The Scandinavian Journal of Economics

Formerly The Swedish Journal of Economics

Vol. **94** 1992

Supplement

Productivity Concepts and Measurement Problems

Editors' Preface by *Ernst R. Berndt, Peter Englund* and *Lennart Hjalmarsson*

Introduction by *Robert M. Solow*

Part I. What is it We Want to Measure?

Articles by *Charles R. Hulten; Zvi Griliches; Dale W. Jorgenson* and *Barbara M. Fraumeni; Michael Intriligator; Jeffrey Bernstein; R. J. Colwell* and *E. P. Davis; Burton Weisbrod*

Part II. What Do the Measurements Indicate?

Articles by *Ernst R. Berndt* and *Bengt Hansson; Hans Bjurek, Urban Kjulin* and *Björn Gustafsson; Lennart Hjalmarsson* and *Ann Veiderpass; Sigbjørn Atle Berg, Finn R. Førsund* and *Eilev S. Jansen; Erik Mellander*

Part III. The Birth of a Discipline

Articles by *Lars Magnusson; Sven-Eric Liedman* and *Mats Persson*

The Scandinavian **Journal of Economics**

(formerly The Swedish Journal of Economics)

The journal is published with financial support from the Central Banks of Denmark, Finland, Iceland, Norway and Sweden.

Ansvarig utgivare

Lennart Hjalmarsson

The Scandinavian Journal of Economics is published four times a year, in March, June, September and December by **Blackwell Publishers**, 108 Cowley Road, Oxford OX4 1JF or 238 Main Street, Suite 501, Cambridge, MA 02142, USA.

Information for Subscribers. New orders and sample copy requests should be addressed to the Journals Marketing Manager at the publisher's address above. Renewals, claims and all other correspondence relating to subscriptions should be addressed to the Journals Subscriptions Department, Marston Book Services, PO Box 87, Oxford OX2 0DT. Cheques should be made payable to Basil Blackwell Ltd.

Subscription prices 1992	UK/Europe	USA & Canada	Overseas (incl Eire)
Institutions	£84.50	US$159.00	£94.00
Individuals	£48.00	US$91.50	£54.00

US Mailing. Second class postage paid at Rahway, New Jersey. Postmaster: send address corrections to The Scandinavian Journal of Economics, c/o Mercury Airfreight International Ltd Inc., 2323 E-F Randolph Avenue, Avenel, NJ 07001.

Back issues. Single issues from the current and previous two volumes are available from Marston Book Services at the current single issue price. Earlier issues may be obtained from Swets & Zeitlinger, Back Sets, Heereweg 347, PO Box 810, 2160 SZ Lisse, Holland.

Manuscripts, books for review and editorial correspondence. See inside back cover.

Microform. The journal is available on microfilm (16 mm or 35 mm) or 105 mm microfiche from the Serials Acquisitions Department, University Microfilms Inc., 300 North Zeeb Road, Ann Arbor, MI 48106, USA.

Advertising. For details contact the Advertising Manager, Bill Drake, telephone (0869) 38477 or write c/o the publishers.

Books for review. Any books for review should be sent to the Editorial Office (see inside back cover).

ISSN 0347-0520

Printed and bound in Great Britain by Page Bros, Norwich

Scand. J. of Economics 94, Supplement, 111–129, 1992

Output and Productivity in Banking

*R. J. Colwell and E. P. Davis**

Bank of England, London, England

Abstract

Concepts in banking output and the empirical literature on bank productivity — which employs output concepts — are critically surveyed. For output, the national accounts, production and intermediation approaches are compared. As regards productivity, both partial and total factor productivity measures, and the DEA and parametric approaches to the latter are assessed. Among the most striking results is the prevalence of technical inefficiency in banking. But it is also suggested that measurement techniques have often outpaced the theory of what is to be measured, notably in fields such as joint production, risk and competition. Alternative approaches to address these issues are suggested.

I. Introduction

Recent developments in financial markets such as deregulation, securitization, internationalization, credit expansion, financial instability and the generally growing importance of financial services in economic activity in the advanced countries have all put an increasingly sharp focus on the activities of banks. What do they produce? Are they efficient? Economists have exposed considerable difficulties in the definition and measurement of the concepts of bank output and productivity. For example, are demand deposits an input or output? Are banks' services best measured by number of accounts and transactions or value of accounts? Methodological issues are predominant in the analysis of productivity — should partial or total productivity be measured? If the latter, by parametric or non parametric methods? This paper seeks to provide a critical survey of the literature in these areas. The principal focus is on empirical studies of bank productivity (Section III), largely abstracting from the problem of efficient

*The views expressed are those of the authors and not necessarily those of the Bank. We thank J. Ganley, C. Goodhart, D. Miles, participants in seminars at the Bank of England, London School of Economics and Uppsala, and two referees for helpful suggestions. The errors remain our own responsibility.

scale[1] and focussing mainly on efficiency in use of inputs — allocative and technical efficiency. However, an assessment of conceptual issues relating to banking output — itself an essential background for the study of productivity — is also provided in Section II.

II. Bank Output Measurement

We begin by identifying conceptual problems regarding bank output and how it may be measured. As well as being of importance in itself, a measure of output is crucial to estimation of productivity. As is well known, measurement of output is problematic in all industries, due to problems such as aggregation and quality. But the output of *financial institutions* presents particular difficulties. As pointed out by Kinsella (1980), each bank is a multi-product firm (posing a problem of aggregation of outputs); many of its services are joint or interdependent — providing one service may entail providing others which cannot be separated or priced separately (for example safekeeping and accounting services in a current account) or which it is cheaper to produce together than separately (economies of scope); not all services are paid for directly (demand deposits) and banking is subject to government regulations that may affect costs, prices or level of output.

At a practical level, the obvious starting point in measuring the sector's output is to look at the way it is treated in the *national accounts*. These accounts seek to measure the value added by different sectors of the economy, reflected in turn in the profits and income from employment arising in each sector. Profits normally exclude interest (or net interest) receipts on the basis that the latter represent transfers of earnings from activities in other sectors. If interest payments only represented such transfers, there would not be a problem. But the "interest" received and paid by banks is in fact a combination of a charge for the use of capital and a charge for various services provided by these firms. The capital charge element nets out, at least when non-financial items in the balance sheet and the extent of any maturity transformation or risk absorption by financial intermediaries are taken into account. However, the exclusion of all interest received and paid leads to an understatement of financial firms' profits, insofar as the "concealed" charges in net interest receipts are also excluded from output (typically only explicit service charges are counted). The understatement is so large that trading profits for the sector, as recorded, are invariably negative. It also leads to an understatement, rather than simply a redistribution, of GDP to the extent that the "concealed"

[1] Most of the literature on bank efficiency focuses on scale economies; see the reviews in Gilbert (1984), Humphrey (1990) and Evanoff and Israilevich (1991).

charges reflect services provided to final rather than intermediate consumers. In looking at the share of the sector in GDP, therefore, it is conventional to include net interest receipts in its value added.[2] In the U.S., these are attributed to depositors; in the U.K., to both depositors and borrowers.

Most banking studies do not use national accounts measures, but instead have tended to adopt either the "production" or the "intermediation" approach.[3] According to the *production approach*, banks are treated as firms which use capital and labour to produce different categories of deposit and loan accounts. Outputs are measured by the number of these accounts or number of transactions carried out on each type of product, while total costs are all operating costs used to produce these outputs. On the other hand, in the *intermediation approach*, banks are viewed as intermediators of financial services rather than producers of loan and deposit account services, and the values of loans and investments are used as output measures; labour and capital are inputs to this process, hence operating costs plus interest costs are the relevant cost measure. Deposits may be either inputs or outputs.

The "intermediation approach" was first used in early cost studies. For example, Alhadeff (1954) measured output in terms of dollar values of earning assets (loans plus investments). The disadvantage of this measure is that other assets, such as trust operations are excluded, thus inflating the unit costs of larger banks. Schweiger and Mcgee (1961) and Gramley (1962) used total deposits and assets respectively to avoid this bias. However, all these studies used real-valued unweighted indexes, which ignore the differential importance of individual bank products, the relative cost of production and the ease with which banks can alter their product mix. This highlights the additional problem of how to account for the multi-product nature of bank activity. Furthermore, production is a "flow" concept expressed as some amount per unit of time, while the amount of assets and deposits are "stock" concepts representing given amounts at a particular point in time. Moreover, it ignores services not proxied by balance sheet magnitudes. To correct for some of these problems, weighted indexes have been used to measure output. A simple example would be Current Operating Revenue; however, Powers (1969) suggested it would be better to use a weighted bank output index, including in output a "charge" weight to each dollar of time deposits based on the difference between the Treasury Bill rate and the time deposit rate, to allow for services provided by the bank in accepting time deposits. Both these

---

[2] Fixler and Zieschang (1991) suggest this measure can be rationalized in terms of a theory of financial firms grounded in a user cost of money concept.

[3] Kolari and Zardkoohi (1987) provide a detailed review of this literature.

weighted measures assume there is no market failure or other distortion (higher loan rates obtained by one bank may imply market power or greater management efficiency and not higher output). This problem had led Greenbaum (1967) to use linear regressions to derive a set of average interest rates charged on various categories or earning assets by a sample of banks. These average rates were used as weights. But his measure was still vulnerable to the criticism of ignoring the effect of inflation on interest rates (which provides an unjustifiable boost to this measure of bank output). Moreover, non-credit output is generally treated crudely in the intermediation approach.

Meanwhile, the "production approach" of measuring numbers of accounts and transactions per period was first introduced by Benston (1965). This method meets some of the problems of the intermediation approach. For example, it removes the inflation bias and is a flow concept. It also allows numbers of accounts and average size of accounts to have differential effects on costs. But this approach suffers from lack of a method of weighting of the contribution of each service to total output (especially given interdependence) and omits many important items of bank services. Later work by Benston *et al.* (1982) weighted numbers of accounts in each activity area by proportionate shares in total operating costs using a Divisia Index, with a separate control provided by including the average size of accounts. The method is still vulnerable to the criticism of ignoring interest costs, which constitute a substantial proportion of banks' total costs. Omission is of particular importance if there is a tradeoff of higher operating costs (e.g. by operating many branches) against interest costs (because of greater locational convenience). In more recent studies, the production approach has only been used by studies focusing on the relative efficiency of branches within a particular bank, rather than across banks. Moreover, these studies have used the "number of transactions" rather than "number of accounts" on the basis that an account may be opened at one branch but transactions on the account may be processed at other branches.[4] Besides intrinsic difficulties, the fact that the "production approach" has not been used for inter-bank productivity studies reflects the difficulties encountered in collating accurate data.[5]

Given these data limitations, the latest bank productivity studies have adopted the "intermediation approach". More specifically, Elyasiani and

---

[4] For instance, Sherman and Gold's (1985) study of a U.S. savings bank measured output as a weighted average of the 17 services most commonly offered by the branches; while Vassiloglou & Giolias (1990) took into consideration the complete range of 72 transactions offered by the Commercial Bank of Greece. Similarly, Tulkens (1990) aggregated 60 operations into eight categories in his assessment of a public Belgian bank.

[5] Comprehensive data are only available for the U.S., and even this has questionable features; see Elyasiani and Mehdian (1990a).

Mehdian (1990a and b) followed Mester (1987) and the early studies outlined above, in assuming that output should be measured as the dollar value of a banks' earning assets; whilst deposits, in addition to labour and capital, should be treated as inputs in the production of assets. In contrast, Field (1990) took a similar view to Powers (1969), in regarding deposits not as an input but as an additional product over which banks compete. Hence he chose to measure output as the value of loans and deposits. Other studies have refined this approach by making distinctions between different types of deposits. For instance, Rangan *et al.* (1988) considered demand, time and savings deposits as outputs, whilst purchased funds such as large CDs, notes and debentures were regarded as inputs. Similarly, Berger and Humphrey (1990) treated produced deposits (demand, retail time and savings accounts) as outputs, but considered purchased funds (federal funds, large CDs and foreign deposits) to be inputs. They explained that this differentiation is necessary because the latter are not highly resource consuming. More recently, Berg (1991) and Berg and Kim (1991) have argued that since purchased funds do not use real resources they do not even qualify as an input. Meanwhile, Berger and Humphrey (1990) suggested that before one of these input/output approaches is adopted, there should be careful consideration of the banking functions seen as most important to the study in hand. They outlined three approaches to this initial identification process, an asset approach (outputs are bank loans and other assets), a user cost approach (outputs contribute to net revenue) and a value added approach (outputs contribute to value added).

To summarize, therefore, three approaches have been distinguished. However, national income measures are little used in the academic literature; and at present the "intermediation approach" appears to be preferred to the "production approach" in inter-bank studies. In the light of Berg, Førsund and Jansen (1989), the choice between these two approaches needs to be carefully considered, since their study of the Norwegian banking market in 1985 found that the number and ranking of efficient banks varies significantly depending on which output measurement is used.

We suggest that current output measures also suffer from various omissions. *Risk* is an additional feature of bank loans, but variations in it are not taken into account in most output measures; a bank may be able to boost output in terms of the balance sheet or profits by increasing risk. No account is taken of diversification. Perhaps it might be more appropriate to use some *ex-post* revenue measure, covering losses over the cycle, with provisions as negative output. Alternatively, as suggested by Charnes *et al.* (1990) provisions and actual loan losses could be counted as inputs. More generally, none of the identified measures of output seem to reflect the

quality of bank services of which risk (of failure) is only one dimension. Other aspects include liquidity and security for deposits; maturity, covenants and secured status for loans. At least some of these can be objectively measured, perhaps using "Hedonic price indices"; see Shaffer and David (1991) for an attempt to measure economies of scale using such techniques.

Other difficulties — whose measurement is more problematic — include the fact that the various measures do not allow for intertemporal *relationships* that are crucial in banking, hence rather than being only an implicit indicator of services provided, the interest rate might indicate an investment by the bank in a long-term relationship; there may be biases to measures of output when *competition* increases;[6] output measures take no account of the importance of *monitoring*, held by theorists to be central to banking; see Diamond (1984). Finally, private and social measures of output may differ due to *externalities*.[7]

III. Productivity

Having outlined the main issues concerning the measurement of bank output, we now go on to discuss work on productivity in banking (which uses the concepts of output as outlined in Section II). We first discuss partial productivity measures before assessing research into total factor productivity. The principal focus is on methodological issues and the main results in the literature.

Partial Productivity Measures

Partial productivity ratios (which relate output to one type of input only), such as output per manhour, are often used as proxies for total productivity, although as will be seen these measures suffer a number of deficiencies. In this respect, a number of studies have argued that useful insights into bank productivity can be gained by considering accounting ratios such as asset size and operating revenue per employee. For instance, Fanning (1981) found that, although on such measures the productivity of the U.K. clearing banks in the early 1980s was improving, it was still

---

[6] If it narrows interest margins, it will reduce national income measures (the more monopoly/oligopoly, the higher indicated output), although if more loans are made, this may be partly offset. The production approach is unaffected unless more loan accounts are opened. The traditional intermediation approach shows a fall in output (higher interest costs) unless this is offset by a larger volume of loans.

[7] These may include effects on economic development and endogenous growth, as well as external economies of scale between institutions. For a survey, see Colwell and Davis (1992).

inferior to international competition, suggesting that overmanning existed in U.K. banking.

Other partial studies have focused on the relative productivity of the banking industry in relation to other sections of the economy. For instance, Baumol and Oates (1972) suggested that the service sector is inherently resistant to the kind of technological progress which has continually increased productivity elsewhere in the economy, particularly in manufacturing.[8] Thus, so long as relative wages in various sectors remain the same, costs in the service sector must rise faster than those elsewhere in the economy ("Baumol's cost disease"). Kinsella (1973) attempted to apply this hypothesis to the banking industry in Ireland 1960–71 by comparing its labour productivity to that of manufacturing and services as a whole. The results of this study did indeed support Baumol's hypothesis, with (his estimate of) bank productivity only rising by 5 per cent over the period, as opposed to 30 per cent for services as a whole and 140 per cent in manufacturing.[9] In the same context, Revell (1980) assessed the cost trends among banks from 18 OECD countries from 1964 to 1977. This study provided a less damning conclusion, for whilst banks were found to have lagged somewhat behind the goods sector in productivity gains, they were not in the desperate position of the services studied by Baumol. The measure used to reach this conclusion was the ratio of operating costs to the balance sheet total, or volume of business. Revell expected this to be constant if productivity in banking was rising as fast as in manufacturing; instead it was seen to rise. Revell suggested that the lag behind manufacturing may be explained by the fact that in banking most of the technological improvements tend to be once and for all measures that cannot easily be repeated; whereas rising productivity in manufacturing is a much more continuous process. In support of Revell's conclusion that bank productivity is not as sluggish as earlier studies suggested, Tschoegl *et al.* (1984) found that, among banks world wide, employee costs per unit assets and per number of branches are falling markedly. Their study controlled for both economies of scale and product mix by picking a sample of world banks in 1979 that were no larger than the largest bank in 1967. They were thus able to conclude that changes observed were due to gains in productivity.[10]

Although these partial productivity studies provide some insights into bank performance, there are a number of *critical problems* which limit their ability to evaluate operating efficiency. In particular, as pointed out

---

[8] Note that this assertion was made before the advent of large scale computerization.

[9] Baumol (1991) tests the hypothesis on the insurance sector.

[10] This study updated an earlier regression analysis of determinants of bank employment by Kaufman (1970).

by Frazer (1982), the accounting ratio analysis favoured by Fanning can only give a useful measure of staff productivity if the banks are doing much the same business in much the same environment. This largely explains why British banks (with a large involvement in labour intensive money transmission services) were found to be low down the international list. More generally, all partial productivity studies are vitiated by their inability to account for the cost of generating changes in, for example, labour productivity; if a bank replaces labour with machines to carry out routine functions, it may raise labour productivity, but the overall costs *ex post* may be similar. As a variant of partial productivity, some studies have focused attention on the "transfer of payments" activity as a proxy for changes in bank productivity over time. Frazer explained that this is one area for which there is reliable long term data. Frazer's study covering 20 years found that the number of payment items handled by major U.K. banks had increased by a factor of 4, whilst staff numbers had only doubled. Meanwhile, in response to Kinsella's claim that bank productivity fell well behind that of manufacturing, Gambs' (1976) study of the U.S. payments system suggested that the growth of productivity in handling cheques was slightly higher than productivity growth in the U.S. economy as a whole between 1967–72. However, although this approach may overcome the problem of meaningful "comparability" it does not account for "Total Factor Productivity (TFP) differences", nor does it correct for factor intensity differentials in terms of physical and financial capital per employee.

Total Factor Productivity

Total Factor Productivity (TFP) is a generalization of the partial factor productivity (PFP) ratio. It extends the concept of PFP by embracing multiple outputs and multiple inputs in a *single* productivity ratio. The central issue of TFP measurement is the methodology adopted to estimate the weights used to combine (or value) inputs and outputs. The advantage of TFP over PFP measures is that it enables *consistent* productivity comparisons to be made across the range of banks' outputs and inputs; whereas *a priori* there is nothing to gauarantee that the equivalent n*m PFP ratios will give a consistent picture of productivity performance. However, calculation of TFP over time and between industries is difficult because proportions of factor inputs do not remain constant over time or between industries and their contribution to output is difficult to unravel. Partly for this reason, most of the work cited below focuses on cross-sectional interbank comparison.

The latest work focusing on TFP measurement has tended to use estimated frontier production functions. This has superseded traditional econometric TFP measures as in Solow (1957), principally because these

measures were based on ordinary least squares (OLS) *average* production functions which distorted efficiency results. That is, proximity to an OLS production function does not necessarily mean productivity is maximized. Also, the OLS approach cannot separate technical efficiency from technological change. In addition, it was unrealistically assumed that competitive market conditions existed and that only a single output was produced; see Berg (1991). Frontier work can be classified according to the way the frontier is specified and estimated. For instance, the frontier may or may not be specified as a parametric function of inputs. Also, an explicit statistical model of the relationship between observed output and the frontier may or may not be specified. Finally, the frontier itself may be specified to be either deterministic or stochastic. From the various permutations that exist, the *deterministic non-parametric frontier*[11] approach has seen most development, and a substantial body of applied work in banking has utilized it. This approach was pioneered by Farrell (1957). His approach is non-parametric in the sense that it is not based on any explicit model of the frontier or of the relationship of the observations to the frontier. Instead a convex hull of the observed input-output ratios is constructed by linear programming techniques; which is supported by a subset of the sample with the rest of the sample points lying within it.

Data Envelopment Analysis (DEA)

A development from Farrell's work is the linear programming based data envelopment analysis (DEA). This is also a non-parametric, deterministic methodology, which was introduced by Charnes *et al.* (1978) for the assessment of efficiency of non-profit-making organizations, where accounting profit measures are difficult to compute (particularly in the public sector). More widely, DEA can evaluate the relative efficiency of a set of organizations in their use of multiple inputs to produce multiple outputs, where the efficient production function is not known or easily specified. It does this by comparing several organizations' (denoted p) observed outputs (Y_{jp}) and inputs (X_{ip}). It identifies the *relatively* more efficient "best practice" subset of firms and the subset of firms that are relatively inefficient (and the magnitude of their inefficiencies) compared to the "best practice" firms. More formally, we mazimize:

$$E_p = \sum_j u_j Y_{jp} \Big/ \sum_i v X_{ip} \qquad (1)$$

subject to $E_p \leqslant 1$ for all p and weights $v_i, u_j > 0$.

[11] On the other hand, from first principles it is difficult to justify deterministic methods. The data itself being noisy, it could be argued that a stochastic analysis is more inherently desirable.

This model is run repetitively with each firm appearing in the objective function once to derive individual efficiency ratings. Each firm will either have a derived efficiency rating either of $E=1$, which implies relative efficiency, or $E<1$, which implies relative inefficiency. (It must be stressed that $E=1$ is a "best practice" unit, which means it is not necessarily efficient but that it is not less efficient compared with other firms in the study. That is to say, DEA is a *relative* efficiency measure; it cannot measure efficiency in an *absolute* sense.) In addition, DEA facilitates the exploration of the nature of inefficiencies at a firm by identifying an efficiency reference set. This is the set of relatively efficient (best practice) firms to which the inefficient unit has been most directly compared in calculating its efficiency rating. DEA, therefore, avoids the need to investigate all units to understand the inefficiencies present.

The principal disadvantage of DEA is that the frontier is defined on the outliers rather than on the whole sample and is thereby particularly susceptible to extreme observations and measurement error. For instance, Berger and Humphrey (1990) explained that small changes in the measurement error or luck of a firm on the frontier may have a significant impact on aggregate inefficiencies because other firms are measured relative to this fully efficient firm. Second, statistical inferences cannot be made using this approach. Berg and Kim (1990) also pointed out that the non-parametric DEA cannot take into account market structure and that this is important given their finding that efficiency scores are not independent of market structure characteristics. Furthermore, inadequacies in data or sample size may vitiate DEA results. Among the counter intuitive results arising from data and sample problems was a suggestion that Continental Illinois was the most efficient U.S. bank just prior to its collapse; see Charnes *et al.* (1990). On the other hand, the same authors developed Polyhedral-cone ratio DEA which generalizes the model outlined above to enable it to incorporate exogenous expert opinion — in their case, characteristics of a set of banks whom experts unanimously agreed were efficient — which enables this misclassification problem to be reduced. Moreover, the limitation of a constant returns to scale assumption used in early DEA work has been overcome by adding a variable returns to scale constraint; cf. Banker (1984).[12]

DEA has emerged as a leading tool for efficiency evaluation in terms of both the number of research papers published and the number of applications to real-world problems. Sherman and Gold (1985) were the first to

[12] However, Berg (1991) argued that the distribution of data needs to be considered before applying such a constraint. This is because if there are few observations at higher levels of output, DEA will almost certainly identify one of them as efficient. Therefore, a constant returns to scale assumption may be more appropriate (though surely such a judgement must rest on whether it is empirically justified).

apply DEA to banking by carrying out an analysis on 14 branches of a U.S. savings bank. They adopted the production approach for measuring bank output, choosing to assess 17 transactions; whilst the inputs monitored were labour, office space and supply costs. The results revealed that six of the 14 branches were relatively inefficient.

However, this paper was criticized for being based upon a very small sample (since one should have as large a cross section as possible to maximize the discriminatory power of DEA). By way of a slight improvement Parkan (1987) applied the DEA technique to 35 branches of a major Canadian Chartered Bank in Calgary. The production approach was again used to measure output. The results from the study suggested that 11 of the 35 branches were found to be relatively inefficient. Meanwhile, a similar study by Vassiloglou and Giolias (1990) found that only nine from 20 branches of the Commercial Bank of Greece in 1987 had a maximum efficiency rating. Tulkens (1990) undertook a larger scale study when he applied the DEA and Free Disposed Hull (FDH)[13] techniques to 773 branches of a Belgian public bank and 911 branches of a private bank in the same country. Under the DEA approach less than 6 per cent of the branches were deemed efficient; whereas 74.6 per cent of the public bank's branches are on the FDH frontier compared to 57.8 per cent of the private bank's branches.

These studies applied DEA across branches within single banks; other studies have extended the application across banks. For instance, Rangan *et al.* (1988) attempted to break down inefficiency of 215 independent U.S. banks into that originating from pure technical inefficiency (stemming from wasted resources) and scale inefficiency (operating at non-constant returns to scale). Such a decomposition was made feasible by Banker's (1984) reformulation of Charnes *et al.* (1978) which added an extra constraint of variable returns to scale instead of constant returns to scale. In contrast to the branch studies, Rangan *et al.* preferred the intermediation approach to output measurement, taking the $ value of three types of loans and two categories of deposits; whilst the inputs used were labour, capital and purchased funds. The results implied that on average the banks in this sample could have produced the same level of output by using 70 per cent of the inputs actually used, largely due to pure technical inefficiencies. In an extension to their work, Rangan *et al.* (1990) separately assessed banks from the unit banking as well as branch banking states. These two organizational forms operate under very different legal environments and, this may, therefore, significantly influence the efficiency measures. However, the results showed there to be no sizeable differences in efficiency between the two groups.

---

[13] Such an approach avoids the DEA assumption of convexity.

Field (1990) applied the DEA method to a cross section of 71 British building societies in 1981. At that time 86 per cent were found to be inefficient, mainly due to scale inefficiencies. A contrast between Rangan *et al.* as well as other U.S. studies and Field is that the former's analysis indicated that the technical efficiency measures is positively related to bank size, and hence the dispersion in firms' efficiency seemed to be accounted for by their size. However, Field found that the overall technical efficiency was negatively correlated with firm size. This may relate to cartelized and oligopolistic market conditions among U.K. building societies in 1981. A further contrast to Field's work is provided by Drake *et al.* (1991) who applied DEA to building societies after deregulation in 1988. They found 37 per cent to exhibit overall efficiency — a marked increase. And they found overall efficiency positively correlated to size.

Elysiani and Mehdian (1990b) used DEA to measure the rate of technological change (RTC) for a sample of 191 large U.S. banks, based on 1980 and 1985 data. The results of this study suggested that the frontier had shifted inward due to technological advancement to the extent that the banks could have produced the same level of output in 1985 with 90 per cent of the inputs they actually used.

Finally, one of the latest studies to apply DEA in a banking context, using the intermediation approach to output measurement, was Berg's (1991) study of bank mergers in the Norwegian banking sector between 1984 and 1989. Berg noted that the accounting profits of acquiring banks were not systematically different from those of the acquired banks. He then computed efficiency scores for merging and non-merging banks, and found that merging banks had on average significantly lower efficiency scores than the industry, but there was no significant difference between acquired and acquiring banks. This suggests that efficiency was not a main reason for the mergers.

Parametric Approaches

An alternative to the DEA approach is the *deterministic parametric frontier.* Aigner and Chu (1968) were the first to develop this. They specified a homogenous Cobb-Douglas production frontier, and required all observations to be on or beneath the frontier. Førsund *et al.* (1980) suggested that their model may be written

$$\ln y = \ln f(x) - u = \alpha_0 + \sum_{i=1}^{n} \alpha_i \ln x_i - u_j, \qquad u \geq 0 \tag{2}$$

where the one-sided error term forces $y \leq f(x)$. The elements of the parameter vector may be derived either by linear programming (minimiz-

ing the sum of the absolute values of the residuals, subject to the constraint that each residual be non-positive) or by quadratic programming (minimizing the sum of squared residuals, subject to the same constraint). Although Aigner and Chu did not do so, the technical efficiency of each observation can be computed directly from the vector of residuals, since u represents technical inefficiency. As with the case for the non-parametric approach, the "estimated" frontier is supported by a subset of the data and is therefore sensitive to outliers. Moreover, the information and data requirements are much more demanding than for DEA (e.g. an assumed functional form for the production function). Third, when using programming methods, the possibility of determining statistical significance of tests of the frontier is foreclosed. However, Afriat (1972) made the model amenable to statistical analysis by making further assumptions. This development generates *deterministic statistical frontiers*, where the assumptions most often made are that the observations on u are independently and identically distributed and that x is exogenous (independent of u).

Elyasiani and Mehdian (1990a) applied the deterministic statistical frontier method, using the corrected ordinary least-square technique (COLS) to a banking study. First, the parameters of the production function were estimated, then the intercept was shifted until no residuals were positive and at least one was zero. Relative to the constructed frontier, they then calculated measurements of efficiency for banks in the sample. The sample of 144 banks from 1985 was selected to include a wide range of U.S. banks in terms of size, geographic locations and status.[14] This study adopted the intermediation approach, which measures bank output as the revenue from loans and investment; whilst inputs were assumed to be labour, capital and two categories of deposits. The results showed that on average banks in the sample generated 64 per cent of potential revenue available. This reflects the degree of inefficiency present, 80 per cent of which was scale related and 20 per cent technical.

An alternative parametric approach is to try to assess productive efficiency in relation to econometric estimation of a cost function; cf. Ferrier and Lovell (1990). The intuition of the cost function approach is that the producer is assumed to seek to produce given outputs at a minimum cost, but may not succeed. In order to capture and measure departure from efficiency it is necessary to derive parameter estimates describing the nature and cost of departures from cost minimizing behaviour as well as a (stochastic translog) cost frontier. The results

[14] Although this is good practice in terms of a large cross section, it does run the risk of not comparing like with like, e.g. rural and urban banks, where indicated efficiency might well differ due to market differences.

suggested that among the sample of U.S. banks, technical inefficiency raises cost by 9 per cent on average, while allocative inefficiency raises cost by 17 per cent. The shortfall was due largely to excessive labour utilization, and did not vary between small and large banks. The study also found modest scale economies. The authors also carried out a DEA analysis, which showed similar qualitative findings although there were differences in the magnitudes of calculated costs of technical and allocative efficiency. It was expected that DEA being non-stochastic would be more sensitive to noise, classifying such errors as inefficiency and hence estimated costs should be higher. In fact estimated cost inefficiency was comparable (technical inefficiency raises cost 16 per cent, allocative inefficiency 5 per cent). This was felt to show that the DEA production frontier is sufficiently flexible to envelop the data more closely than the translog production frontier, a second order approximation in logs.

Berger and Humphrey (1990) modified this econometric approach by estimating a "thick frontier" cost function using data from banks in the lowest average cost quartile, which are assumed to represent those banks with greater than average efficiency. The differences in predicted average costs between the lowest and highest cost quartiles are deemed to reflect inefficiencies. This approach avoids DEA's susceptibility to extreme observations, as well as the questionable assumptions (such as the one sided error distribution) needed for other parametric tests; and although it requires a subjective judgement as to where to apply the upper and lower efficiency thresholds, the authors found that their quartile segmentation assumption did not substantially violate the data. In their application of this method to all (13,951) insured commercial banks in the U.S. in 1984, they found inefficiencies in the order of 25 per cent, with technical inefficiencies (proportionate overuse of all inputs) dominating allocative inefficiencies (improper mix of inputs). In an evaluation of the importance of market structure to findings on bank efficiency, Berg and Kim (1991) also estimated a thick frontier. However, their ranking of banks (in the Norwegian market) was based on cost/output ratios rather than average total costs. They found that the average bank was about 11 per cent more efficient under the Cournot model of independent behaviour than when the conjectual variation model of interdependent behaviour is applied.

Summary

These applications of frontier analysis to banking offer two types of results. First, in terms of the type and magnitude of inefficiencies, it is suggested that technical inefficiency is more important than allocative inefficiency and that such technical inefficiency can be up to 30 per cent of costs. Indeed, Berger and Humphrey (1990) go further and suggest technical efficiency also dominates scale and scope economies.

Second, varying efficiency levels exist side by side in the market place. This begs the question of how such a market structure can exist; in particular, how can managers continue to underutilize factor inputs? A first hypothesis focuses on differences in the product range and product mix among commercial banks. As a consequence of the deregulatory movement in banking in the 1980s, banks have generally tried to create a niche for themselves in the market which would specifically suit their abilities and character. This partial specialization and concentration to create a "niche" may have reduced the intensity of competition allowing coexistence of banks with varying degrees of efficiency. Second, apparently inefficient banks may survive for regulatory reasons. For example, restrictions on interstate banking means banking markets in the U.S. have been local (or regional) rather than national. Therefore, there is likely to be little sensitivity to pricing policies of non-local banks.[15] More generally, mispriced "insurance" of banks via the lender of last resort and deposit insurance may effectively subsidize small, risky and perhaps inefficient banks. We suggest that a third possibility is that the level of economic activity may vary across the sample (spatially); those banks in depressed areas appear to be inefficient, whereas averaged over the cycle productivity might be similar between banks. Such a pattern could result from indivisibilities and fixed costs in production (it may be difficult to partly close a bank or branch — and skilled labour may be retained in a downturn despite the implicit reduction in productivity). Finally, extending the suggestion of regional markets, there may be market power which is not adequately captured in the measures; for example, firms with higher reputations may be able to charge higher prices for the same product and are under correspondingly less pressure to improve productivity.

Some omissions from productivity studies can again be suggested. Few of the productivity studies have a role for *banking capital* as an input,[16] although it is recognized that capital backing is essential to the stability of the institution. Moreover, as the Basle Accord on capital adequacy now requires such capital to be held at a ratio of above 8 per cent of risk adjusted assets, part of the production process is now fixed-coefficients and the maximum "productivity" of such capital is given. In effect, prudential requirements in the industry restrict the ability (and desirability) of banks maximizing efficiency. Meanwhile, when capital is inadequate, balance sheet growth must be restrained or spreads widened — with conflicting implications for output and productivity. Some of the

---

[15] In support of this hypothesis, Berger and Humphrey (1990) found efficiency lower in the restrictive "unit banking" states.

[16] In most other sectors counting capital as an input would be to double count fixed capital, goodwill, etc. but arguably in banking it performs a separate service of providing stability.

comments made for output also apply to productivity. For example, if a bank can increase measured productivity at a cost in terms of risk by taking on loans or accounts of low quality, is productivity correctly measured?[17] "Productivity" in this sense is often stimulated by heightened competition; and some would argue it is made possible by "regulatory insurance", which reduces market incentives to investigate risk.

IV. Conclusions

This survey suggests that output and productivity in banking remain difficult to assess theoretically, and even harder to measure. Nor is this merely due to the problem of quality change, although this clearly has important implications. It arises at a more basic level from disagreement over the nature of bank output — a concept to which at least three approaches can be distinguished, each with their own advantages as well as serious disadvantages. At the core of the problem, we suggest, lies the complexity of banking as an activity (featuring many interconnected products) and poor data; a complete picture would also take into account the additional quality dimensions (such as risk) not present in such an acute form in most other industries and would also assess the impact of regulation on the industry. The difficulties with output make assessment of productivity more problematic. Partial factor productivity measures are overlaid by differences in product mix or joint production, while total factor productivity studies, on which some progress has been made, can still only compare banks or branches cross sectionally within a defined market. But there are no straightforward substitutes for measures of productive efficiency. For example, an efficient bank may have a good record for profitability and market share, but so may a bank with market power. Hence these are not adequate discriminatory variables in assessing efficiency.

The importance of banking to the modern economy — and the magnitude of the potential market failures related to it — make research leading to further progress in these areas all the more urgent. We conclude by suggesting alternative approaches that could be pursued to address the issues. One approach would be to suggest that banking output and productivity are cases of "measurement without theory". This would suggest empirical work should be given a lower priority than development of the theory of financial intermediation and its application to output. Such a theory should coherently link the services of the financial sector

[17] Excessive balance sheet expansion by banks in several countries in the late 1980s, which led on to bad debt problems, make this issue highly relevant. See Davis (1992) for a theoretical and empirical investigation of debt growth and financial stability.

(payments/liquidity; allocation of saving; risk management; price information) and thus offer consistent measures of output and productivity. Alternatively, the existing measures could be retained but applied to specialized financial institutions (loan offices, centralized mortgage lenders, money market mutual funds) rather than banks, to show 'true" prices and output without cross subsidies and joint production.

References

Afriat, S.N.: Efficient estimation of production frontiers. *International Economic Review 13*, 568–98, 1972.

Aigner, D. J. & Chu, S.F.: On estimating the industry production function. *American Economic Review 58*, 826–39, 1968.

Alhadeff, D. A.: *Monopoly and Competition in Commercial Banking.* University of California Press, Berkeley, 1954.

Banker, R. D.: Estimating most productive scale size using data envelopment analysis. *European Journal of Operational Research 17*, 35–44, 1984.

Baumol, W. J. & Oates, W. E.: The cost disease of the personal services and the quality of life. *Enskilda Banken Quarterly Review 2*, 44–54, 1972.

Baumol, W. J.: Technological imperatives, productivity and insurance costs. *Geneva Papers on Risk and Insurance 59*, 154–65, 1991.

Benston, G. J.: Branch banking and economies of scale. *Journal of Finance 20*, 312–31, 1965.

Benston, G. J., Hanweck, G. A. & Humphreys, D. B.: Scale economies in banking: A restructuring and reassessment. *Journal of Money, Credit and Banking 14*, 435–56, 1982.

Berg, S. A.: Mergers, efficiency and productivity growth in Norwegian banking 1984–89. Norges Bank Research Paper, Oslo, 1991.

Berg, S. A., Førsund, F. R. & Jansen, E. S.: Bank output measurement and the construction of best practice frontiers. Norges Bank Research Paper 1989/6, Oslo, 1989.

Berg, S. A. & Kim, M.: Oligopolistic interdependence and banking efficiency: An empirical evaluation. Norges Bank Research Paper 1991/5, Oslo, 1991.

Berger, A. N. & Humphrey, D. B.: The dominance of inefficiencies over scale and product mix economies in banking. *Finance and Economics Discussion Series* No. 107, Federal Reserve Board, Washington DC, 1990.

Charnes, A., Cooper, W. & Rhodes, E.: Measuring the efficiency of decision making units. *European Journal of Operational Research 6*, 429–44, 1978.

Charnes, A., Cooper, W. W., Sun, D. B. & Huang, Z. M.: Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks. *Journal of Econometrics 46*, 73–91, 1990.

Colwell, R. J. & Davis, E. P.: Output, productivity and externalities, the case of banking. Working Paper, Bank of England, London, 1992.

Davis, E. P.: *Debt, Financial Fragility and Systemic Risk.* Oxford University Press, 1992.

Diamond, D.: Financial intermediation and delegated monitoring. *Review of Economic Studies 51*, 393–414, 1984.

Drake, L. & Weyman-Jones, T. G.: Technical and scale efficiency in U.K. building societies. Economic Research Paper 91/6, Loughborough University, 1991.

Elyasiani, E. & Mehdian, S.: Efficiency in the commercial banking industry, a production frontier approach. *Applied Economics 22*, 539–51, 1990a.

Elysiani, E. & Mehdian, S.: A non-parametric approach to measurement of efficiency and technological change: The case of large U.S. banks. *Journal of Financial Services Research 4*, 157–68, 1990b.

Evanoff, D. D. & Israilevich, P. R.: Productive efficiency in banking. *Economic Perspectives*, Federal Reserve Bank of Chicago, 1991.

Fanning, D.: Productivity: The human asset approach to bank rankings. *The Banker*, 31–4, Nov. 1981.

Farrell, M. J.: The measurement of productive efficiency. *Journal of Royal Statistical Society 120*, Sec. A, 253–81, 1957.

Ferrier, G. D. & Lovell, C. A. K.: Measuring cost efficiency in banking; Econometric and linear programming evidence. *Journal of Econometrics 46*, 229–45, 1990.

Field, K.: Production efficiency of British building societies. *Applied Economics 22*, 415–25, 1990.

Fixler, D. & Zieschang, K. D.: Measuring the nominal value of financial services in the national income accounts. *Economic Inquiry 29*, 53–68, 1991.

Førsund, F., Lovell, C. A. K. & Schmidt, P.: A survey of frontier production functions and of their relationship to efficiency measurement. *Journal of Econometrics 13* (Supplement), 5–25, 1980.

Frazer, P.: How not to measure bank productivity. *The Banker*, 103–5, Aug. 1982.

Gambs, C. M.: The cost of the U.S. payments system. *Journal of Bank Research 6*, 240–4, 1976.

Gilbert, R. A.: Bank market structure and competition. *Journal of Money, Credit and Banking 16*, 617–45, 1984.

Gramley, L.: A study of scale economies in banking. Federal Reserve Bank of Kansas City, 1962.

Greenbaum, S.: Competition and efficiency in the banking system: Empirical research and its policy implications. *Journal of Political Economy 75*, 461–79, 1967.

Humphrey, D. B.: Why do estimates of bank scale economies differ? *Economic Review*, Federal Reserve Bank of Richmond, 38–50, 1990.

Kaufman, G.: Bank employment; a cross section analysis of the largest banks. *Journal of Money, Credit and Banking 2*, 101–11, 1990.

Kinsella, R. P.: Baumol's cost-disease and the banks. *Skandinaviska Enskilda Banken Quarterly Review 2*, 61–5, 1973.

Kinsella, R. P.: The measurement of bank output. *Journal of the Institute of Bankers in Ireland 82*, 173–83, 1980.

Kolari, J. & Zardkoohi, A.: *Bank Costs, Structure and Performance.* Lexington Books, New York, 1987.

Mester, L.: Efficient production of financial services: scale and scope economies. *Business Review*, Federal Reserve Bank of Philadelphia, Jan./Feb. 1987.

Parkan, C.: Measuring the efficiency of service operations: An application to bank branches. *Engineering Costs and Production Economics 12*, 237–42, 1987.

Powers, J. A.: Branch versus unit banking: Bank output, and cost economics. *Southern Economic Journal 36*, 153–64, 1969.

Rangan, N., Grabowski, R., Aly, J. & Pasurka, C.: The technical efficiency of U.S. banks. *Economics Letters 28*, 169–75, 1988.

Rangan, N., Grabowski, R., Pasurka, C. & Aly, H.: Technical, scale and allocative efficiencies in U.S. banking: An empirical investigation. *Review of Economics and Statistics 52*, 211–8, 1990.

Revell, J. R. S.: *Costs and Margins in Banking. An International Survey.* OECD, Paris, 1980.

Schweiger, I. & McGee, J. S.: Chicago banking. *Journal of Business*, 1961.

Shaffer, S. & David, E.: Economics of superscale in commercial banking. *Applied Economics 23*, 283–93, 1991.

Sherman, J. D. & Gold, F.: Bank branch operating efficiency: Evaluation with data envelopment analysis. *Journal of Banking and Finance 9*, 297–316, 1985.

Solow, R. M.: Technical change and the aggregate production function. *Review of Economics and Statistics 39*, 312–20, 1957.

Tschoegl, A. E. & Choi, S. R.: Bank employment in the world's largest banks: An update. *Journal of Money, Credit and Banking 16*, 359–62, 1984.

Tulkens, H.: Non-parametric efficiency analyses in four service activities: Retail banking, municipalities, courts and urban transit. CORE DP 9050, Université Catholique de Louvain, 1990.

Vassiloglou, M. & Giolias, D.: A study of the relative efficiency of bank branches: An application of data envelopment analysis. *Journal of Operational Research Society 41*, 591–7, 1990.

Instructions to Contributors

The editorial policy of *The Scandinavian Journal of Economics* is to foster original economic research of high standard in the Nordic countries and make it known to an international readership.

Manuscripts submitted to the Editor should be in *English*. *Four copies* are requested. Articles should be no longer than 15 printed pages (approximately 20 double-spaced typewritten pages, including references, figures and tables). An *abstract* not exceeding 100 words should also be enclosed.

References to articles and books (including the first names of all authors cited) should be listed at the end of the article; references in the text to this list should be made *by year* (in parentheses). *Footnotes* and *formulas* should be numbered consecutively throughout the text and acknowledgements denoted by an asterisk. *Figures* should be suitable for direct photographic reproduction and allowances should be made for necessary reduction. *Tables* should be clearly labeled. Supplementary explanations of mathematical methods and statistical tests which, although not intended for publication, could facilitate the referees' work, should be enclosed.

The *proofs* should be carefully checked by the author and returned within a week. The publisher reserves the right to charge for alterations made at the proof stage.

The authors of articles receive 25 *offprints* free of charge. An Offprint Order Form is provided for orders for additional copies.

Manuscripts submitted after January 1, 1993 should be accompanied by a submission fee of SEK 300, payable to The Scandinavian Journal of Economics. (i) Nordic submissions: Please remit in Swedish kronor only to Swedish postal giro account no. 15 54 51-8; (ii) Non-Nordic submissions: Please remit in Swedish currency only to the above postal giro account or by check, payable in SEK, drawn on a Swedish bank.

Manuscripts (4 copies), books for review and editorial correspondence should be addressed to:

The Editor
THE SCANDINAVIAN JOURNAL OF ECONOMICS
Department of Economics
University of Stockholm
S-106 91 Stockholm, Sweden

Phone +46-8 16 30 42
Fax +46-8 15 90 61