Output, Productivity and Externalities—the Case of Banking

R J Colwell

and

E P Davis

Paper presented at conference on 'Welfare, Quality and Productivity in the Service Industries', Uppsala, Sweden, May 1991. The views expressed are those of the authors and not necessarily those of the Bank. The authors thank Joe Ganley, C Goodhart, D Miles and participants in seminars at the Bank of England, London School of Economics and Uppsala for helpful suggestions. The errors remain their own responsibility. The authors would also like to thank the Bulletin Group for editorial assistance in producing this paper.

Issued by the Economics Division, Bank of England, London, EC2R 8AH to which requests for individual copies should be addressed: envelopes should be marked for the attention of the Bulletin Group. *(Telephone: 071-601-4030.)*

©Bank of England 1992
ISBN 1 85730 085 8
ISSN 0142-6753

Contents page

Abstract

Concepts in banking output, and the empirical literature on bank productivity—which employs output concepts—are critically surveyed. Related issues concerning externalities from banking activity, which entail a deviation of private from social measures of banking output, are outlined. For output, the national accounts, production and intermediation approaches are compared. As regards producivity, both partial and total factor productivity measures, and the DEA and parametric approaches to the latter are assessed. The externalities from banking are shown to include contributions to economic development, external economies of scale between institutions, and contagious effects of failures. Among the most striking results is the prevalence of technical inefficiency in banking. In addition, externality issues are rarely considered in combination nor assessed empirically. But more generally, 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.

Introduction

Recent developments in financial markets such as deregulation, securitisation, internationalisation, 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? Are there side effects to the growth of financial services?

The answers to such questions are of interest from a variety of points of view. At a macro level, data suggest that output of banking and finance in major OECD countries has grown strongly in recent years, as a share of GDP (Table 1). What does it imply? Comparisons of simplistic measures of productivity suggest Luxembourg and UK banks are more productive than others (Table 2)—but does this make them more competitive? And some commentators suggest there are market failures in banking, causing it to draw labour from more socially productive uses elsewhere [Tobin (1984)]. How can such suggestions be evaluated?

Table 1
Financial and business services output and employment

Share of GDP (%)

	United Kingdom	United States	Japan	Germany	France
1979	15.3	19.9	14.6	10.1	15.7
1989	21.9	25.9	16.9	11.6	21.5

Share of employment (%)

	United Kingdom	United States	Japan	Germany	France
1979	6.9	10.2	4.7	3.2	12.4
1989	11.6	14.0	5.2	3.4	15.7

Source: OECD.

Table 2
Productivity measures in banking 1986

Dollars in millions

	Assets per employee	Employees per branch	Assets per branch
Luxembourg	$4.8	39.9	$189.3
United Kingdom	2.5	28.2	70.5
Belgium	2.2	13.9	29.8
Netherlands	1.8	25.2	44.3
West Germany	1.5	30.0	44.4
France	1.2	23.1	28.5
Ireland	0.9	23.4	21.1
Denmark	0.7	16.7	11.7
Spain	0.7	9.6	6.7
Italy	0.6	32.6	19.5
Portugal	0.8	36.3	12.6
Greece	0.3	27.8	7.9

Source: European Banking Federation, December 1986.

Narrowing the focus to a more micro level, in the case of productivity bank managers may desire improved efficiency as a means of widening margins, which increase profitability and hence retentions, from which capital may be accumulated. Also, the higher a bank's degree of efficiency, the lower the possibility of it failing or being subjected to a takeover. Customers are interested in bank productivity because it can translate into lower service charges and lower loan rates, as well as a higher quality of service. And regulators wish to know whether banks have sufficient productive efficiency to be viable in a competitive environment. Similar issues arise for output (a workable definition of which is in any case needed to measure productivity) and externalities.

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? Meanwhile externality issues in banking, while often analysed individually, are rarely assessed together to give an overall view of the external contribution of banking, nor how they could be measured.

This paper seeks to provide an overview of the issues in these areas. However, partly reflecting the pattern of the literature as well as the interests of a Central Bank, it approaches the three issues in different ways. In the first section, output is addressed largely on a conceptual basis, the survey outlining the main views in the literature and some recent applications, as well as offering some criticisms of its own. In the second section, productivity is mainly approached in terms of the empirical methodology, (since most productivity studies rely on established definitions of output) although some conceptual issues are addressed, and criticisms attempted. Note that we largely abstract in this section from the problem of efficient scale (1) and focus instead on efficiency in use of inputs—allocative and technical efficiency. In the third section, externalities in banking usually identified separately in the literature are drawn together in a fairly eclectic manner and some novel approaches suggested. It is noted that very little empirical work has been done in this area. In the conclusion, we note some key policy issues related to banking, the resolution of which is made more problematic by the theoretical and empirical difficulties outlined in the paper.

1 Bank output measure

This section seeks to identify conceptual problems regarding bank output and how it may be measured. As well as being of relevance in itself, a measure of output is

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

crucial to estimation of productivity. As is well known, the outputs of *primary and secondary industries* can be measured in terms of physical quantities or money values deflated by appropriate price indices (to allow for non-homogeneous outputs).[1] However, output in the form of *services* (including financial services) cannot be measured by physical quantities. Moreover, quality problems in measuring services output are acute—does for example a switch from corner shops to supermarkets show a loss in quality (convenience) or gain (variety of goods available). But the output of *financial institutions* presents particular difficulties. In the case of banks, as well as providing customers with low risk assets, credit and payments services, banks act as intermediaries in channelling funds from savers to borrowers and provide non-monetary services such as protection of valuables, accounting services and running of investment portfolios. Not all services are paid for directly ('free' services may offset zero interest on demand deposits). 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); 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*, from which the statistics quoted in Table 1 are derived. 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, in so far 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' charges reflect services provided to final rather than intermediate consumers. In looking at the share of the sector in GDP, therefore, it is conventional

[1] Not that this is a straightforward calculation; for example, new products and quality charges make it difficult to calculate changes in output (or productivity) from a base year, particularly when working with volume as opposed to value series.

to include net interest receipts in its value added.[1] In the United States, these are attributed to depositors; in the United Kingdom, 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; Kolari and Zardkoohi (1987) provide a detailed review of this literature. 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 (see below).

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 indices, 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. (It should be noted that many authors, such as Kinsella (1980) adopted these measures for want of better information.)

To correct for some of these problems, weighted indices 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 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

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

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 by removing 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 (eg 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.(1) Besides intrinsic difficulties, the fact that the 'production approach' has not been used for interbank productivity studies reflects the difficulties encountered in collating accurate data.(2)

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

(1) For instance, Sherman and Gold's (1985) study of a US savings bank measured output as a weighted average of the 17 services most commonly offered by the branches; while Vassiloglou and 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 8 categories in his assessment of a Belgian public bank.

(2) Comprehensive data are only available for the United States, and even this has questionable features (Elyasiani and Mehdian (1990a)).

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, while 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.

Berger and Humphrey (1990b) drew attention to the need, before making a decision on which method to use, to firstly identify which banking functions are most important for the purpose of the study being undertaken. They outlined three approaches to this initial identification process. Under the *asset approach*, banks are considered only as financial intermediaries between liability holders and those who receive bank funds, and bank outputs are considered to be just loans and other assets (see Sealey and Lindley (1977)). The *user cost approach* determines whether a financial product is an input or output on the basis of its net contribution to bank revenue. If the financial returns on an asset exceed the opportunity cost of funds or if the financial costs of a liability are less than the opportunity cost, then the instrument is considered to be a financial output (see Hancock (1985)). Under the *value added approach*, those factors having substantial value added are employed as important outputs (see Berger, Hanweck and Humphrey (1987)).

To summarise, 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 interbank studies. In the light of Berg, Forsund 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.

Additional comments

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 by increasing risk. Should output be 'sustainable' and hence discounted for risk? And should any account be taken of diversification? Note that revenue takes variations in *ex-ante* risk premia into account and hence output increases more if risk premia are increased than if they are not. 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. Note that in this connection, the national accounts measure counts all of the spread as depositor or lender services, with no return to risk bearing.

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. For example, in the United Kingdom there have been considerable changes in characteristics of deposits (interest-bearing current accounts, no notice on time deposits, cheque books with time deposits). Custom made products are common in securities markets. At least some of these can be objectively measured, perhaps using 'Hedonic price indices' (Deaton and Muellbauer (1980)), although integration into measures of output could be problematic (see Shaffer and David (1991) for an attempt to measure economies of scale using such techniques). Of course, the increase in explicit charging and 'unbundling' of financial services previously provided jointly makes output measurement easier.

Third, the various measures do not allow for intertemporal relationships that are crucial in banking. 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.

Fourth, what happens to measures of output when *competition* increases? If it narrows interest margins, it will reduce national income measures,[1] 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.

It may be suggested that recent developments in the *theory of intermediation* may offer insights into bank output. The traditional theory of banking relates to economies of scale (Gurley and Shaw (1960)). But more recent studies have focused on information asymmetries between borrowers and lenders. These arise from the inability of investors to screen the quality of enterpreneurs and firms (Leland and Pyle (1977)) and to monitor their performance (Diamond (1984)). There may be economies of scale in monitoring making delegation of monitoring to banks desirable. Banks may have informational advantages arising from ongoing credit relationships; from access to the borrower's deposit history (Fama (1985)); and from use of transaction services (Lewis (1989)). If monitoring is the crucial activity of banks, should more account be taken of it in output measures? Is it a cost (required to provide services) or a service in itself? As with risk, is the best way of measuring it

[1] The more monopoly/oligopoly, the higher indicated output.

in terms of outturns, in that successful monitoring will reduce loan losses as a proportion of the balance sheet?

Finally, there may be a deviation of social from private output due to externalities, as discussed in Section 3.

2 Productivity

Having outlined the main issues concerning the measurement of bank output, the paper now goes on to discuss work on productivity in banking (which uses the concepts of output as outlined in Section 1). We first discuss partial productivity measures before assessing research into total factor productivity. The principal focus in this section is on methodological issues and the main results in the literature.

Application of partial productivity measures to banking

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 UK clearing banks in the early 1980s was improving, it was still inferior to international competition, suggesting that overmanning existed in UK banking.

Other 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.(1) 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% over the period, as opposed to 30% for services as a whole and 140% in manufacturing.(2)

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 while banks were found to have lagged somewhat behind the goods sector in

(1) Note that this assertion was made before the advent of large-scale computerisation.

(2) Baumol (1991) tests the hypothesis on the insurance sector.

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. His 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. He was thus able to conclude that changes observed were due to gains in productivity.(1)

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 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.

Activity-based studies of bank productivity

As an alternative to partial productivity, some studies have focused attention on the 'transfer of payments' activity as a proxy for changes in bank productivity over time. Frazer (1982) 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 UK banks had increased by a factor of 4, while 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 US payments system suggested that the growth of productivity in handling cheques was slightly higher than productivity growth in the US 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.

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

Total Factor Productivity

Total Factor Productivity (TFP) is a generalisation 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 guarantee 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.

Frontier measures of productivity

The study of production frontiers in economics has been active for over three decades, but only recently has it attracted widespread attention. For instance, applications to the banking sector have been clustered in the last five years. The 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*[1] approach has seen most development, and a substantial body of applied work in banking has utilised 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 organisations, where accounting profit measures are

[1] 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.

difficult to compute (particularly in the public sector). More widely, DEA can evaluate the relative efficiency of a set of organisations 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 organisations' (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 maximise:

$$E_p = \sum_j u_j \quad Y_{jp} \,/ \sum_i v_i \quad X_{ip} \qquad (1)$$

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

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.

Advantages and limitations of DEA [1]

Frontier analysis, such as DEA has superseded traditional econometric TFP measures [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 maximised. Also, the OLS approach cannot separate technical efficiency from technological change. In addition, it was unrealistically assumed that competitive market conditions existed (see Berg (1991)) and that only a single output was produced. This contrasts with DEA which is a true *frontier*, where no functional form is imposed on the data and all outputs and inputs can be handled simultaneously for TFP.

However, 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

[1] An exhaustive overview of the advantages of DEA and its interpretation as a TFP measure can be found in Ganley (1989).

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. In addition, it may be suggested that application of a frontier analysis such as DEA to the private sector was not justifiable due to the presence of freedom to redeploy resources to another industry. At least, this should involve the introduction of prices or other weighting devices for the evaluation of otherwise non-comparable alternatives rather than unweighted quantities of output. Berg and Kim (1991) 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 US bank just prior to its collapse (Charnes *et al* (1990)). On the other hand, the same authors developed Polyhedral-cone ratio DEA, which generalises 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 (Banker 1984). 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 meaningful.

Applications of DEA to banking

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

However, this paper was criticised for being based upon a very small sample (since one should have as large a cross section as possible to maximise 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. Parkan limited the number of outputs and inputs to six by applying a weighting scheme to aggregate some of the initially proposed variables and eliminating others. This was done because DEA provides a better contrast in comparing branches with respect to their efficiency when the number of branches is significantly larger than the sum of the number of inputs and outputs. The results from the study suggested that eleven of the thirty-five branches were found to be relatively inefficient. Meanwhile, a similar study by Vassiloglou and Giolias (1990) found that only nine from twenty 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)[1] 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% of the branches were deemed efficient; whereas 74.6% of the public bank's branches are on the FDH frontier compared to 57.8% 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 US 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 dollar value of three types of loans and two categories of deposits; while the inputs used were labour, capital and purchased funds. The results showed that the average value of efficiency for the sample was 0.70. This implied that on average the banks in this sample could have produced the same level of output by using 70% of the inputs actually used. Hence a significant amount of inefficiency seemed to exist, almost all of which appeared to be due to pure technical inefficiencies. This result may be compared to that of Elyasiani and Mehdian (1990a) who used a deterministic statistical form of frontier analysis to make inferences about technical and scale inefficiencies in a random sample of 144 US banks in 1985 (see below). Rangan *et al* found that most of the inefficiency was due to technical inefficiency whereas the Elyasiani *et al* results suggested scale inefficiency was the dominant factor.

Rangan *et al* (1990) sought to extend their own work. The sample was composed of banks from the unit banking as well as branch banking states. These two organisational forms operate under very different legal environments and, this may, therefore, significantly influence the efficiency measures. In order to investigate this issue, the pooled sample was split into two subsamples of banks that are allowed to operate branches (212) and those that are prohibited from operating branches (110). Separate production frontiers were then calculated for each subsample. However, the

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

results showed there to be no sizable differences in efficiency between the two groups.

Field (1990) applied the DEA method to a cross section of 71 British building societies in 1981. At that time 86% were found to be inefficient, mainly due to scale inefficiencies. A contrast between Rangan *et al* as well as other US 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 cartelised and oligopolistic market conditions among UK 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% to exhibit overall efficiency—a marked increase. But again they found overall efficiency positively correlated to size.

Elysiani and Mehdian (1990b) used the non-parametric DEA approach to measure the rate of technological change (RTC) for a sample of 191 large US banks. They derived RTC relative to production frontiers 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 1980 with 90% 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. Berg went on to assess whether the mergers might have helped the participating banks catch up with the industry's average performance by calculating Malmquist input-based productivity indices. However, the rates of productivity growth do not change significantly within the first three years of merging.

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. Forsund *et al* (1980) suggested that their model may be written

$$\ln y = \ln f(x) - u$$

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

where the one-sided error term forces $y \leq f(x)$. The elements of the parameter vector may be 'estimated' either by linear programming (minimising the sum of the absolute values of the residuals, subject to the constraint that each residual be non-positive) or by quadratic programming (minimising 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 (eg 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 are estimated, then the intercept is shifted until no residuals are positive and at least one is 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 US banks in terms of size, geographic locations and status.[1] This study adopted the intermediation approach, which measures bank output as the revenue from loans and investment; while inputs were assumed to be labour, capital and two categories of deposits. The results showed that on average banks in the sample generated 64% of potential revenue available; where the latter is defined to be the level of revenue generated by best practice when no technical or scale inefficiency exists. This reflects the degree of inefficiency present, 80% of which was scale related and 20% technical.

An alternative parametric approach is to try to assess productive efficiency in relation to econometric estimation of a cost function [Ferrier and Lovell (1990)]. The intuition of the cost function approach is that the producer is assumed to seek to

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

produce given outputs at a minimum cost, but may not succeed. In order to capture and measure departures from efficiency it is necessary to derive parameter estimates describing the nature and cost of departures from cost minimising behaviour as well as a (stochastic translog) cost frontier. The results suggested that among the sample of US banks, technical inefficiency raises cost by 9% on average, while allocative inefficiency raises cost by 17%. The shortfall was due largely to excessive labour utilisation, 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%, allocative inefficiency 5%). This was felt to show that linear programming production frontier is sufficiently flexible to envelop the data more closely than the translog production frontier, a second order approximation in logs. Finally, the ranking of individual banks by inefficiency differed somewhat between the approaches, which could be due to the noise problem or alternatively due to specification error or the role of mean values in the econometric investigation.

Berger and Humphrey (1990) adapted 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 United States in 1984, they found inefficiencies in the order of 25%, 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% more efficient under the Cournot model of independent behaviour than when the conjectual variation model of interdependent behaviour is applied. (Note that, in contrast to standard game theory, this model was centred on the assertion that other firms retaliate in response to output, rather than price, changes.)

Summary of results

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 (or scale) inefficiency and that such technical inefficiency can be up to 30% of costs. Indeed, Berger and Humphrey (1990) would go further and suggest technical inefficiency 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 underutilise factor inputs? Elyasiani and Mehdian (1990a) argued that several hypotheses may be offered in response, each of which presents a manifestation of some market imperfection. (Note that in principle each of these hypotheses could be tested by including an extra discriminatory variable in the equation, but then the power of the equation to identify best practice would be sharply reduced.) One 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. Each group of banks has in practice concentrated on a particular subset of banking activities and acquired a comparative advantage over others in those activities. This partial specialisation and concentration to create a 'niche' may have reduced the intensity of competition allowing coexistance of banks with varying degrees of efficiency. (This view does not apply to Field's (pre-deregulation) sample of building societies.) Again, apparently inefficient banks may survive for regulatory reasons. For example, restrictions on interstate banking means banking markets in the United States have been local (or regional) rather than national. Therefore, there is likely to be little sensitivity to pricing policies of non-local banks.[1] More generally, mispriced 'insurance' of banks via the lender of last resort and deposit insurance, may effectively subsidise small, risky and perhaps inefficient banks. Alternatively there could be broader prudential constraints on maximisation of efficiency, as discussed in the next section.

We suggest that a third possibility is that the level of economic activity may vary across the sample (spatially); those banks in depressed areas would 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; and more generally firms with

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

higher reputations may be able to charge higher prices for the same product, and are under correspondingly less pressure to improve productivity.

Additional comments

Few of the productivity studies have a role for *banking capital* as an input,[1] although it is recognised 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% 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 maximising efficiency. Meanwhile, when capital is inadequate, balance sheet growth must be restrained or spreads widened—with conflicting implications for output and productivity.

Labour input may be mismeasured; if unpaid overtime is rife in banking, productivity measures may be inaccurate. And quality of inputs may be changing if some banks are using more highly skilled workers but this is not reflected in labour costs.

Some of the comments made in Section 1 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?[2] As discussed in the next section, such behaviour can lead to bank failures and externalities in the rest of the banking sector. '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.

3 Externalities

In this third section we examine the incidence of external benefits and costs in banking, and assess the policy issues raised. To the extent such externalities arise, they may modify conclusions regarding output and productivity that were drawn above. Following Layard and Walters (1978), we define externalities as cases when the consumption or production decisions of one agent affect the consumption or production opportunities open to another directly, rather than through the prices which she faces.[3] The general conclusion is that banking is particularly susceptible to externalities, (and this is partly reflected in regulation applied to banks), but measurement is extremely problematic. Nevertheless, particularly given the

[1] 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.

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

[3] Changes in decisions of others due to induced price changes ('pecuniary externalities') are excluded from the discussion.

important policy issues raised, the field warrants further study. The following cases are distinguished; innovation; the role of finance in economic development; external economies of scale; systemic risk and its effects on the wider economy; problems relating to information asymmetry; and free riding.

Innovation, where the inventor of a new product is unable to capture all the returns to his invention which thus benefits other firms, is in principle a major source of market failure in finance, where products tend to be homogeneous and easily copied. This phenomenon is often suggested to lead to under-investment in innovation; the basic argument [Arrow (1962)] is based on the public-good status of knowledge, which even a patent system can only partly overcome. The inventor is unable to appropriate all the returns to research activity, so will carry out less. Since patents are difficult to enforce in finance, this argument should, on the face of it, apply strongly. Dasgupta and Stiglitz (1980) put forward a modified version of insufficiency—that circumstances may arise in which research is insufficiently risky in a competitive equilibrium. Firms are biased in favour of less risky projects as long as the interest rate is positive. This may offer some explanation why financial markets have not produced all Arrow-Debreu contingent securities such as long-term futures contracts—though obviously there are other, more intrinsic reasons.

However, it can also be argued that there will be pressures for over-investment in research in finance in terms of innovation, part of the argument for which relies on externalities [Davis (1988)]. In effect, the private gain of innovation in finance may be much higher than the social benefit. (In technical terms, there is an externality which drives private gain in excess of social benefit, deriving from the failure of a firm to take into accounts that its gain is another firm's loss.) Also, there may be 'races'—entailing duplication of effort to produce a new product first. Despite the face that social gains to introducing such innovations earlier than would otherwise be the case are plausibly rather low, private gains of obtaining business following the introduction of innovations may be high. The importance of such innovation lies in obtaining and keeping an investor base, gaining reputation and expertise. The duplication of effort in production of such innovations is arguably a deadweight loss, which increases the costs of intermediation (unless such research is purely a substitute for advertising). In addition, the deviation of private for social returns in finance may divert skilled labour and capital from other parts of the economy where social productivity may be higher (Tobin (1984)). Similar arguments apply to financial analyses such as reports on the budget or analyses aiming to uncover mis-valued securities with the intention of realising private gains either by attracting investors or trading themselves (Brealey (1985)). Finally, Silber (1975) suggested that innovations occur as a means by which financial firms seek to circumvent operational constraints, such as regulations. This may be against the social good, impairing, for example, monetary policy.

One can go further and argue that there may be strong externalities to the introduction of new instruments close to existing ones in characteristics space. If the market for the new instrument collapses—the classic case being the perpetual FRN (floating-rate note) market—the existing instrument may also become moribund—the dated FRN. This occurred although the problem with the perpetual FRN—which some have suggested resulted partly from the fact that investors and traders had not fully understood its equity characteristics—found no echo with dated FRNs, pricing of which as a debt instrument is straightforward. This, one can suggest a case of contagion as much as is contagion between failing and sound, but comparable, financial institutions. Such externalities to product differentiation may be less common in goods markets.

Of course, as well as innovating itself, financial services also benefit from spillovers from inventions elsewhere. Bresnahan (1986) sought to measure the welfare benefits to financial services from adoption of new technologies in the computer industry, and found large social gains not captured by manufacturers of computers.[1] Meeting the same problem as outlined in sections 1 and 2—namely that measures of real output and hence of effects of new technology on productivity in banking are highly problematic—Bresnahan instead inferred the value spilled over from the area under the derived demand curve for computers (which, because it is an intermediate good, entails both increases in producers' and consumers' surplus). A key assumption of his approach is that the financial services sector is competitive and thus will act as an agent of customers. Hence purchases of computers by the sector can be treated as if they were made by customers, and gains inferred from the derived demand for computers.

A view of innovation as having positive externalities can be related more generally to theories of *endogenous growth* [see, for example, Schumpter (1942) and Romer (1989a)] where it is argued that innovation is a key element of the growth process. Romer suggests, first, that technological change (in terms of improved design of production capacity) lies at the heart of economic growth, providing the incentive for capital accumulation, and together with capital accumulation accounts for much of the observed increase in productivity. Second, technological change arises largely, from intentional actions of agents responding to market incentives. Third, any such improvement can be generalised at no cost and is incompletely excludable. The second condition generates a need for imperfect competition so as to compensate the firm for R&D expenditures; the third highlights spillovers or external effects of such activity. As regards process innovation (eg organisation of cheque clearance, invention of ATMs) as well as certain product innovations (subject to the critique above) banking would seem to fit into Romer's theory similarly to innovations in any other industry.

[1] Note that such spillovers exist in principle for most innovations, and are not peculiar to computers or banking.

The role of finance in *economic development* has been discussed by Gurley and Shaw (1960), Goldsmith (1978), Abraham (1986) and surveyed more recently by the World Bank (1989). The basic argument is that financial innovation generally, and the development of a banking system in particular, has an external effect on the organisation of economic relationships in an economy, thus affording a key precondition for sustained economic development. However, Abraham (1986) concluded that financial innovation alone cannot be relied upon to generate economic development, but rather it depends on the interplay of many factors. Progress in modelling the externalities of banking in economic development has been made by Bencivenga and Smith (1991). They develop an overlapping generations endogenous growth model, where agents face random future liquidity needs and accumulate capital and a liquid but unproductive asset. Introduction of intermediaries can shift the composition of saving towards capital, as well as preventing unnecessary capital liquidation, so intermediation is growth promoting. The key features of banks in this context are that they provide liquidity services to a large number of depositors, hence due to the law of large numbers demand for withdrawals is fairly predictable; they hold liquid reserves against withdrawal demand; they issue liabilities that are more liquid than their assets; and they reduce the need for self-financing of investment. These mean banks reduce investment in liquid assets relative to a situation where each individual must self insure; and also reduce the need for enterpreneurs to liquidate capital due to liquidity needs. Both of these will under certain conditions raise the growth rate. Between them these changes imply spillover externalities from banking leading to social increasing returns to scale in production.

Another possible link between finance and development could be via a development of the theory of endogenous growth, focusing on human capital (Romer (1989b)), with the skills associated with banking being a key input to the growth process. Alternatively, one could refer to the theory of intermediation based on information asymmetry as surveyed by Gertler (1988) and referred to below in terms of Bernanke's (1983) work on transmission of the Great Depression. In the same way that bank failure has a negative external effect on agents unable to obtain credit elsewere, establishment of banks provides the precondition for the financing of such agents, and hence the growth of the company sector, which provides a key impetus to development.

A related case to development concerns *foreign direct investment* (fdi) in banking where it is often suggested that foreign banks may offer external benefits to domestic financial systems in terms of introduction of new instruments, financial services and financing concepts. They may offer managerial efficiency that will be an example to domestic firms. They may offer more efficient intermediation per se. Implicitly, fdi may be a way for developing countries to assimilate spillovers from innovations in advanced countries, as long as they have the human capital to do this. These arguments suggest that developing countries might consider subsidising the development or establishment of foreign financial institutions.

The observed spatial organisation of banking (where banks often cluster together) suggests a potential importance of *external economies of scale* where the externality is the effect one bank on other banks' cost or demand functions, for which it is not renumerated or charged. The best example is probably the development of international financial centres [Davis and Latter (1989), Grilli (1989), Davis (1990)] where patterns of establishment and casual empiricism clearly imply the presence of such economies. These include, for example, the fact that firms participate in organised markets whose liquidity (enabling rapid execution of large orders with minimum disturbance to prices) and efficiency (in establishing prices which reflect all available information) increase with the number of participants. Firms, whether in the same or related activities, need and benefit from close business contacts with each other. A pool of skilled labour develops. In such an environment business may grow in a self-sustaining manner; implicitly, a major financial centre may be a form of 'natural monopoly'. The benefits arising from contacts and participation in markets may increase progressively with the number of firms in the locality, and firms continue to be attracted to the centre because of the numbers already there. Business becomes concentrated and competing centres may find it hard to become established. (While 'concentration' on one centre is a global welfare improvement.) Competing governments may seek to subsidise new entrants to their centres, but increasing returns suggest this is unremunerative.

The case of *systemic risk* offers an example of negative externalities to banking. To recap briefly, banks provide liquidity insurance to risk averse consumers facing private liquidity risks, while reflecting the preferences of borrowers, banks' assets are long term and illiquid, ie banks transform illiquid assets into short-term liquid liabilities. The risk sharing deposit contract and illiquid liabilities (whose value is unknown) leaves banks vulnerable to panic runs even if they are not insolvent, because banks must pay withdrawals on demand until insolvency is declared, and hence depositors who withdraw funds first minimise the risk of not being paid in full [Diamond and Dybvig (1983)]. The externality arises from a danger of systemic failure due to contagious bank runs. Widespread failure of banks may create a further strong negative externality to agents in the real sector as the production process is interrupted and assets are prematurely liquidated. In addition, a significant proportion of borrowers, due to private information held by banks and banks' unique role as monitors and evaluators of loan contracts, can *only* obtain credit from banks. Bernanke (1983) suggested this was the major transmission mechanism of the Great Depression. (For an assessment of experience of systemic risk in recent decades, see Davis (1992).)

Systemic risk and its potential effects on the wider economy are of course the bases for bank regulation, deposit insurance and the lender of the last resort (where regulation 'protects' the deposit insurer/lender of last resort against moral hazard created by mispricing of the insurance provided). Measurement of potential externalities related to systemic risk would clearly be helpful to design of such regulation.

Problems relating to *information asymmetry and investor protection* may also have external aspects. As is well known, if it is difficult or costly for the purchaser of a good or service to obtain sufficient information on the quality of the product in question, there is a tendency for individual sellers to try to raise their profit margins and their market share by cutting production costs and lowering product quality (fraud, high risk, deception etc). Buyers (who are unable to distinguish between sellers) may respond by withdrawing from the market altogether, hence there is an externality to other sellers. Such phenomena are of particular importance for financial services such as banking, because clients are often seeking advice or safe keeping for a sizable proportion of their wealth, contracts are often one-off and involve a commitment over time. This is the basis for investor protection legislation and associated regulatory structures.

Both systemic risk and information asymmetry may entail *free riding*; a condition where an institution chooses deliberately to take advantage of the good reputation of other firms by selling low-quality services in order to increase its profitability. For example, in the case of systemic risk, this would entail pursuing a deliberately high risk strategy, free riding on the good reputation of other banks. Such behaviour may damage the reputation of other firms even if the severe consequences outlined above are not realised, thus increasing their costs of funds or reducing demand for their services. This externality formed the basis of 'clubs' of banks in certain countries such as the United Kingdom prior to deregulation, which entailed tight regulation of entry standards and co-operation between incumbents to maintain quality. The risk is, of course, of monopolistic behaviour. In general, deregulation renders such 'clubs' ineffective (because free riders cannot be excluded) and hence shifts the burden of maintaining standards and stability to the regulator.

This review of externalities differs from that of output and productivity in that empirical work is sparse in this area, especially taking all the externalities together rather than focussing on one type. Therefore, this section of the survey has been largely an enumeration of theoretical issues for future empirical research. However, it is suggested that their importance to policy makes further research assessing such externalities all the more urgent.

Conclusions

This survey suggests that welfare implications of the development of 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 with a defined market. Issues such as risk and regulation again arise, for example the role of regulation in limiting the degree to which banks can pursue technical efficiency. 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.

Finally externalities, which this survey suggests are of particular importance in banking, are only at the stage of theoretical modelling, and are rarely either considered in combination or evaluated empirically.

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. Among the issues of policy relevance on which such progress could cast light are the following:

- Evaluation of the impact of financial regulation on the industry and the economy. Are there severe efficiency/stability tradeoffs?

- Assessment of the relative merits of specialised and diversified institutions.

- Welfare implications of financial liberalisation.

- Implications of bank mergers for efficiency.

- Measurement and evaluation of the contribution of banking to output and welfare. Is a high share of banking in GDP due to comparative advantage and efficiency, or entry barriers and inefficiency?

A number of these issues are of particular relevance to former command economies in Eastern Europe contemplating the design of a financial system from scratch.

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 (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 specialised financial institutions (loan offices, centralised mortgage lenders, money market mutual funds) rather than banks, to show 'true' prices and output without cross subsidies and joint production. Such an assessment would give building blocks to a better understanding of banks.

References

Abraham J P (1986), 'Financial innovation and economic growth', in a symposium on 'Europe and the future of financial services', 5–7, November 1986. Commission of the European Communities.

Afriat S N (1972), 'Efficient estimation of production frontiers', *International Economic Review*, 13, pages 568–98.

Aigner D J and Chu S F (1968), 'On estimating the industry production function', *American Economic Review*, 58, pages 826–39.

Alhadeff D A (1954), 'Monopoly and competition in commercial banking', University of California Press, Berkeley.

Arrow K J (1962), 'Economic welfare and the allocation of resources for invention' in R Nelson (ed), 'The rate and direction of inventive activity: Economic and social factors', *NBER*, Princeton University.

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

Baumol W J and Oates W E (1972), 'The Cost Disease of the Personal Services and the Quality of Life', *Enskilda Banken Quarterly Review*, 2, pages 44–54.

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

Bencivenga V R and Smith B D (1991), 'Financial intermediaton and endogenous growth', *Review of Economic Studies*, 53 pages 195–209.

Benston G J (1965), 'Branch banking and economies of scale', *Journal of Finance*, 20, pages 312–31.

Benston G J, Hanweck G A and Humphreys D B (1982), 'Scale Economies in Banking: a restructuring and reassessment', *Journal of Money, Credit and Banking*, 14, pages 435–56.

Berg S A (1991), 'Mergers, efficiency and productivity growth in Norwegian banking 1984–89', *Norges Bank Research Paper.*

Berg S A, Forsund F R and Jansen E S, (1989), 'Bank output measurement and the construction of best practice frontiers', *Norges Bank Research Paper*, No 1989/6.

Berg S A and Kim M (1991), 'Oligopolistic interdependence and banking efficiency: an empirical evaluation', *Norges Bank Research Paper*, No 1991/5.

Berger A N, Hanweck G A and Humphrey D B (1987), 'Competitive viability in banking: scale, scope and product mix economies', *Journal of Monetary Economics*, 20, pages 501–20.

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

Berger A N and Humphrey D B (1990b), 'Measurement and efficiency issues in commercial banking', *Finance and Economics Discussion series*, No. 151, Federal Reserve Board, Washington DC.

Bernanke B S (1983), 'Non-monetary effects of the financial crisis in propagation of the Great Depression', *American Economic Review*, 73, pages 257–76.

Brealey R A (1985), 'The changing structure and regulation of UK securities markets', *London Business School Working Paper*, No. IFA, pages 76–85.

Bresnahan T F (1986), 'Measuring the spillovers from technical advance; mainframe computers in financial services', *American Economic Review*, 76, pages 742–55.

Charnes A, Cooper W and Rhodes E (1978), 'Measuring the efficiency of decision making units', *European Journal of Operational Research*, 6, pages 429–44.

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

Dasgupta P and Stiglitz J E (1980), 'Uncertainty, industrial structure and the speed of R&D', *Bell Journal of Economics*, 11, pages 1–12.

Davis E P (1988), 'Industrial structure and dynamics of financial markets; the primary eurobond market', *Bank of England Discussion Paper*, No. 35.

Davis E P (1990), 'International financial centres; an industrial analysis', *Bank of England Discussion Paper*, No. 51.

Davis E P (1992), 'Debt, financial fragility and systemic risk', forthcoming, The Claredon Press, Oxford.

Davis E P and Latter A R (1989), 'London as an international financial centre', *Bank of England Quarterly Bulletin*, Vol. 29, pages 516–28.

Deaton A and Muellbauer J (1980), 'Economics and consumer behaviour', Cambridge University Press, Cambridge.

Diamond D (1984), 'Financial intermediation and delegated monitoring', *Review of Economic Studies*, 51, pages 393–414.

Diamond D and Dybvig P (1983), 'Bank runs, deposit insurance and liquidity', *Review of Economic Studies*, 51, pages 393–414.

Drake L and Weyman-Jones T G (1991), 'Technical and scale efficiency in UK building societies', *Economic Research Paper*, No 91/6, Loughborough University.

Elyasiani E and Mehdian S (1990a), 'Efficiency in the commercial banking industry, a production frontier approach', *Applied Economics*, 22, pages 539–51.

Elysiani E and Mehdian S (1990b), 'A non-parametric approach to measurement of efficiency and technological change: the case of large US banks', *Journal of Financial Services Research*, 4, pages 157–68.

Evanoff D D and Israilevich P R (1991), 'Productive efficiency in banking', *Economic Perspectives*, Federal Reserve Bank of Chicago.

Fama E (1985), 'What's different about banks?', *Journal of Monetary Economics*, 15, pages 29–39.

Fanning D (1981), 'Productivity: the human asset approach to bank rankings', *The Banker*, November 1981, pages 31–4.

Farrell M J (1957), 'The Measurement of Productive Efficiency', J*ournal of Royal Statistical Society*, Vol. 120 Sec A, pages 253–81.

Ferrier G D and Lovell C A K (1990), 'Measuring cost efficiency in banking; econometric and linear programming evidence', *Journal of Econometrics*, 46, pages 229–45.

Field K (1990), 'Production efficiency of British building societies', *Applied Economics*, 22, pages 415–25.

Fixler D and Zieschang K D (1991), 'Measuring the nominal value of financial services in the National Income accounts', *Economic Inquiry*, 29, pages 53–68.

Forsund F, Lovell C and Schmidt P (1980), 'A survey of frontier production functions and of their relationship to efficiency measurement', *Journal of Econometrics,* 13 (supplement), pages 5–25.

Frazer P (1982), 'How not to measure bank productivity', *The Banker,* August 1982, pages 103–5.

Gambs C M (1976), 'The Cost of the US Payments System', *Journal of Bank Research,* 6, pages 240–4.

Ganley J A (1989), 'Relative efficiency measurement in the public sector with Data Envelopment Analysis', PhD thesis, University of London, September.

Gertler M (1988), 'Financial structure and aggregate economic activity: an overview', *Journal of Money, Credit and Banking,* 20, pages 559–96.

Gilbert R A (1984), 'Bank market structure and competition', *Journal and Money, Credit and Banking,* 16, pages 617–45.

Goldsmith R W (1985), 'Comparative national balance sheets', University of Chicago Press.

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

Greenbaum S (1967), 'Competition and efficiency in the banking system: empirical research and its policy implications', *Jornal of Political Economy,* 75, pages 461–79.

Grilli V (1989), 'Europe 1992: issues and prospects for the financial markets', *Economic Policy,* 9, pages 387–422.

Gurley J and Shaw E (1960), 'Money in a theory of finance', Brookings, Washington.

Hancock D (1985), 'The financial firm: production with monetary and non-monetary goods', *Journal of Political Economy,* 93, pages 859–80.

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

Kaufman G (1970), 'Bank employment; a cross section analysis of the largest banks', *Journal of Money, Credit and Banking,* 2, pages 101–111.

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

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

Kolari J and Zardkoohi A (1987), 'Bank costs, structure and performance', Lexington Books, New York.

Layard R and Walters A (1978), 'Microeconomic theory', McGraw-Hill, New York.

Leland H and Pyle D (1977), 'Information asymmetries, financial structures and financial intermediaries', *Journal of Finance*, 23.

Lewis M K (1989), 'Theory and practice of the banking firm', Mimeo, University of Nottingham.

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

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

Powers J A (1969), 'Branch vs Unit Banking: Bank Output, and Cost Economics', *Southern Economic Journal*, 36, pages 153–64.

Rangan N, Grabowski R, Aly H and Pasurka C (1988), 'The Technical efficiency of US banks', *Economic Letters*, 28, pages 169–75.

Rangan N, Grabowski R, Pasurka C, Aly H (1990), 'Technical, scale and allocative efficiencies in US banking: an empirical investigation', *Review of Economics and Statistics*, 52, pages 211–18.

Revell J R S (1980), 'Costs and margins in banking. An international survey', OECD, Paris.

Romer P M (1989a), 'Endogenous technological change', National Bureau of Economic Research, *Working Paper No. 3210*.

Romer P M (1989b), 'Human capital and growth; theory and evidence', National Bureau of Economic Research, *Working Paper No. 3173*.

Sealey C and Lindley J (1977), 'Inputs, outputs and a theory of production and cost at depository financial institutions', *Journal of Finance*, 32, pages 1,251–66.

Schumpeter J A (1942), 'Capitalism, socialism and democracy', New York: Harper and Brothers.

Schweiger I and McGee J S (1961), 'Chicago banking', *Journal of Business*.

Shaffer S and David E (1991), 'Economics of superscale in commercial banking', *Applied Economics*, 23, pages 283–93.

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

Silber W L (1975), 'Financial innovation', Lexington Books, 1975.

Solow R M (1957), 'Technical Change and the Aggregate Production Function', *Review of Economics and Statistics*, 39, pages 312–20.

Tobin J (1984), 'On the efficiency of the financial system', *Lloyds Bank Review*.

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

Tulkens H (1990), 'Non-parametric efficiency analyses in four service activities: retail banking, municipalities, courts and urban transit'. Centre for Operations Research and Econometrics, Universite Catholique de Louvain, *Discussion Paper No. 9050.*

World Bank (1989), 'World development report 1989', Oxford University Press, New York.

Vassiloglou M and Giolias D (1990), 'A study of the relative efficiency of bank branches: an application of Data Envelopment Analysis', *Journal of Operational Research Society*, 41, pages 591–97.

Bank of England Working Paper Series

Publication date in italics

1	Real interest parity, dynamic convergence and the European Monetary System *(June 1992)*	Andrew G Haldane Mahmood Pradhan
2	Testing real interest parity in the European Monetary System *(July 1992)*	Andrew G Haldane Mahmood Pradhan
3	Output, productivity and externalities—the case of banking *(August 1992)*	R J Colwell E P Davis