Assessing the impact of regulation on bank cost efficiency

Douglas D. Evanoff

Introduction and summary

The purpose of financial regulation is to improve upon the performance of financial markets relative to how they would perform driven solely by the forces of the private marketplace. For example, in the 1930s it was decided that, left unchecked, competition in the pricing of U.S. banking services could become so intense that it would actually be harmful to the functioning of the markets. This resulted in the introduction of interest rate and price restrictions to provide banks with an inexpensive source of funds. In addition, to insure that local market participants were not forced from the market as a result of "excessive" competition, entry barriers and branching restrictions were introduced.

Such regulation, however, frequently results in unintended behavior and market inefficiencies. The price restrictions aimed at providing an inexpensive source of bank funding resulted in disintermediation and significant bank expenditures to circumvent the restrictions. The entry barriers resulted in inferior service levels and the generation of local market power by incumbent institutions as competing service providers were unable to use an efficient entry mechanism. These unintended effects often prompt re-regulation to realize the original intent of the regulation, but without the resulting inefficiencies. However, re-regulation typically results in additional responses by bankers aimed at avoiding the effect of the regulation.

In responding to regulation, banks are altering the production process. The theoretical foundation for most bank cost studies is based on the maintained assumption of cost minimization with respect to market input prices in competitive markets.[1] However, extensive evidence suggests that this is not the behavior practiced by regulated firms. Regulated firms frequently alter the production process from what it would be absent the regulation. Banking firms are subject to extensive regulation in nearly all facets of operations, raising the possibility that the assumption of cost-minimizing behavior in response to market input prices may be particularly inappropriate for this industry.

Our objective is to evaluate whether industry regulations distort firm behavior and, as a result, generate productive inefficiency in the mix of inputs used by banks (for example, physical capital to labor ratios). We estimate this *allocative* inefficiency using a generalized cost function that allows for cost-minimization behavior, taking into account the above-mentioned distortions resulting from regulation. From a theoretical viewpoint, the generalized model is superior to the standard model. We test to see if there is also a statistical difference. We evaluate the impact of accounting for the regulatory distortions on various cost characteristics.[2] In addition to generating a measure of inefficiency resulting from banks using a suboptimal mix of inputs, we obtain a measure of the level of inefficiency resulting from the underutilization or mismanagement of inputs, that is, *technical inefficiency*. Finally, we analyze the effect of relaxing the regulatory constraints.

For a sample of large U.S. banks, we find statistically significant input price distortions, and resulting allocative inefficiency, which we attribute to regulation. We reject the standard cost model in favor of a more general one, which allows for cost minimization

Douglas D. Evanoff is a vice president and senior financial economist at the Federal Reserve Bank of Chicago. *The author acknowledges helpful comments on earlier drafts from Herb Baer, Dave Humphrey, William C. Hunter, David Marshall, Larry Mote, and Rasoul Rezvanian. The analysis presented here resulted from earlier work coauthored with Philip Israilevich. Excellent data and research assistance was provided by Betsy Dale, Velma Davis, Scott Johnson, Peter Schneider, and Gary Sutkin.*

subject to *effective* input prices that can differ from market prices as a result of regulation. Findings from our analysis of the 1972–87 period suggest that for our sample of banks, scope economies and minor scale economies existed. Scope economies exist if the cost of joint production is less than the cost resulting from independent production processes; scale economies exist if, over a given range of output, per unit costs decline as output increases. In addition, technology played a significant role in reducing costs, and regulatory induced allocative inefficiency existed. Although statistically significant, the allocative efficiency distortions appear to be relatively minor. The advantages of the generalized cost model become apparent, however, when we compare the 1972–79 period, one of significant regulation, with the 1984–87 period, which is considered the deregulated environment. Our findings suggest that the banking environment changed significantly between these two periods. Allocative inefficiency was a factor in 1972–79, but was nearly nonexistent in the later period. Banks apparently responded to the deregulated environment by altering their production process to fully exploit scale economies, and reaped significant returns from technological change. We conclude that the heavy regulation of the earlier period had a significant adverse effect on bank efficiency.

Productive efficiency: The basics

Basic economic theory assumes that production occurs in an environment in which an attempt is made to maximize profits by operating in the most efficient manner possible. The competitive model suggests that firms that fail to do so will be driven from the market by more efficient ones. These competitive forces generate an industry of firms producing efficiently with respect to the scale and scope of operation and the mix and quantity of inputs used. However, when market imperfections weaken competitive forces, inefficient firms may continue to prosper. True firm behavior may vary from that implied by the competitive model. Firms may find they are not required to operate as efficiently as possible because they are protected from the discipline of the market by either natural or regulatory forces. Inefficiencies can then arise and the characteristics associated with the competitive model (efficient scale, scope, and input utilization) no longer hold.

Variations from productive efficiency can be broken down into input- and output-induced inefficiencies. Assuming a given level of output, input inefficiency implies the firm is not optimally using the factors of production. That is, the given level of output is not produced at the lowest possible cost. Output efficiency requires the production of both the optimal level and the optimal mix of outputs.

Overall input inefficiency resulting from the suboptimal use of inputs can be divided into *allocative* and *pure technical* inefficiency. Allocative inefficiency occurs when inputs are combined in suboptimal proportions. Regulation is typically given as a major reason for this. An extreme example would be if regulations mandated that regulated firms use a particular process to produce a commodity. For example, no machinery can be used. Even if the inputs other than capital were used as effectively as possible, the ban on machinery would most likely result in a production process that would be less efficient than the unrestricted process.

Pure technical inefficiency occurs when more of each input is used than should be required to produce a given level of output. This type of inefficiency is more difficult to explain, but it is typically attributed to weak competitive forces that allow inefficient firms to remain in the market despite their inferior productivity. Pure technical inefficiency implies that firms employ the proper mix of inputs, but mismanage them. Combining allocative and pure technical inefficiency, we get the overall inefficiency resulting from the improper use of inputs. The distinction between the two types of inefficiency is important because they may be caused by different forces and, therefore, be correctable by different means. For example, the explicit repeal of regulations may result in an increase in allocative efficiency, while a general increase in the level of competition permitted (perhaps through reductions in entry barriers) may increase pure technical efficiency.

Productive efficiency also requires optimizing behavior with respect to outputs. Here, optimal behavior necessitates production of the level and combination of outputs that correspond to the production process with the lowest per unit cost. An optimal output level is possible if economies and diseconomies of scale exist at different output levels. Economies of scale exist if, over a given range of output, per unit costs decline as output increases. Increases in per unit cost correspond to decreasing returns to scale. A scale-efficient firm will produce where there are constant returns to scale; that is, changes in output result in proportional changes in costs. Many recent bank mergers have been justified on the basis of potential scale economies realized by the new combined entity.[3] Because it involves the choice of an inefficient level, scale inefficiency is considered a form of technical inefficiency. Thus, total technical inefficiency includes both pure technical and scale inefficiency, or inefficient levels of both inputs and outputs.

Additional cost advantages may result from producing more than one product. For example, a firm may be able to jointly produce two or more outputs more cheaply than producing them separately. If the cost of joint production is less than the cost resulting from independent production processes, economies of scope exist. Diseconomies of scope exist if the joint production costs are actually higher than the cost of specialized or stand-alone production of the individual products. In banking, potential scope economies are typically precluded by regulatory limitations on bank activities.

Finally, pure technical inefficiency is entirely under the control of, and results directly from the behavior of, the producer, whereas output inefficiency and allocative inefficiency may be unavoidable from the firm's perspective. For example, a firm optimally using inputs may find that per unit cost declines over the entire range of market demand. While increasing production would generate cost savings or efficiencies, the characteristics of market demand may not justify it. Failure to exploit scope advantages may also result from factors outside of the control of the firm. Clearly, in banking the array of allowable activities is constrained by regulation. This may preclude potential gains from the joint production of various financial services. Further, as mentioned earlier, allocative inefficiency may occur as a direct result of regulation. For example, during the 1970s, banks were restricted with respect to the explicit interest rates they could pay depositors. As market rates rose above allowable levels, banks frequently substituted implicit interest payments in the form of non-price payments or improved service levels—for example, a free toaster with the opening of a new savings account, or more offices per capita or per area.[4] This resulted in an overutilization of physical capital relative to other inputs. In this case, regulation was the driving force behind the resulting allocative inefficiency.

Generalized model of bank costs

To generate our cost and efficiency estimates, we use a methodology developed by Lau and Yotopoulos (1971) and Atkinson and Halvorsen (1980). This *shadow price* model has been employed in previous studies to account for regulatory-induced market distortions, for example, Atkinson and Halvorsen (1984), and Evanoff, Israilevich, and Merris (1989). In this model, firms optimize with respect to the shadow price or effective price of inputs, which includes any non-price aspects such as regulatory burden.

We apply the more general shadow price (SP) model with additional variables specific to banking. (See the studies listed above for a detailed discussion of the methodology and the technical appendix for a summary of the formal derivation of the cost relationships.)

From basic microeconomics, the condition required for optimization behavior in the standard cost model is for the firm to produce at the level where the ratio of the marginal products of the inputs employed[5] (that is, the ratio of the changes in output associated with marginal changes in inputs) is equal to the ratio of the prices of the inputs.

$$1)\quad \frac{f_i}{f_j} = \frac{P_i}{P_j} \text{ for } i \neq j = 1, \ldots, m.$$

In equation 1, f_i denotes the marginal product of input i, and P_i is the price of input i. Given that the firm takes input prices and the level of output as given, it can then derive the optimal combination of inputs to minimize costs.[6]

The standard model typically assumes that the optimal combination of inputs is determined by prices observed in the marketplace. Therefore, the observed and optimal costs are equivalent. However, if additional constraints exist, such as those imposed by regulation, the true cost of the input need not equal the observed price. There may be non-price costs induced by regulations, and these will also be accounted for in the firm's optimization process. As discussed earlier, when deposit rate ceilings were imposed, banks were limited in their ability to compete directly for funds. The banks then used non-price competition in an attempt to elude the restriction.[7] One result of this in the 1970s was the significant proliferation of bank offices in states allowing broad branching as banks attempted to "compete with brick and mortar." The decision to introduce more branch offices was not driven entirely by the market price of physical capital. More physical capital was used than would have been suggested by the market price alone, because the perceived return on these capital expenditures differed from that implied by the market prices. In other words, the effective price of physical capital was less than the market price. In determining the true effective price of inputs, these additional regulatory constraints must be taken into account. This possibility is captured by the more generalized cost model presented in the technical appendix.

The input combination generating cost minimization, therefore, equates the ratio of marginal products to the ratio of the *effective* prices of the inputs, including the non-price costs. It is these effective or shadow prices that are influenced by regulation and drive behavior. In the technical appendix, we show how one can derive the bank's cost function using shadow prices rather than observed prices. The resulting shadow cost function is a more comprehensive representation of costs to be minimized and is the appropriate representation of the production process. In the absence of binding regulatory constraints, shadow and actual prices are equal and the shadow model reduces to the standard cost model. However, if market and shadow prices are not equivalent (likely to be the case in a heavily regulated industry like banking), one needs to account for the additional regulatory constraints.

The shadow prices of bank inputs are not directly observable. Therefore, we assume that the shadow prices are proportional to market prices:

$$2)\quad P_i^* = k_i P_i \text{ for } i = 1, \ldots, m,$$

where P_i is the price for input i, k_i is a measure of the extent of the factor price distortion, and there are m inputs used in the production process.[8] If regulation is nonbinding, all shadow prices equal the respective market prices, $k_i = 1$ for all i, and the shadow cost function reduces to the standard formulation.

In applying the generalized model to large U.S. banks, we make certain assumptions concerning the bank production process. We include variables generally thought to affect costs in the banking industry, including measures for the number of bank offices, the holding company structure, and the role of technological change. We assume that banks produce four outputs: the dollar value of commercial and industrial loans, installment loans, real estate loans, and investment securities. Banks produce these outputs using labor, physical capital, and financial funding.[9] More details of the empirical specification are given in the technical appendix.

We estimated the model for the years 1972–87 for the largest banks in the U.S. that were members of a holding company over the entire period. This is a period during which regulation of the industry was evolving as certain restrictions became quite binding and industry participants were arguing for regulatory relief. The final data set consists of 164 banks and 2,624 observations. Our expectation was that these institutions were probably in the best position to avoid adverse effects from regulation, thus making our findings conservative. Inefficiency could be less for these institutions than for smaller banks, because they may have more astute management; be more cost conscious; and be more involved in wholesale banking, whereas most regulations concentrate on the retail side of banking.[10]

The Bank Call Report was the major data source. Costs, defined as the sum of expenditures on labor, funds, and physical capital, the number of banking offices, and the type of bank holding company organization (that is, single or multibank) were obtained from Call Report data. We used a time trend to account for technical progress. We assigned state level wage trends for each year to each bank according to the location of its home office. We approximated the price of physical capital, P_K, from Call Report data as the ratio of physical capital expenditures, measured as additions to plant and equipment, furniture, and physical premises, to the book value of net bank premises, furniture, and physical equipment. We also calculated the price of funds from Call Report data as an average cost of funds. We obtained the input price for labor, P_L, from the Bureau of Labor Statistics.

Empirical results

We used the standard market price (MP) model and the more general SP model to estimate costs for the sample of U.S. banks.[11] We find that the standard model can be statistically rejected in favor of the more general SP model. Our cost estimates suggest that observed input prices differ from effective prices. As expected, we find that the price of physical capital was distorted downward relative to that of both labor and financial funding, suggesting that the regulatory-induced production constraints are binding. In particular, the cost of capital relative to labor is only 58.8 percent of what it would be in the absence of regulatory distortions. We also find that the cost of funds is biased downward slightly relative to that of labor.[12] The cost of funds relative to labor is 97.6 percent of what it would be absent distortions.

Table 1 shows a number of production characteristics and additional comparisons between the standard and generalized models. The calculated scale elasticity measure suggests the existence of economies of scale that are significant in a statistical sense. In particular, according to the first row in table 1, a 1 percent increase in the scale of output increases costs by only 0.981 percent. The results suggest a "U" shaped average cost relationship (the scale elasticity measure equaling a value of 1.0 at the minimum value of the average cost relationship), with 58 percent of the observations falling in the range in which statistically significant scale economies exist and 35 percent falling in the range of significant diseconomies. We also find significant scope economies for the two broad categories of outputs analyzed—loans and investment securities. Specifically, the second row of table 1 indicates that when loans and securities are

TABLE 1

Shadow and market price models, statistical results

Cost characteristic	Shadow price model	Market price model
Scale elasticity	0.981 (.0033)	.983 (.0033)
Scope economies	.280	.282
Technical change	–0.076 (.0033)	–0.069 (.0034)
Allocative efficiency	0.01	–

Technical notes: The scale elasticity measure is $\frac{\partial \ln C}{\partial \ln Q_i}$, where C and Q denote costs and output, respectively, with values less than 1.0 indicating potential per unit cost savings from increased output. These scale elasticities are computed by evaluating the model at the mean of the sample.

The scope economies measure is $\frac{\Sigma_i C(Q_i) - C(Q)}{C(Q)}$, where C and C_i denote the cost of joint production of the outputs and the cost of producing i on a stand-alone basis, respectively. Scope economies show the extent to which costs are lower (for example, 28 percent) as a result of jointly producing the outputs. For this measure, two output categories were considered—loans and investment securities.

The technical change measure, $\frac{\partial C}{\partial T}$ captures how much cost changed (for example, 7 percent) per year over the period analyzed as a result of technical change.

The allocative efficiency measure captures how much costs could be decreased if the inefficiency were eliminated.

Standard errors of the estimates are presented in parentheses. A translog function was used to model the cost structure. Therefore, when generating the scope measure, zero output values were replaced with small values to avoid arithmetic errors.

Source: Evanoff and Israilevich (1990).

produced jointly, costs are 28 percent lower than when they are produced separately. That is, there are cost benefits from jointly producing the two categories of output.[13]

We find the role of technological change to be significant, suggesting that technical advances over the period, proxied with a time trend, significantly aided the production process.[14] In particular, the general model implies that the cost of a given level of output decreased at a rate of 7.6 percent per year. The technological advances also resulted in changes in the production process by altering the mix of inputs used. Banking firms began economizing on labor relative to the other inputs (physical capital and financial funding). Additionally, technical advances tended to flatten the average cost curve, that is, to decrease the advantages or disadvantages resulting from the scale of operation. Finally, the more restricted standard cost model (which ignores regulatory distortions) understates the rate of technical progress. That model implies the cost advantages resulting from technical progress were approximately 10 percent less than those found with the more general cost model.[15] Thus, in addition to finding differences in market and effective input prices that will tend to alter banks' input use decisions, the generalized model finds differences in other characteristics of the bank production process.

Bank efficiency

We evaluate the extent of allocative inefficiency resulting from regulatory restrictions by deriving the difference between shadow costs assuming no regulatory distortions (k_i = 1.0) and shadow costs assuming the estimated factor price distortions, that is,

$$3)\quad I_A = \ln C^S(\hat{k}_i) - \ln C^S(k_i = 1)\forall i,$$

where I_A is allocative inefficiency and C^S is the shadow cost relationship. (The statistic I_A is displayed in the final row of table 1.)

Although our estimates suggest the perceived price of capital is distorted downward, we find the resulting inefficiency from regulatory distortions to be relatively small. According to the final row of table 1, the cost of distortions is less than 1 percent of total costs.[16] That is, on average, allocative distortions resulted in costs being approximately 1 percent higher than they otherwise would have been. This finding of limited allocative inefficiency is somewhat similar to the findings of previous studies, for example, Berger and Humphrey (1990) and Aly et al. (1990). Only Ferrier and Lovell (1990) found significant allocative effects. Their analysis, however, combines different types of financial institutions (credit unions, savings and loans, and commercial banks) and may be influenced by data measurement problems. They also find labor to be overused relative to capital, which is precisely the opposite of what we have argued should occur as a result of regulation.

We also analyze the extent of pure technical inefficiency for the sample of banks, investigating whether banks overutilize *all* inputs once the optimal *combination* of inputs is determined. We do this by comparing the estimated cost structure to the best practice cost structure, or the *cost frontier*. To find the cost frontier, or the level at which firms would be operating if there was no pure technical inefficiency, we use an approach developed by Berger and Humphrey (1990), which compares the efficiency levels of high- and low-cost banking firms. We arrange the data in quartiles according to total cost per dollar of output and separately estimate SP models for the high- and low-cost banks. We then compare the costs of the average bank in the two groups, holding factor prices and market characteristics constant.[17] We find that technical inefficiency accounts for approximately 21 percent of costs. That is, elimination of this inefficiency could decrease costs by 21 percent. This effect is slightly smaller than that found in previous studies, but the difference may be due to sample structure—as stated earlier, one would expect the large banks in this sample to operate more efficiently.

Although we reject the more restrictive model relative to the more general one, our findings suggest that the biases induced by the misspecification are relatively minor for most cost characteristics. The exception is the measure of technical progress, which is understated by approximately 10 percent in the standard model.

Comparison of regulatory periods

Given the relatively low level of allocative inefficiency, one is tempted to say that regulatory distortions were minor over the period studied. This would make arguments for industry deregulation less persuasive, since the constraints are not shown to distort behavior appreciably. Additionally, in spite of the statistical significance of the differences found using the two models, one may question the net benefits of the SP specification because biases from the MP model appear relatively minor. While our results can be interpreted as representing the average distortion over the 17-year period, regulatory stringency was not constant over this period. For example, the 1980 Depository Institution Deregulation and Monetary Control Act and the 1982 Garn-St Germain Act relaxed constraints on industry prices, products, and geographic expansion—each considered a significant industry restriction, for example, see Evanoff (1985). Other studies have found that deregulation in the early 1980s did affect firm behavior, for example, LeCompe and Smith (1990). Below, we account for the changing regulatory environment and evaluate whether industry productive behavior varied over the period.

To account for the influence of industry deregulation, we divide the 17 years into the following three periods: 1972–79, characterized by significant regulation; 1984–87, considered the deregulated environment; and 1980–83, thought to be a period of adjustment in response to the newly relaxed restrictions. During the adjustment period, banks presumably adjusted their input mix. They may have closed offices previously opened as a substitute for explicit interest payments or altered their use of funds relative to the earlier period. We compare the productive behavior of large banks during the 1972–79 (restrictive) and 1984–87 (less restrictive) regulatory environments by separately estimating the SP model for the two periods. Table 2 presents a comparison of the resulting cost function characteristics.[18] We find substantial differences for the two periods. As expected, the price distortions and resulting inefficiency are significantly greater for the more restricted 1972–79 period than for the later period.[19]

We find that for the average bank in the sample, scale economies existed in the early period. According to the first row of table 2, the scale elasticity measure is significantly below one. However, the scale economies were fully exhausted after deregulation. One interpretation of this would be that the banks, faced with fewer production constraints and increased competition in the deregulated period, were able to alter their operations to capture the benefits from scale. That is, they could more effectively "grow their business" to exploit scale advantages, or they could take advantage of scale economies via mergers and acquisitions.

The findings concerning the role of technology are particularly interesting. Although technical change over the *entire* period was estimated to be approximately 7 percent (table 1), it appears that most of the cost savings were realized *after* deregulation. During the eight-year regulated period, technology decreased costs by only 5 percent, while over the significantly shorter deregulated period, it lowered costs by nearly 26 percent (see the third row of table 2).[20] What caused

TABLE 2

Cost characteristics for regulated and deregulated periods

Cost characteristic	1972–79	1984–87
Scale elasticity	0.981 (0.0045)	1.01 (0.0067)
Scope economies	.885	.891
Technical change	–0.050 (0.0045)	–0.258 (0.056)
Allocative efficiency	0.021	0.001
Observations	1,312	656

Note: See technical notes, table 1.
Source: Evanoff and Israilevich (1990).

the change? There is reason to believe it was a result of the deregulation.

Deregulation increased the banks' ability and incentives to take advantage of more efficient production techniques. We know that the technology was different in the two periods, because each period has a unique cost relationship. To evaluate how banks would have behaved in the later period with the old regulatory framework and technology (that is, the technology from the first period) still in place, we imposed the old technology on the data and recalculated technological change. We find that technology would have decreased costs in 1984–87 by approximately 9 percent, significantly *less* than the cost savings actually realized. Inefficiency would also have been significantly greater than that realized in the later period. In particular, the fourth row of table 2 shows that allocative inefficiency was greater than 2 percent in the earlier period, compared with 0.1 percent in the later period. It appears, therefore, that banks responded to deregulation by altering their production techniques to reap significant benefits from technology that could not be realized in the regulated environment. Finally, according to table 2, there was essentially no difference in economies of scope between the two periods. This is consistent with the results in table 1, where our estimates of scope economies were not affected by regulatory distortions.

Conclusion

We have analyzed costs for a sample of large banks, which may be more resilient than most banks to regulation. Nevertheless, we find statistically significant input price distortions, which appear to be due to regulatory constraints. We reject the standard market price model in favor of a more general one that allows for cost minimization subject to shadow factor prices, which can differ from market prices as a result of regulation. Our analysis incorporates the multiproduct production process and employs the intermediation approach to measuring bank output and costs—that is, banks serve as an intermediator of financial services. Findings from our analysis of the 1972–87 period suggest that for this sample of banks, scope economies and minor scale economies existed, technology played a significant role in reducing costs, and the standard market price model should be rejected relative to the more general shadow price model. However, for this time period, the distortions created by using the market price model appear relatively minor.

The advantages of the shadow price model relative to the market price model are highlighted in a comparison of the pre- and post-deregulated periods in banking. Our findings suggest that the banking environment changed significantly. Allocative inefficiency was a factor in the early time period, but was nearly nonexistent after deregulation. Banks apparently responded to the deregulated environment by altering their production process to fully exploit scale economies, and reaped significant returns from technological change. Scope advantages existed in each period.

We have evaluated the effect of regulation on the production process, particularly efficiency, of large commercial banks. The effect may be significantly different for alternative samples. Future studies of bank costs should consider the role of inefficiencies induced by regulation and determine whether the production process has changed over time. Our analysis suggests the change has been significant.

TECHNICAL APPENDIX

The generalized cost model

In the neoclassical cost model, firms are assumed to minimize costs in the Lagrangian-constrained cost function given by:

$$4) \quad L = P'X - \mu[f(X,Z) - Q],$$

where P and X are $(m \times 1)$ vectors of input prices and quantities, respectively; $f(X,Z)$ is a well-behaved neoclassical production function; Z is a vector of exogenous variables; Q is a vector of outputs; and μ is a Lagrangian multiplier. From the first-order conditions for cost minimization, the marginal rate of technical substitution between inputs i and j is equal to the ratio of prices of the two inputs. That is,

$$5) \quad \frac{f_i}{f_j} = \frac{P_i}{P_j} \text{ for } i \neq j = 1, \ldots, m,$$

where $f_i \equiv \partial f / \partial X_i$ is the marginal product of input i, and P_i is the price of input i. Given input prices, and the predetermined level of output as the only constraint, the optimal combination of inputs can be derived to minimize costs.

Now assume that additional regulatory constraints exist. The Lagrangian-constrained cost function to be minimized becomes:

$$6) \quad L = P'X - \mu[f(X,Z) - Q] - \sum_{h=1}^{n} \lambda_h R_h(P, X),$$

where R_h for (h = 1, ..., n) are constraints arising from regulation, and λ_h are Lagrangian multipliers. From the first-order conditions for cost minimization, the marginal rate of technical substitution between inputs *i* and *j* is equal to the ratio of effective prices of the two inputs. That is,

$$7)\quad \frac{f_i}{f_j} = \frac{P_i - \sum_{h=1}^{n} \lambda_h \partial R_h / \partial X_i}{P_j - \sum_{h=1}^{n} \lambda_h \partial R_h / \partial X_j}$$

$$= \frac{P_i^*}{P_j^*} \text{ for } i \neq j = 1, \ldots, m,$$

where P_i^* is the effective or *shadow price* of input *i*.

In the absence of binding regulatory constraints, equation 7 reduces to the neoclassical condition, whereby the marginal rate of technical substitution equals the ratio of *market prices* of inputs:

$$8)\quad \frac{f_i}{f_j} = \frac{P_i}{P_j} = \frac{P_i^*}{P_j^*} \text{ for } i \neq j = 1, \ldots, m.$$

This special case is nested within the more general shadow price relationship (equation 7).

Since the shadow prices of the inputs are not directly observable, following Lau and Yotopolous (1971) and Atkinson and Halvorsen (1980, 1984), the shadow prices are approximated by

$$9)\quad P_i^* = k_i P_i \text{ for } i = 1, \ldots, m,$$

where k_i is an input-specific factor of proportionality. As noted by Atkinson and Halvorsen (1980, 1984), the shadow price approximations can be interpreted as first-order Taylor's series expansions of arbitrary shadow price functions. When regulation is nonbinding, all shadow prices equal the respective market prices, $k_i = 1$ for all *i*, and the shadow cost function reduces to the more restricted function.

Differing from the restrictive function only in the input-price variables, the shadow cost function is given by

$$10)\quad C^S = C^S(kP, Q, Z),$$

where kP is a vector of shadow prices of inputs. Applying Shephard's Lemma, the set of derived input demand functions is

$$11)\quad X_i = \frac{\partial C^S}{\partial (k_i P_i)}.$$

Using equation 11, the firm's total *actual* cost is

$$12)\quad C^A = P'X = \sum_{i=1}^{m} P_i \frac{\partial C^S}{\partial (k_i P_i)}.$$

The shadow factor cost shares are obtained by logarithmic differentiation of C^S:

$$13)\quad M_i^S = \frac{\partial \ln C_i^S}{\partial \ln (k_i P_i)} = \frac{k_i P_i X_i}{C^S} \text{ for } i = 1, \ldots, m.$$

Rearranging equation 13,

$$14)\quad X_i = \frac{M_i^S C^S}{k_i P_i} \text{ for } i = 1, \ldots, m,$$

and substituting equation 14 into equation 12 gives,

$$15)\quad C^A = C^S \sum_{i=1}^{m} \frac{M_i^S}{k_i}.$$

Taking logarithms,

$$16)\quad \ln C^A = \ln C^S + \ln \sum_{i=1}^{k} \frac{M_i^S}{k_i}.$$

Using equations 14 and 15, actual factor-cost shares can also be obtained,

$$17)\quad M_i^A = \frac{P_i X_i}{C^A} = \frac{M_i^S k_i^{-1}}{\sum_{i=1}^{m} M_i^S k_i^{-1}} \text{ for } i = 1, \ldots, m$$

Equations 16 and 17 comprise our model.

For estimation purposes, we specify the shadow cost function in translog form. Total shadow cost is specified to be linearly homogeneous in shadow prices. The level of k_i cannot be estimated, given that the equations for total actual cost and factor cost shares are homogeneous of degree zero in k_i. The shadow price factor for labor, k_L, is set equal to unity and the shadow price factors for the remaining inputs are estimated. Therefore, we test for relative price efficiency only, not absolute efficiency.

The total shadow cost function measure in translog form is

$$
\begin{aligned}
18)\ \ln C^S = {} & \alpha_0 + \sum_i \beta_{Qi} \ln Q_i \\
& + 0.5 \sum_i \sum_j \beta_{Q_i Q_j} (\ln Q_i \ln Q_j) \\
& + \sum_i \gamma_{iQ_j} \ln Q_j \ln(k_i P_i) + \sum_i \beta_i \ln(k_i P_i) \\
& + 0.5 \sum_i \sum_j \gamma_{ij} \ln(k_i P_i) \ln(k_j P_j) + \phi_T \ln T \\
& + 0.5 \phi_{TT} (\ln T)^2 + \sum_i \theta_{Q_i T} \ln Q_i \ln T \\
& + \sum_i \gamma_{iT} \ln(k_i P_i) \ln T + \beta_B \ln B + 0.5 \beta_{BB} (\ln B)^2 \\
& + \sum_i \theta_{Q_i B} \ln Q_i \ln B + \theta_{TB} \ln T \ln B \\
& + \sum_i \gamma_{iB} \ln(k_i P_i) \ln B + \beta_H H + \sum_i \theta_{HQ_i} H \ln Q_i \\
& + \theta_{HT} H \ln T + \theta_{HB} H \ln B + \sum_i \gamma_{iH} \ln(k_i P_i) H; \\
& \forall i, j = K, L, F, \text{and } Q_i, Q_j = \text{the four outputs;}
\end{aligned}
$$

where $\gamma_{ij} = \gamma_{ji}$.

Linear homogeneity in shadow prices implies the following adding-up restrictions on parameters:

$$
\begin{aligned}
19)\ & \sum_i \beta_i = 1 \text{ and } \sum_i \gamma_{iQ_j} = \sum_i \gamma_{iB} = \sum_i \gamma_{iT} = \sum_i \gamma_{iH} = 0 \\
& \sum_i \gamma_{ij} = 0; \\
& \forall i, j, \text{ and } Q_j.
\end{aligned}
$$

Shadow cost shares for the translog specification are derived by logarithmic differentiation of C^S in equation 18:

$$
\begin{aligned}
20)\ M_i^S & = \frac{\partial \ln C^S}{\partial \ln(k_i P_i)} \\
& = \beta_i + \sum_j \gamma_{iQ_j} \ln Q_j + \sum_j \gamma_{ij} \ln(k_j P_j) + \gamma_{iT} \ln T \\
& \quad + \gamma_{IB} \ln B + \gamma_{ih} H \\
& \quad \forall i, j, \text{ and } Q_j.
\end{aligned}
$$

From equations 16, 18, and 20, total actual (observed) costs are

$$
\begin{aligned}
21)\ \ln C^A = {} & \ln C^S + \ln(\sum_i [\beta_i + \sum_j \gamma_{ij} \ln(k_j P_j) + \sum_j \gamma_{iQ_j} \ln Q_j \\
& \sum_i \gamma_{iT} \ln T + \sum_i \gamma_{iB} \ln B + \sum_i \gamma_{iH} H] k_i^{-1}) \\
& \forall i, j, \text{ and } Q_j.
\end{aligned}
$$

Using equations 17 and 20, the actual (observed) cost shares are given by

$$
\begin{aligned}
22)\ M_i^A = {} & [\beta_i + \sum_j \gamma_{ij} \ln(k_j P_j) + \sum_j \gamma_{iQ_j} \ln Q_j + \gamma_{iT} \ln T \\
& + \gamma_{iB} \ln B + \gamma_{iH} H] k_i^{-1} / \sum_i [\beta_i + \sum_j \gamma_{ij} \ln(k_j P_j) \\
& + \sum_j \gamma_{iQ_j} \ln Q_j + \gamma_{iT} \ln T + \gamma_{iB} \ln B \\
& + \gamma_{iH} H] k_i^{-1} \\
& \forall i, j, \text{ and } Q_j.
\end{aligned}
$$

Equation 21 and two of the share equations 22, appended with classical additive disturbance terms, constitute the set of equations to be jointly estimated.[21] Cost estimates were derived using the iterated seemingly unrelated regression technique.

NOTES

[1]For a technical discussion, see Diewert (1974).

[2]We consider the impact on scale and scope economies and the role of technology.

[3]There is also significant disagreement on the existence of these economies; see Evanoff and Israilevich (1990). There has also been a common misinterpretation in the literature of precisely what constitutes scale efficiency, see Evanoff and Israilevich (1995). For an alternative analysis of the impact of regulation on bank efficiency, see DeYoung (1998) and accompanying articles. Other recent analyses of bank efficiency include Berger and Mester (1997) and Berger and Humphrey (1997).

[4]See Evanoff (1985) for further discussion of non-price competition.

[5]This ratio is the marginal rate of technical substitution between the inputs.

[6]That is, the predetermined level of output is the only constraint imposed on the firm.

[7]See Brewer (1988), Chase (1981), Lloyd-Davies (1975), Pyle (1974), and Startz (1979).

[8]Technically, these shadow price approximations can be interpreted as first-order Taylor's series expansions of arbitrary shadow price functions. It should be emphasized that we are testing for relative price efficiency (whether $k_l = k_k$) and not absolute efficiency whether all *k*s actually equal one.

[9]We are using an "intermediation approach" in defining bank outputs, that is, we measure output as the dollar value of produced assets and include the interest expense of funds in our measure of costs. This accounts for the most fundamental role of banks: to intermediate and transform liabilities into assets. This is in line with much of the recent bank cost literature, although an alternative "production approach" has been used by others when evaluating small commercial banks. For a discussion of the alternative approaches and their differences, see Berger, Hanweck, and Humphrey (1987).

[10]Rangan et al. (1988), Berger and Humphrey (1990), and Elyasiani and Mehdian (1990a) found large banks to be more efficient. Elyasiani and Mehdian attribute most of the differential to scale advantages, and Rangan et al. attribute it to pure technical efficiency differences. Neither study, however, tested for allocative efficiency. Using a nonparametric approach, Aly et al. (1990) did not find allocative efficiency to be related to bank size.

[11]Detailed estimates are available from the author upon request and are summarized in Evanoff and Israilevich (1990). The sample includes both unit and branch banks. This was done to preserve the attributes of the panel sample, as some states changed their restrictions on geographic expansion during the period studied. Analysis of a sample of branch banks produced similar results, albeit distortions of a smaller magnitude.

[12]With $k_L = 1.0$, the estimated factor price distortions were $k_{capital} = .588$ and $k_{funds} = .976$, with both estimates being statistically different from a value of 1.0.

[13]It is also most likely that this partially results from the substantial "fixed" costs. We should emphasize that cost complementarities and scope economies are not synonymous; scope is a broader concept. Additionally, estimates of scope economies should be interpreted cautiously, since these require evaluation of the cost function at values significantly distant from the sample. Empirically this has been shown to be a particular problem with the translog functional forms we have used here.

[14]Using similar data, an aggregate output measure, and production expenses only (excluding funding cost), an earlier study found a more significant influence from technology (Evanoff, Israilevich, and Merris, 1989). This suggests, as expected, that technical advances have aided the physical production process significantly more than the funds gathering process.

[15]In our empirical analysis, we simultaneously estimate a cost equation and input share equations. The finding of labor-saving technology for banks is derived from the input share equations.

[16]Again, contrasting these findings to those using an aggregate output measure and production costs suggests, as expected, that regulatory-induced inefficiencies affect the production process more than the funds collection process.

[17]See Berger and Humphrey (1990) for a complete description of the procedure. Our methodology differs slightly because we do not have to assume that the low-cost quartile firms are both technically and allocatively efficient. We can account for allocative inefficiencies by using the SP model. In theory, we believe this is preferred since even well managed (technically) efficient institutions can be adversely affected by regulation. However, quantitatively the difference may be small, given our finding of limited allocative inefficiency. Also, by using a panel data set, we do not encounter the problem of limited observations for the subsample of large banks. Detailed results from the estimates summarized here are available from the author on request.

[18]Further details are available from the author.

[19]Statistical tests indicated the two periods should be viewed separately.

[20]We also found that the effect of technology on input shares was significantly different between the two periods. While technology was funds-using in both periods, the effect was much larger in the deregulated period. Similarly, technology was significantly more capital-saving in the deregulated period, that is, when firms could compete directly via prices instead of employing alternative (capital-intensive) means to compete.

[21]One share equation is dropped because of the singularity of the variance-covariance matrix of the error terms for the three-equation system resulting from the adding-up conditions on the share equations. We arbitrarily drop the capital-share equation. The empirical results are invariant to the choice of share equation deleted and to the shadow price chosen for normalization.

REFERENCES

Aly, Hassan Y., Richard Grabowski, Carl Pasurka, and Nanda Rangan, 1990,"Technical, scale, and allocative efficiencies in U.S. banking: An empirical investigation," *Review of Economics and Statistics*, Vol. 72, May, pp. 211–218.

Atkinson, Scott E., and Robert Halvorsen, 1984, "Parametric efficiency tests, economies of scale, and U.S. electric power generation," *International Economic Review*, Vol. 25, October, pp. 647–662.

________, 1980, "A test of relative and absolute price efficiency in regulated utilities," *The Review of Economics and Statistics*, Vol. 62, February, pp. 185–196.

Bell, Fredrick W., and Neil B. Murphy, 1968, "Costs in commercial banking: A quantitative analysis of bank behavior and its relation to bank regulation," Federal Reserve Bank of Boston, research report, No. 41.

Benston, George J., 1965, "Branch banking and economies of scale," *Journal of Finance*, Vol. 20, May, pp. 312–331.

________, 1972, "Economies of scale of financial institutions," *Journal of Money, Credit, and Banking*, Vol. 4, May, pp. 312–341.

Berger, Allen N., and David B. Humphrey, 1997, "Efficiency of financial institutions: International survey and directions for future research," *European Journal of Operational Research*, Vol. 98, April, pp. 175–212.

________, 1991, "The dominance of inefficiencies over scale and product mix economies in banking," *Journal of Monetary Economics,* Vol. 28, pp. 117–148.

Berger, Allen N., and Loretta Mester, 1997, "Inside the black box: What explains differences in the efficiencies of financial institutions," *Journal of Banking and Finance*, Vol. 21, July, pp. 895–947.

Berger, Allen N., Gerald A. Hanweck, and David B. Humphrey, 1987, "Competitive viability in banking: Scale, scope, and product mix economies," *Journal of Monetary Economics*, Vol. 20, December, pp. 501–520.

Brewer, Elijah III, 1988, "The impact of deregulation on the true cost of savings deposits," *Journal of Economics and Business*, Vol. 40, November, pp. 79–98.

Chase, Kristine, 1981, "Interest rate deregulation, branching, and competition in the savings and loan industry," *Federal Home Loan Bank Board Journal*, Vol. 14, November, pp. 2–6.

Clark, Jeffrey A., 1988, "Economies of scale and scope at depository financial institutions: A review of the literature," *Economic Review*, Federal Reserve Bank of Kansas City, September/October, pp. 16–33.

DeYoung, Robert, 1998, "The efficiency of financial institutions: How does regulation matter?: Introduction," *Journal of Economics and Business*, Vol. 50, March/April, pp. 79–84.

Diewert, W. E., 1974, "Applications of duality theory." in *Frontiers in Quantitative Economics*, M. Intriligator and D. Kendricks (eds.), Vol. 2, Amsterdam: North-Holland.

Elyasiani, Elyas, and Seyed Mehdian, 1990a, "Efficiency in the commercial banking industry, a production frontier approach," *Applied Economics*, Vol. 22, April, pp. 539–551.

________, 1990b, "A nonparametric approach to measurement of efficiency and technological change: The case of large U.S. commercial banks," *Journal of Financial Services Research*, Vol. 4, July, pp. 157–168.

Evanoff, Douglas D., 1985, "Financial industry deregulation in the 1980s," *Economic Perspectives*, Federal Reserve Bank of Chicago, Vol. 9, No. 5, September/ October, pp. 3–5.

Evanoff, Douglas D., and Philip R. Israilevich, 1995, "Scale elasticity versus scale efficiency in banking," *Southern Economic Journal*, Vol. 61, April, pp. 1036–1046.

________, 1990, "Deregulation, cost economies, and allocative efficiency of large commercial banks," Federal Reserve Bank of Chicago, working paper, No. 90-19.

Evanoff, Douglas D., Philip R. Israilevich, and Randall C. Merris, 1990, "Relative price efficiency, technical change, and scale economies for large commercial banks," *Journal of Regulatory Economics*, Vol. 2, pp. 281–298.

Ferrier, Gary, and C. A. Knox Lovell, 1990, "Measuring cost efficiency in banking: Econometric and linear programing evidence," *Journal of Econometrics*, Vol. 46, October/November, pp. 229–245.

Gilligan, Thomas, and Michael Smirlock, 1984, "An empirical study of joint production and scale economies in commercial banking," *Journal of Banking and Finance*, Vol. 8, March, pp. 67–76.

Gilligan, Thomas, Michael Smirlock, and William Marshall, 1984, "Scale and scope economies in the multiproduct banking firm," *Journal of Monetary Economics*, Vol. 13, May, pp. 393–405.

Humphrey, David B., 1990, "Cost and technical change: Effects from bank deregulation," paper presented at the Franco-American Economic Seminar on Productivity Issues at the Micro Level, National Bureau of Economic Research Summer Institute, July 23–26.

________, 1987, "Cost dispersion and the measurement of economies in banking," *Economic Review*, Vol. 73, May/June, pp. 24–38.

Hunter, William C., and Stephen G. Timme, 1986, "Technical change, organization form, and the structure of bank production," *Journal of Money, Credit, and Banking*, Vol. 18, May, pp. 152–166.

Hunter, William C., Stephen G. Timme, and Won Keun Yang, 1990, "An examination of cost subadditivity and multiproduct production in large U.S. banks," *Journal of Money, Credit, and Banking*, Vol. 22, November, pp. 504–525.

Lau, Lawrence J., and Pan A. Yotopoulos, 1971, "A test for relative efficiency and application to Indian agriculture," *American Economic Review*, Vol. 61, March, pp. 94–109.

Lawrence, Colin, and Robert Shay, 1986, "Technology and financial intermediation in multiproduct banking firms: An econometric study of U.S. banks, 1979–1982," in *Technological Innovation, Regulation, and the Monetary Economy*, Colin Lawrence and Robert Shay (eds.), Cambridge: Ballinger Publishing Company.

LeCompte, Richard L. B., and Stephen D. Smith, 1990, "Changes in the cost of intermediation: The case of savings and loans," *Journal of Finance*, Vol. 45, September, pp. 1337–1346.

Lloyd-Davies, Peter, 1977, "The effect of deposit rate ceilings upon the density of bank locations," *Proceedings of a Conference on Bank Structure and Competition*, Chicago: Federal Reserve Bank of Chicago, pp. 1–18.

Noulas, Athanasios, Subhash Ray, and Stephen Miller, 1990, "Returns to scale and input substitution for large U.S. banks," *Journal of Money, Credit, and Banking*, Vol. 22, February, pp. 94–108.

Pyle, David H., 1974, "The losses on savings deposits from interest rate regulation," *Bell Journal of Economics and Management Science*, Vol. 5, Autumn, pp. 614–622.

Rangan, Nanda, Richard Grabowski, Hassan Y. Aly, and Carl Pasurka, 1988, "The technical efficiency of U.S. banks," *Economic Letters*, Vol. 28, pp. 169–175.

Shaffer, Sherrill, 1985, "Competition, economies of scale, and diversity of firm sizes," *Applied Economics*, Vol. 17, June, pp. 467–476.

Shaffer, Sherrill, and Edmond David, 1986, "Economies of superscale and interstate expansion," Federal Reserve Bank of New York, unpublished research paper, October.

Shephard, R. W., 1970, *Theory of Cost and Production Functions*, Princeton: Princeton University Press.

Startz, Richard, 1979 "Implicit interest on demand deposits," *Journal of Monetary Economics*, Vol. 5, October, pp. 515–534.

Toda, Yasushi, 1977, "Substitutability and price distortion in the demand for factors of production: An empirical estimation," *Applied Economics*, Vol. 9, September, pp. 203–217.