (those with \$50 to \$75 million in assets in 1984), purchased funds were 12 percent of the value of core deposits plus purchased money. For medium-sized banks (with \$300 to \$500 million in assets), the purchased funds proportion rises to 19 percent. And for large banks (with \$2 to \$5 billion and then over \$10 billion in

[1] Purchased funds are purchased federal funds, CDs of \$100 thousand or above, and foreign deposits (which are almost always over \$100 thousand). Core or produced deposits are demand deposits and small denomination (i.e., less than \$100 thousand) time and savings deposits. The costs of equity and subordinated notes and debentures are small and are almost always excluded from bank cost studies.

assets), the proportion rises further to 36 and finally 60 percent. At unit state banks for the same four size groupings, the purchased funds proportions are 16, 31, 61, and 78 percent. Thus the percentage composition of interest and operating expenses varies considerably across banks and is closely related to bank size, which is the key to the problem which arises when operating costs are used.

Purchased funds have very low operating expenses per dollar raised; their only significant cost is interest expense. In contrast, core or produced deposits generate the major portion (49 percent) of all operating (capital, labor, materials) expenses. Since purchased funds are a strong substitute for core deposits, the interest expense of purchased funds is also a substitute for the operating and interest expenses of core deposits. To accurately gauge how bank costs really change with size thus requires that purchased funds and core deposit interest expenses be included with operating costs. Taken together, these components allow one to determine the average cost actually faced by a bank even as its funding mix is altered. In this way, changes in the funding mix do not bias the results.

This point is illustrated by comparing the actual *average operating cost* (operating expenses divided by total assets) for 1984 with the *average total cost* (operating plus interest expenses divided by total assets) for the same year across 13 size classes of banks (see Figure 2). The branching state bank comparison is shown in Panel A with the unit state bank comparison in Panel B.[2] Operating cost per dollar of assets is seen to fall more rapidly than total costs per dollar of assets. Thus if only operating costs are used in a statistical analysis of bank scale economies, as some investigators have done, greater scale economies (or lower diseconomies) will typically be measured when an equation like (1) is estimated and a curve is fitted to these raw data points.[3]

Hunter and Timme, 1986, obtained this result when they alternatively used operating costs and then operating plus interest costs in their statistical estimates of scale economies for 91 large bank holding companies over 11 years (1972-82). They found significant operating cost scale economies (using only operating costs) but no significant total cost scale economies (when interest expenses were included). Their study covered large banks separately and did not include any small or medium-sized institutions.

While operating costs are of some interest in themselves, it would be misleading to conclude that reductions in the ratio of operating costs to assets accurately reflects inherent bank scale economies. If this were true then a bank with a wholesale orientation (large purchased funds, small core deposits) would always experience lower costs solely because of lower operating costs per dollar of assets. But lower operating costs per dollar of assets are typically offset by having greater interest costs per dollar of assets through more intensive reliance on purchased funds instead of core deposits. Thus the proper comparison of costs, and measurement of scale economies, must rely on total costs rather than only on operating costs by themselves. When this is done, then differences in a bank's funding mix will not bias the results.[4]

IV.
Bank Output Measurement: Number of Accounts versus Dollars in the Accounts

Another important difference in published studies concerns the definition of bank output (Q), a key independent variable in equation (1). In most other industries, the measurement of output is not a problem. Output is a flow concept measured in physical terms, either because the physical unit is homogeneous and can be easily observed or because there is a convenient index of the value of the output flow which can be deflated by an appropriate output price index. In banking, neither of these alternatives exists and data availability dictates how bank output is defined. Output flow information is not available for each individual bank so information on the stock of output is used instead. Generally, researchers assume that the unobserved output flow is proportional to the observed output stock. Thus use of stock information in statistical analyses is presumed to give results similar to those obtainable using flow data.

[2] The top line in each comparison is the mean average total cost curve (solid line). To make this comparison clearer, the scale for average operating costs—right side of the figure—has been shifted up so that the two curves will appear to start from the same point for the first size class. The scale for average total costs is on the left side.

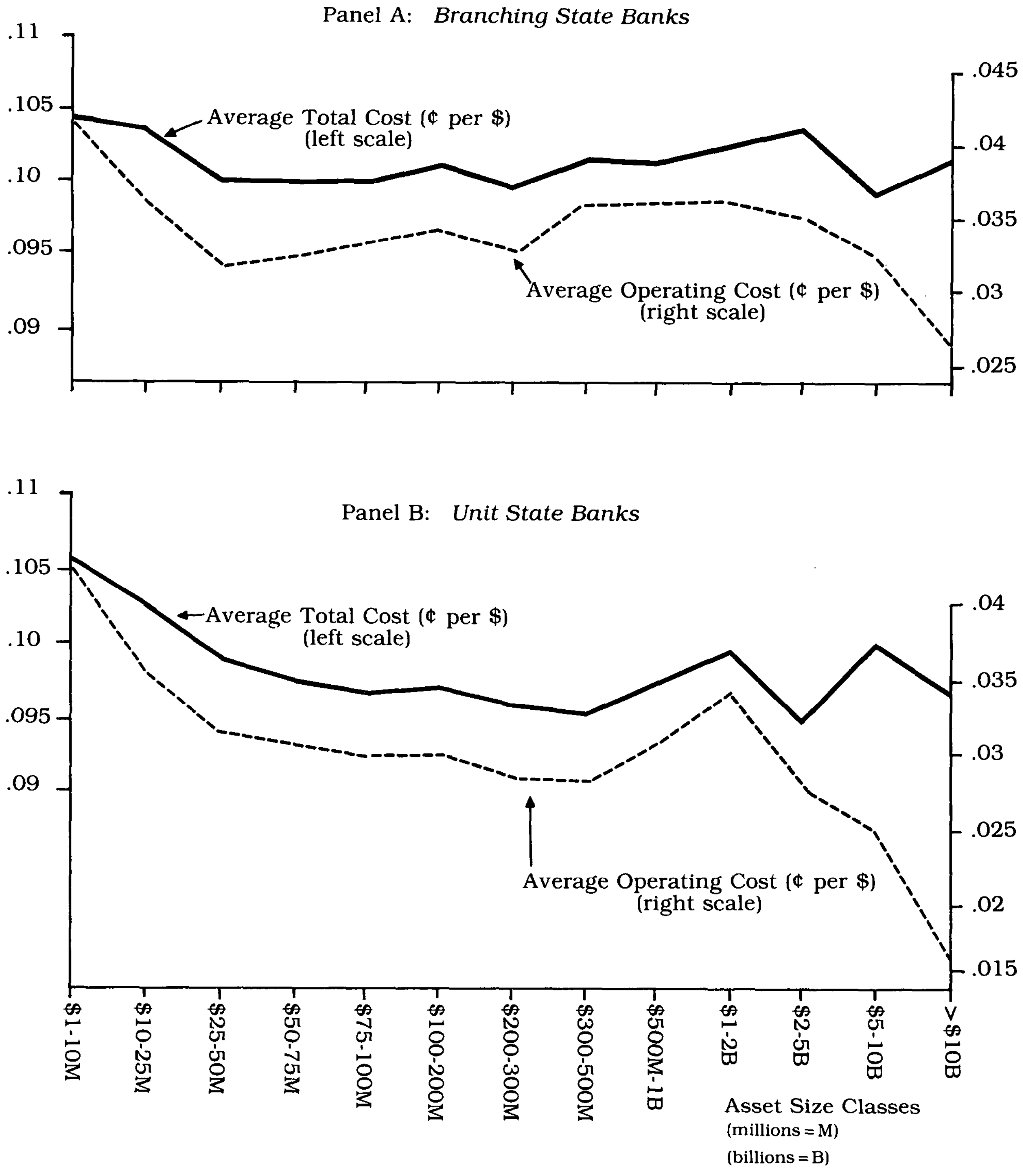
[3] The same sort of bias toward finding scale economies when only operating costs are used also exists for thrift institutions. This can be seen in the raw data presented in Verbrugge, McNulty, and Rochester, 1990, Table 1.

[4] If the U.S. banking system were considerably more consolidated, as could occur if full interstate branching were permitted, then the importance of purchased funds would of course be reduced. Once this occurs, looking at operating cost per dollar of assets could be more revealing. There would be less substitution of purchased funds for produced deposits and the funding mix bias that exists in current studies using only operating cost would be attenuated.

Figure 2

Comparing Actual Average Operating and Average Total Cost

(1984 data points)



Panel A: *Branching State Banks*

Panel B: *Unit State Banks*

Data on the number of deposit and loan accounts, an output stock measure, also are not available for all banks. Nevertheless, some information is given in the Federal Reserve's annual *Functional Cost Analysis* (FCA) survey. This survey covers 400 to 600 banks but typically excludes the very largest (those with more than $1 billion in assets). Also, the same banks are not in the sample each year.[5] Alternatively, the value of dollars in the various deposit and loan accounts, another output stock measure, is publicly available for each individual bank in every year in the *Report of Condition and Income* (Call Report).

Some researchers have made a strong argument for using the number of accounts as an indicator of bank output (Benston and Smith, 1976). Fortunately, it turns out that the scale economy results are reasonably robust to the use of either the number of accounts or the dollar value in the accounts. That is, using both of these alternative representations of bank output in the same model for the same year leads to similar scale economy results (Benston, Hanweck, and Humphrey, 1982; Berger, Hanweck, and Humphrey, 1987). This occurs because these two approximations to bank output, while numerically quite different, are highly correlated, both in the U.S. and elsewhere (see Berg, Forsund, and Jansen, 1990).

A preferable measure for bank output would measure the flow of some physical aspects of bank output rather than just the stock of accounts serviced or their dollar values. While the Bureau of Labor Statistics compiles such a measure annually, it applies only to the aggregate of all banks in the U.S. (BLS, 1989). This aggregate flow measure is a specially weighted index of the number of checks processed (for demand deposit output), the number of savings account deposits and withdrawals (for savings and small-denomination time deposit output), the number and type of new loans made (for various loan outputs), and the number of trust accounts serviced (for trust output).[6]

Over a recent 10-year period (1977-86), the BLS aggregate measure of bank output rose by 40.4 percent. Over the same period, a cost share-weighted index of the *value* of demand deposits, savings and small time deposits, real estate, installment, and commercial and industrial loans (all deflated by the GNP deflator) rose by 43.8 percent (Humphrey, forthcoming). These 5 output stock categories accounted for around 75 percent of bank value-added during the 1980s and so clearly reflect the majority of services produced by banks (in a flow sense). Importantly, the similar growth rates indicate, at the aggregate level at least, that the flow and stock measures of bank output closely correspond to one another. This suggests that use of a stock measure of bank output (the only one available at the individual bank level for all banks) may be a reasonable approximation of the unobserved flow measure for recent time periods. Thus it would seem that little bias has been introduced in past scale economy studies when a stock of output measure is used in place of a flow measure. Also, either the stock of accounts or the stock of dollars in those accounts seems to give qualitatively similar scale economy results (when properly used in the same model).

A related issue, often noted in the literature, concerns the similarity of the survey bank data from the FCA versus that for the population of all banks in the Call Report. The only published study addressing this issue concluded that while there were statistically significant differences between the FCA sample and the Call Report population data (in terms of portfolio composition, capital/asset ratio, and total cost/asset ratio), these differences were quantitatively small. In fact, FCA banks in 1970 experienced mean average costs which were 6 percent lower than the average costs for the mean of the non-FCA bank size-matched sample (Heggestad and Mingo, 1978). Updating this comparison for 1984, but using all banks, we find that the mean difference is now only 3 percent, and most of this arises for banks with the highest costs. Thus, FCA data should not lead to markedly different scale economy results compared to use of data on all banks, or on only large banks not covered in the FCA sample.

V.
A Linear versus a Quadratic Functional Form

Historically, bank scale economies were typically estimated using a linear functional form for equation (1), such as the log-linear Cobb-Douglas form.[7] Such forms were commonly used in cost or production analyses in areas where the research emphasis was

[5] The sample has varied by as much as 15 to 20 percent each year. Also, credit unions and thrift institutions (such as MSBs) can and do participate in the FCA survey. In 1984, the participation rate of thrift and credit unions was almost 17 percent of the total sample.

[6] The FCA data also provide physical flow information, similar to that used by the BLS, but these data are available only for banks in the survey, not for all banks.

[7] Greenbaum, 1967, is an important exception as he used a simple quadratic equation and, as a result, found a U-shaped average cost curve (in contrast to studies using a Cobb-Douglas form).

on factor shares in the distribution of income and on estimating the various sources of output growth over time. Unfortunately, one property of the log-linear Cobb-Douglas form is that the *same* cost economies or diseconomies will be measured for *all* banks in the sample regardless of their size. Put differently, all banks will either have scale economies, scale diseconomies, or constant costs. A U-shaped long-run cost curve, similar to that illustrated in Figure 1, cannot be estimated when only Q enters the regression equation (1). What is needed is a specification that includes Q and Q^2, making (1) a quadratic equation.

Earlier studies, such as the comprehensive analyses of Benston, 1965 and 1972, and Bell and Murphy, 1968, used a Cobb-Douglas form and found that scale economies existed in many banking services.[8] Overall, these economies were relatively small. The average scale economy value was .92.[9] This means that for each 10 percent increase in bank output, costs rise by only 9.2 percent, so average costs would be estimated to fall as a bank gets larger. A scale economy value greater than one—say 1.05—would have suggested a 10.5 percent rise in costs for each 10 percent increase in output (thus reflecting scale diseconomies).

Recently, more flexible functional forms have been developed and used. One of the most common is the translog form, which is a quadratic form. That is, the translog has linear output terms, like the Cobb-Douglas, but also squared output terms. As a result, the translog form can estimate a U-shaped cost curve if one exists in the data. If a U-shaped cost curve were in fact estimated, it would show scale economies at smaller banks and diseconomies at larger ones, like that illustrated in Figure 1. Unlike the Cobb-Douglas form, quadratic forms capture variations of scale economies across different sizes of banks.

Studies using the translog form, such as Gilligan, Smirlock, and Marshall, 1984, Lawrence and Shay, 1986, or Benston, Hanweck, and Humphrey, 1982, generally find that bank cost curves are weakly U-shaped. Scale economies exist in banking but seem to be limited to the relatively smaller banks. Either constant costs (for banks in branching states) or some scale diseconomies (for those in unit banking states) seems to apply to larger institutions. Since under certain restrictions the translog reduces to the Cobb-Douglas form, it is possible to see if these restrictions significantly reduce the ability of the model to fit the underlying data. In these tests, the Cobb-Douglas has been rejected in favor of the more general translog form. That is, the restrictions the Cobb-Douglas form places on the translog model (equal scale economies for all sizes of banks and all elasticities of factor input substitution equal to 1.0) are rejected.

Use of the translog instead of the Cobb-Douglas is one way these restrictions can be relaxed. Another way is through a specialized adjustment (called a Box-Cox adjustment) to the Cobb-Douglas model, as applied by Clark, 1984, and Lawrence, 1989. With such an adjustment, Clark finds only scale economies in his small and medium-sized unit bank data set (the largest bank had only $425 million in assets). In contrast, Kilbride, McDonald, and Miller, 1986, find scale economies at small unit banks but diseconomies at large ones using the same technique as Clark. Since the Kilbride, et al. study differs in two respects—it covered a later time period (1979-83 versus Clark's 1972-77) and added large unit banks up to $10 billion in assets to the unit bank sample—it is not clear which change led to the reversal in Clark's results: the different time period covered, the inclusion of large banks, or both.

Recently, Lawrence, 1989, generalized the Box-Cox adjustment of the Cobb-Douglas model by adding the possibility of multiple outputs—either multiple classes of loans or loans plus certain types of deposits. Both the Clark and the Kilbride, et al., studies had used a single composite measure of bank output. With this adjustment, both the multiple output translog and the single output Cobb-Douglas forms can be tested to see which form best fits the data. The single output Cobb-Douglas form, even with a Box-Cox adjustment, was rejected in favor of the multiple output translog. Thus it appears that both the possibility of U-shaped cost curves and cost complementarities among different bank outputs are important generalizations of the single output Cobb-Douglas form (which cannot reflect either of these more flexible specifications). In sum, a functional form that permits the estimated average cost curve to be U-shaped, rather than monotonic, is preferred. Thus a quadratic form dominates a linear form when

[8] Squared terms of some independent variables were used in Benston's regressions but only rarely applied to the output variables. Thus U-shaped cost curves could not, except in these infrequent cases, be estimated.

[9] Simple averages of Benston's, 1965, direct and indirect expense scale economies were .87 and .98, respectively (Table 26, p.544). As indirect expenses were 43 percent of total operating expenses, this yielded a weighted average scale economy of .87(.57) + .98(.43) = .92. Bell and Murphy obtained an overall scale economy of .93 (Table 4, p.8).

measuring bank scale economies and typically yields different scale economy conclusions as well.

Closely related to the choice of a proper functional form is the assumed constancy of the estimated relationship for all sizes of banks. More precisely, all banks in a particular sample are presumed to lie on the same average cost curve. While some studies estimate scale economies for only large banks and others estimate these economies for small and medium-sized institutions, few have systematically tested to see if all banks lie on the same curve, and therefore face the same technology. This hypothesis has been rejected statistically (Lawrence, 1989), likely due to the large samples which produce a very peaked sampling distribution. However, contrasts of published results for large and small banks separately suggest that scale economy values may not differ much in an economic sense. That is, the relatively flat U-shaped cost curves identified using all banks are replicated when only large banks are used separately (e.g., Noulas, Ray, and Miller, 1990). In either case, it is clear that on average the very largest banks do not appear to have a significant cost advantage due to scale economies compared to most smaller institutions.

VI.
Scale Economies at the Office or Banking Firm Level

When only bank-incurred costs are being minimized, scale economies for the average banking office and the average banking firm—both derived from equation (1)—should be the same. But when costs include both the production and the *delivery* of output to the customer, as occurs in banking, these two measures can differ. In effect banks minimize both bank and customer-incurred costs together, but only the bank portion is observed. Some banks will find it profitable to do more delivery—branching—than others. These banks will save customers' transportation and transaction costs (Nelson, 1985, Evanoff, 1988) but will add to bank costs, and so look to be less efficient compared to others which provide less delivery. As customer costs are unobserved, differences in delivery strategies can give the appearance of higher than minimum bank costs, even though profits may be maximized in either case. In this situation, scale economies can be measured at the office level (as seen in the results of Lawrence and Shay, 1986, who only measure office economies) while diseconomies can be measured at the firm level (as found in Hunter and Timme, 1986, and Berger, Hanweck, and Humphrey, 1987).

Some insight into resolving this difficulty, however, may be obtained by observing how banks behave when they have virtually no branches. Here the office *is* the firm. This is the result when scale economies are estimated for banks in unit banking states.[10] Scale diseconomies are regularly observed for the larger unit banks. Because these banks have (except in rare instances) no branch network to provide "convenience" to customers, these diseconomies must therefore be related to production inefficiencies alone, not to the extra expense of providing consumer convenience. In contrast, banks operating in branching states and hence providing customer convenience through a branching network have lower scale diseconomies at the firm level and slight economies at the office level (for all sizes of banks). Thus it appears that permitting a bank to branch will itself lower costs for the larger banks. The implication is that branching, far from being an extra cost of customer convenience, actually *lowers* both bank and customer costs. Branching permits a banking firm to lower costs by producing services in more optimally sized "plants" or offices rather than producing virtually all of the output at a single office, as occurs in unit banking states.[11] Thus the customer convenience aspect of branching would appear to be largely a side effect of a bank's desire to lower scale diseconomies by choosing a more optimal configuration of production facilities.

For banks in branching states, which in 1988 included all but Colorado, Illinois, Montana, and Wyoming, the average number of accounts per banking firm rises steadily with bank size, while the average number of accounts per office remains steady after a certain minimum is reached. This fact implies that branching banks can add output (deposits and loans) in either of two different ways: by adding additional offices in new market areas (which attract new accounts and balances) or by adding new accounts and balances to existing offices. The data indicate that the former method of output expansion, which includes internal growth as well as mergers, dominates the latter (Benston, Hanweck, and Humphrey, 1982, Table 1).

[10] Early on, published studies lumped banks in unit banking and branching states together. This is inappropriate since more recent studies have shown that these two classes of banks are significantly different from one another in terms of how costs vary with size. It should be noted that banks in unit banking states do at times have a limited number of branches while unit banks—those with no branches—exist in branching states.

[11] Two studies which contrast unit and branching bank scale economies are Benston, Hanweck, and Humphrey, 1982, and Berger, Hanweck, and Humphrey, 1987. Other studies generally parallel these results for banks in these two different regulatory environments.

To determine economies at the average banking office, the number of branches is included as an explanatory variable in equation (1) and scale economies at the office level are obtained from a partial derivative of the estimated total cost equation with respect to scale (or output) alone. For economies at the average banking firm, the same model is estimated but the total derivative of the equation with respect to both scale and number of branches is used. Equivalently, the variable measuring the number of branch offices can be deleted from (1) to obtain scale economies at the firm level. The results typically indicate that the average office still has some realizable scale economies whereas for the firm, these economies have either disappeared or have turned into slight diseconomies.

Researchers have in the past estimated scale economies for the average banking office and then conclude that large banks (banking firms) have lower costs. They do so without realizing there can be a difference between the office and firm results. In fact, most of the early studies of bank scale economies are deficient in this regard because they typically specified the number of branches as an independent variable in their estimating equation and then proceeded to derive scale economies as the partial derivative of costs with respect to output. But this derivation only gives scale economies when the number of banking offices is held constant and thus reflects only one of the two ways that bank output expansion can affect costs. A better approach is to compute scale economies both ways, and be clear about what concept is being measured, or to compute only those economies which apply to the banking firm as a whole—the relevant concept for policy purposes. That is, most policy issues in banking, whether relating to interstate banking, foreign bank competition, or bank costs faced by users, are a function of the relation between costs and firm size, not costs and the size of the average office. The prices of banking services necessarily reflect all banking costs, so the former, not the latter, is the appropriate point for scale economy evaluation.

VII.
Time Periods With High Versus Low Interest Rates

The time period chosen for a cross-section study of scale economies can affect the estimated slope of the average cost curve. The reason is that total costs—the appropriate cost concept to use when measuring scale economies—will vary over the interest rate cycle and alter the slope of the estimated cost curve.

Each of the three major components of average cost—purchased funds interest cost, core deposit interest cost, and the prices of factor inputs which comprise operating cost—are influenced by the interest rate cycle in cross-section data sets, but by differing amounts and with different lags. For example, average operating cost rises, with a lag, with the rate of inflation while the average cost of purchased funds rises immediately and fully reflects the level of market interest rates. In contrast, the average interest cost of core deposits almost always rises by less than the rise in market rates and usually with a lag. Since larger banks rely more on purchased funds, it is easy to see that large banks will necessarily have higher average costs than smaller banks when interest rates are high. This holds even if equal average costs would prevail across all banks when interest rates are at their "normal" level. Similarly, the reverse can hold if interest rates are at an exceptionally low level.[12]

Simply put, the slope of the average cost curve and estimates of bank scale economies can differ when they are based on single year cross-section data simply because the level of the market interest rate varies over time. Since the vast majority of scale economy estimates are in fact derived from single-year cross-section studies, interest rate variations can be an important consideration in explaining why some studies show more or less scale economies than others. Such variations are especially important when studies conducted in the 1960s and early 1970s, periods of relatively low interest rates, are contrasted with studies of the late 1970s and early 1980s, periods of unusually high rates. But even over 1980-84 when rates were high there was enough variation in the market interest rate to alter the slope of the average cost curve, shifting around the large banks so that small scale economies became small diseconomies (Humphrey, 1987, Figures 4a and 4b).

To abstract from this problem, time-series studies are needed since they can control for the year-to-

[12] If core deposits could be easily and rapidly substituted for purchased funds when market rates were relatively high, and vice versa when these rates were low, then the slope of the average cost curve would not be dependent on the interest rate cycle in the manner just described. But since such substitution is quite limited in practice, and because core deposits are typically treated as quasi-fixed inputs to the banking firm (Flannery, 1982), the effects of the interest rate cycle on cross-section scale economy estimation are operative.

year variation in the level of interest rates.[13] It turns out that those few time-series studies that do exist show constant costs for large banks—a flat average cost curve—when evaluated using the average interest rate over the sample period (Hunter and Timme, 1986). When a broader sample of banks are used over time, slight economies are measured for small banks (around .95) and slight diseconomies for the largest banks (around 1.05).[14] Overall, these time-series results are quite similar to many, but not all, of the studies that used cross-section data for a single year.[15] Thus, while the time period can affect the slope of the average cost curve and therefore the estimate of the associated scale economy, in practice the bias appears to have been relatively small. In any event, the safest course is to rely on generalizations of a number of single year cross-section results (as Mester, 1987, and Clark, 1988, have done) rather than generalize from only a single one. The close correspondence between many cross-section studies and the few time-series studies which exist supports this conclusion.

VIII. Single Versus Multiple Bank Outputs

Until quite recently, scale economy estimates were based on how costs varied with changes in a single, aggregate (stock) measure of bank output. That is, Q rather than the separate and different bank outputs (Q_i) that make up Q were specified in equation (1). A problem with this approach is that there are at least two quite different reasons why costs may vary with an aggregate measure of output and only one of them reflects scale economies. The other reflects economies of scope, or cost changes related to the number and joint production nature of the different outputs produced. Scope economies occur when costs fall as product mix is expanded, allowing fixed costs to be spread over a larger number of different outputs.

In single-output studies, there is the possibility that economies associated with output levels have been confounded with economies associated with joint production. One may avoid this problem by specifying a multiproduct estimating framework (using a number of different Q_is), rather than relying on an aggregate index of the different outputs (where Q is a weighted sum of the Q_is). In this way, the two separate influences—scale and scope—can be separated.[16]

A number of studies have tested the (functional separability) conditions needed to justify a single index of bank output and have rejected them statistically (Kim, 1986). Even so, as often happens, statistical rejection has not led to economic rejection: the scale economy results from single output studies are quite similar to those found in multiproduct analyses. That is, slight but significant economies are measured at the office level (.96 to .98) for all sizes of banks whereas the average cost curve describes a relatively flat U-shape at the level of the banking firm, this shape indicating significant economies at small banks (around .94) but significant diseconomies at the largest (around 1.06).[17] As a result, biases that could be due to commingling scope economies with scale economies appear in practice to be slight. Banks produce very similar product mixes, on average, so that the importance of measured scope economies using current *observed* production is apparently small enough not to bias the scale economy results obtained specifying single versus multiple outputs.[18] In sum, there are strong theoretical reasons to (1) reject studies of scale economies that have aggregated all bank outputs into a single index and (2) use an explicit multiproduct specification in its place. In practice, however, the overall

[13] Making the average interest rate an independent variable in equation (1) will control for the small variation in this rate across banks in a cross-section analysis but will not control for the bias introduced if the level of interest rates are atypically high or low for the time period studied.

[14] These results are from unpublished work by the author using a panel of almost 700 banks over 1977-88 that accounted for $2 trillion out the $3 trillion in total U.S. banking assets.

[15] A large number of cross-section studies are summarized in the comprehensive surveys of bank scale economies done by Mester, 1987, and Clark, 1988. Their conclusions are similar to those here in that scale economies seem to exist for small banks while constant costs or slight diseconomies are measured at the largest.

[16] Strictly speaking, the relationship between scale and scope economies is $S_{1,2} = (W S_1 + (1-W) S_2)/(1-S_c)$ where $S_{1,2}$ is the measure of overall economies of scale (in a two-output situation), S_1 and S_2 are the product-specific scale economies of the two outputs, S_c is the scope economy measure, and W is a weight which is similar to the share of variable costs in total cost for output 1 (See Bailey and Friedlaender, 1982, pp. 1031-32). Thus, the measure of overall economies of scale is related to scope economies in the usual aggregate (single) output situation. Even if S_1 and S_2 show constant costs, the overall scale measure ($S_{1,2}$) can falsely reflect economies or diseconomies depending on the value of scope economies (S_c).

[17] These results hold for both banks in unit banking and branching states, with the exception that the results noted in the text for the firm also apply to the average office in unit states (Berger and Humphrey, 1990).

[18] This result refers to the small expansion path subadditivity results in Hunter, Timme, and Yang, 1988, and Berger, Hanweck, and Humphrey, 1987. Scope economies are a special case of subadditivity and the complete specialization needed to reflect the scope concept is rarely seen in banking.

measure of scale economies is little affected by this adjustment.[19]

IX.
All Banks Are Efficient Versus Only Those on the Frontier

A final source of bias in the estimation of bank scale economies is the possibility that the economies exhibited by the set of most efficient or "best practice" banks can differ from those exhibited by all banks, efficient and inefficient. The potential for such bias exists because scale economies measured using all banks may be affected by other inefficiencies, unrelated to scale. These other factors would give a distorted picture of the true scale effects obtainable if all banks were as well managed and efficiently organized as those best practice banks with the lowest average costs.

This possibility arises because substantial cost differences, likely reflecting inefficiencies, seem to exist in banking (Humphrey, 1987). When all banks are stratified by size and then divided up into quartiles based on their levels of average costs for various years during the 1980s, the mean variation in average cost between the highest and lowest average cost quartiles of banks is 34 (31) percent for branching (unit) state banks. Since the mean variation in average cost across size classes was only 8 (12) percent, the variation between quartiles is seen to be 4 (2) times the variation across size classes. This pattern indicates that relative efficiency differences between similarly sized banks far exceed those obtainable by only altering bank size.[20]

To put these results differently, if a $500 million asset bank experienced a drop in its average cost from the mean of the highest to the mean of the lowest average cost quartile, costs would have fallen by 31 to 34 percent. Such a cost reduction would be equivalent to a scale economy value of .69 to .66. Since this figure far exceeds most estimates attributable to scale economies (e.g., .95), it is seen that even the existence of substantial scale economies at higher cost banks will *not* enable them to become competitive with smaller *or* larger banks that happen to be in the lowest cost quartile. Thus the competitive implications of scale economies at large banks are qualified by the existence of offsetting differences in cost levels or relative efficiency for all sizes of banks.[21]

Surprisingly, given the large differences in average costs between low- and high-cost banks, the scale economy results for banks in the lowest cost quartile (and therefore on the efficient cost frontier) are very similar to those obtained when all banks are pooled together (Berger and Humphrey, 1990). Thus while there are considerable differences in cost efficiency across banks, these differences do not significantly affect the scale economy results or conclusions of the previous section. Frontier analyses, which focus on low-cost or efficient banks, give the same results as the more traditional studies which estimate scale economies for all banks in a sample.

X.
Summary and Conclusions

There are important economic and political issues related to the size of scale economies in banking. Measurement of these economies is an empirical issue and, when many studies exist, it is possible to sort out the likely reasons for seemingly conflicting results. Such an understanding of the data and the results of different research designs permits the derivation of a consensus position useful for policy purposes.

Seven common differences in existing bank scale economy studies have been identified and discussed. These are summarized in Table I. Of the seven, only three (numbers 1, 3, and 4) led to problems sufficiently serious to warrant discounting the conclusions of studies incorporating them. Analyses which relate operating costs—not total costs—to variations in bank output contain a bias due to differences in the funding mix across banks. As a result, these analyses are typically biased toward finding scale economies when

[19] One benefit of a multiproduct specification, however, is that scale economies for each output can be determined separately and contrasted. The scope economy results derived from a multiproduct specification have, however, been disappointing as there has been a lack of consistency in the value of scope economies estimated. It has been shown that one reason for the markedly different scope economy results in different studies is a limitation in the translog functional form itself (virtually the only form used today in banking studies). When a form that better fits the data is used instead, consistent values for scope economies result regardless of the point of evaluation (Pulley and Humphrey, 1990).

[20] These differences are not due to chance occurrences of high or low costs among banks as they exist for the same banks during different time periods, when chance variations would be expected to average out. As well, low-cost banks consistently have higher profits (and vice versa). Thus whatever is happening on the cost side rolls over to the revenue side as well, rather than being the result of high-cost banks producing a different output which is offset by higher revenues (Berger and Humphrey, forthcoming).

[21] Similar conclusions apply to thrift institutions (Verbrugge, McNulty, and Rochester, 1990).

Table I

Summary of Differences Among Bank Scale Economy Studies

Common Differences:	Bias Found:
1. Cost Definition (operating versus total cost)	Use of operating cost gives bias toward finding scale economies.
2. Output measurement (number of accounts versus dollars in the accounts)	Either output measure gives similar results.
3. Functional form (linear versus quadratic)	Linear (Cobb-Douglas) form gives bias toward finding scale economies.
4. Point of scale economy evaluation (single office versus banking firm)	Evaluation for average banking office not relevant for policy purposes.
5. Time period used (high versus low interest rates)	Bias exists but is minor.
6. Commingling scale with scope (single versus multiple outputs)	Similar scale economy results with either single or multiple outputs.
7. Efficiency differences (average bank versus those on frontier)	No effect on scale economy results.

none may exist after proper account is taken of all costs associated with producing bank outputs. Thus, believable scale economy estimates should be based on models using total costs, not just operating costs. As well, a quadratic functional form such as the translog that permits a U-shaped cost curve to be estimated if it exists in the data, is always favored over a linear function such as the Cobb-Douglas. This eliminates the majority of the earlier studies in which the (log linear) Cobb-Douglas form was used and scale economies were regularly (mis)identified. Lastly, only those scale economies evaluated at the level of the banking firm are pertinent to the policy issues at hand since it is the size of the banking firm, not the size of the average office, which captures the full cost efficiency associated with the two ways that bank output can be expanded. While some problems are encountered in using different measures of bank output, selecting different time periods for estimation, commingling scale with scope economies, and pooling efficient with inefficient banks, the resulting scale estimates obtained in these four cases are reasonably robust to these different treatments.

Overall, a consensus conclusion of the preferred studies on bank scale economies suggests that the average cost curve in banking reflects a relatively flat U-shape at the firm level, with significant economies at small banks (around .94) but small and significant diseconomies at the largest (around 1.06). This relatively flat U-shape also holds even when large banks are viewed separately. The implication is that the slight diseconomies identified for all large banks together represents an average for some of the smaller large banks possessing economies and the very largest which seem to possess diseconomies.

From these results, some practical conclusions may be inferred. First, there would seem to be little benefit of a cost-reducing nature from a marked increase in bank size alone, although significant benefits from loan diversification would exist for giant nationwide banks. Second, the measured scale or cost economies are small in comparison to existing differences in cost levels between similarly sized banks. This finding implies that even if cost economies were pervasive, which they are not, they would have a much smaller competitive impact than has been heretofore presumed. The large and persistent cost differences between banks of a similar size and product mix suggest that greater competition within the banking industry would be beneficial but that this need not be associated with bank size. One way to enhance competition is to permit easier entry into and exit from the industry. A step in this direction will come with full interstate banking during the next decade when geographical restrictions on entry are to be removed.

References

Bailey, Elizabeth, and Ann Friedlaender, "Market Structure and Multiproduct Industries," *Journal of Economic Literature*, *20* (September 1982), 1024-48.

Bell, Frederick, and Neil Murphy, *Costs in Commercial Banking: A Quantitative Analysis of Bank Behavior and its Relation to Bank Regulation*, Research Report No. 41, Federal Reserve Bank of Boston, Boston, MA (1968).

Benston, George, "Branch Banking and Economies of Scale," *National Banking Review*, *2* (June 1965), 507-49.

————, "Economies of Scale of Financial Institutions," *Journal of Money, Credit, and Banking*, *4* (May 1972), 312-41.

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

Benston, George, Gerald Hanweck, and David Humphrey, "Scale Economies in Banking: A Restructuring and Reassessment," *Journal of Money, Credit, and Banking*, *14* (November 1982), 435-56.

Berg, Sigbjorn, Finn Forsund, and Eilev Jansen, "Technical Efficiency of Norwegian Banks: the Non-Parametric Approach to Efficiency Measurement," Working Paper, Research Department, Bank of Norway, Oslo, Norway (January 1990).

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

Berger, Allen, and David Humphrey, "The Dominance of Inefficiencies Over Scale and Product Mix Economies in Banking," Working Paper, Board of Governors of the Federal Reserve System (May 1990).

Berger, Allen, and David Humphrey, "Measurement and Efficiency Issues in Banking," *Output Measurement in the Services Sector*, Conference on Research in Income and Wealth, National Bureau of Economic Research, University of Chicago Press (forthcoming 1991).

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

————, *Report of Condition and Income*, Washington, D.C. (Various years).

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

Clark, Jeffrey, "Estimation of Economies of Scale in Banking Using a Generalized Functional Form," *Journal of Money, Credit, and Banking*, *16* (February 1984), 53-68.

————, "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.

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

Evanoff, Douglas, Philip Israilevich, and Randall 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 (June 1989).

Flannery, Mark, "Retail Bank Deposits as Quasi-Fixed Factors of Production," *American Economic Review*, *72* (June 1982), 527-36.

Gilligan, Thomas, Michael Smirlock, and William Marshall, "Scale and Scope Economies in the Multi-Product Banking Firm," *Journal of Monetary Economics*, *13* (May 1984), 393-405.

Greenbaum, Stuart, "A Study of Banking Costs," *National Banking Review*, *4* (June 1967), 415-34.

Heggestad, Arnold, and John Mingo, "On the Usefulness of Functional Cost Analysis Data," *Journal of Bank Research*, *8* (Winter 1978), 251-56.

Humphrey, David, "Cost Dispersion and the Measurement of Economies in Banking," Federal Reserve Bank of Richmond *Economic Review*, *73* (May/June 1987).

————, "Cost and Technical Change: Effects of Bank Deregulation," *Journal of Productivity Analysis*, (forthcoming 1991).

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

Hunter, William, Stephen Timme, and Won Yang, "An Examination of Cost Subadditivity and Multiproduct Production in Large U.S. Banks," Working Paper, Department of Finance, Georgia State University, Atlanta, GA (December 1988).

Kilbride, Bernard, Bill McDonald, and Robert Miller, "A Reexamination of Economies of Scale in Banking Using a Generalized Functional Form," *Journal of Money, Credit, and Banking*, *18* (November 1986), 519-26.

Kim, Moshe, "Banking Technology and the Existence of a Consistent Output Aggregate," *Journal of Monetary Economics*, *18* (September 1986), 181-95.

Langer, Martha, "Economies of Scale in Commercial Banking," Working Paper, Banking Studies Department, Federal Reserve Bank of New York (December 1980).

Lawrence, Colin, and Robert Shay, "Technology and Financial Intermediation in a Multiproduct Banking Firm: An Economic Study of U.S. Banks 1979-1982," in C. Lawrence and R. Shay (Editors), *Technical Innovation, Regulation and the Monetary Economy*, Ballinger, Boston, MA (1986), 53-92.

Lawrence, Colin, "Banking Costs, Generalized Functional Forms, and Estimation of Economies of Scale and Scope," *Journal of Money, Credit, and Banking, 21* (August 1989), 368-79.

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

Nelson, Richard, "Branching, Scale Economies, and Banking Costs," *Journal of Banking and Finance, 9* (June 1985), 177-91.

Noulas, Athanasios, Subhash Ray, and Stephen Miller, "Returns to Scale and Input Substitution for Large U.S. Banks," *Journal of Money, Credit, and Banking, 22,* (February 1990), 94-108.

Pulley, Lawrence, and David Humphrey, "Correcting the Instability of Bank Scope Economies from the Translog Model: A Composite Function Approach," Working Paper, College of William and Mary, Williamsburg, VA (June 1990).

Verbrugge, James, James McNulty, and David Rochester, "For Thrifts, Bigger Doesn't Necessarily Mean Better," *Journal of Retail Banking, 12* (Summer 1990), 23-33.