CHAPTER 3

LITERATURE SURVEY

In the previous chapter, we discussed the theoretical underpinnings of our study. The theories of production, cost and profit have been explained and their application to the study of economies of scale in a firm/an industry is elaborately discussed. In this chapter, we propose to critically examine the important works in the area of economies of scale in the banking industries, both in India and abroad.

Firstly, we have discussed the studies on cost function approach to economies of scale in banking industry. We have focussed on the definition of output and cost variables, on the methodologies and on functional forms used in these studies. Secondly, we reviewed the studies having the profit function approach to the economies of scale.

3.1 Cost Function Studies

3.1.1 The Output in Banking Industry

To examine the relationship between size and cost and the effect of changes in the level of production on the cost per unit of output the measurement of size assumes importance. In the traditional theory of the firm, the concept of size or scale is normally denoted by output of homogeneous product per unit of time. Accordingly, size or scale is also defined some times in terms of assets or number of employees etc. However, it is

argued that within the same industry, the volume of output is the simplest measure of size of the plant and firm.

There are some features, unlike manufacturing industry, peculiar to the banking industry. The physical nature of most of the inputs and the outputs is the same, namely, money and there is not much investment in fixed assets. The question, therefore arises,as to the unit by which the scale or size of operation of the banking firm should be measured. The definition of output of commercial banks continues to be a controversial subject, in particular, because of its importance in the estimation of economies of scale. No general consensus seems to have arisen regarding the appropriate definition. This lack of consensus is reflected in the diversity of measures of output employed in the economies of scale literature. In large part, the disagreement over the appropriate definition of bank output can be attributed to the multiproduct nature of the commercial banks and the subsequent lack of agreement on proxies for both, particular and general measures of lending and non-lending banking services (J.A. Clark, 1984).

The early studies employed an unweighted stock from liability or asset side of the balance sheet, i.e. when a bank is viewed as a collection of liabilities or what produces bank revenue. These studies include Alhadeff(1964), Schweiger and McGee(1961), Horvitz(1963), Grebler and Brigham(1963) and Rangarajan and Mampilly(1972)*.

---

* This being an Indian study, has been dicussed separately later in the chapter.

Alhadeff (1954) used a host of variables while examining the economies of scale in American banking industry, the most important of them being total earning assets (loans + investments). He applied tabular inspection method to relate these variables to operating expenses.

Horvitz (1963) employed the same method and models as used by Alhadeff, but used different source of data and time period. Regarding output measure, Horvitz argued" the essential element of banking is converting the raw material of deposits into loans and investment. A bank, in essence, produces loans and investments. Loans and investments are the banking output most nearly analogous to the product of the manufacturing firm". (pp.4)

Schweiger and McGee (1961) used total deposits as a proxy of the output. Instead of using deposit directly as an independent (size) variable, they defined nine deposit classes (under $ 2 million, $ 2.5 million .....over $ 500 million) and ranked them in order from one to nine. Each bank was then assigned a number according to the deposit class it fell into and the assigned numbers were introduced in lieu of deposits. Further, one more variable, output-mix, was included in the cost function using multiple regression analysis.

Grebler and Brigham (1963) used total assets to measure the bank size as a proxy for output. The dependent variable (cost) was measured as total operating cost per thousand dollars of total assets.

The second approach to output assumed that each of the bank's services was is produced via technically independent production functions,i.e. the bank was viewed as a series of separable Cobb-Douglas production functions. The studies on this approach include Bell and Murphy (1968) and Benston (1965).

Both these studies defined output in terms of what banks do that cause them to incur operating cost. They divided commercial bank output into six relatively homogeneous services: demand deposits, time deposits, real estate loans, instalment loans, business loans and securities. The direct cost of each one was analysed separately. They recognised that from a cost stand point, all these services may be regarded as different products. Thus, a meaningful analysis of bank costs should consider each factor separately. The indirect costs were divided into administrative, business developments and occupancy. The cost functions were estimated for each, using multiple regression analysis in both the studies.

When separate cost functions were estimated for each output, it was implied that the marginal cost of producing one output was independent of another output, i.e. it was assumed that bank production was non-joint and producing one output was independent of the costs of producing other outputs.

The third approach attempted to construct a weighted index of bank output using informations from the income statement as well as the balance sheet.This approach has been used by

Greenbaum(1967), Edgar, Hatch and Lewis(1971) and Benston, Hanweck and Hamphrey(1982).

Stuart I. Greenbaum (1967) computed an average yield, bj, for each of the sixteen types of earning assets by regressing :-

$$Yi/Ai = bo + \sum_{j=1}^{n} bj\,[Zij/Ai] + \sum_{j=n+1}^{m} bj\,Xik + Ui$$

where

Yi	=	ith bank's gross operating income directly attributable to lending
Ai	=	Total assets of ith bank
Zij	=	j types of earning assets of ith bank
Xik	=	k banking structure variable of ith bank
K	=	Population in bank's area, dummy variable for state,number of banks's area.

The computed, yields bjs, were multiplied by the year end amount of each of the sixteen types of assets and the product added to non-lending output to determine the output of the `i'th bank. The reason for including non-lending output was that it also contributed to the community's welfare and comprised as much as ten percent of a bank's operating income. Thus, he suggested that the difference between total operating income and income generated by earning assets be added to the weighted index of revenue from lending activities to capture output from non-lending activities.

Edgar, Hatch and Lewis (1971) formulated a different concept of bank output. As EHL stated, "the basic economic function of a financial intermediary in general and bank in particular is to intermediate between the borrowers and lender of funds, giving services to both. Banks act as financial intermediaries by

collecting funds as deposits and transmitting these to borrowers, principally, by allowing accounts to be overdrawn. Consequently, banks can be viewed as hiring men and materials and organising these in conjuction with capital equipments in order to produce services to the users of bank accounts irrespective of whether these accounts are in credit or in debit." (pp.20)

They classified the services produced by banks into three headings -

a. Services provided when accepting deposits, allowing funds to be withdrawn and negotiating and servicing loans.

b. Services related to the operating of community's payment mechanism, including the provision of foreign exchange.

c. Miscellaneous services connected with the ownership of a bank account - business information, trade advice and travel services.

In order to produce these services, banks hire men, purchase materials and use-up capital equipments. These activities generate operating costs which are a function of the services produced.

Regarding output-mix variables used in the cost function, EHL argued, "it is feasible to consider economies of scale in a multi product industry, but, it is essential in a time series analysis to be sure that the product mix has remained constant over time. If this cannot be assumed, then we are faced with isolating, part of the product-mix in order to produce a cost output relationship for a given product."

The measure of output used takes the form -

Value of deposits + advances deflated by implicit GNP	X	Index of bank account turnover

The term in the brackets is thought to be indicative of the number and size of accounts as number of accounts and deflated deposits/advances were closely correlated. The implicit deflation for GNP was used for the purpose of variation in the purchasing power of money. The measure of the turn-over of accounts for each bank was an index of the ratio of the value of debits for all banks to the value of deposits plus advances for all banks.By combining the two terms given above, the measure of output was regarded as an approximation to the value of debits for each bank,deflated for purchasing power.

Benston, Hanweck and Humphrey (1982) measured the output in terms of what banks do that cause operating expenses to be incurred. BHH argued that the appropriate way to aggregate quantities of bank output was to use a statistical index number formula which approximated the results that would have been obtained by a flexible aggregate function. BHH selected the divisia multilatural statistical index number developed by Caves, Cristensen and Diewert(1982).

In recent years we have a few studies by Gilligan and Smirlock (1984), Gilligan, Smirlock and Marshall (1984) and L.S. Mester (1987), which explicitly take into account the multiproduct nature of the firm by using a theory developed by Bamoul, Panzar and Willing [1982]. It gives attention to the issue of product

specific and over all scale economies and cost saving realised from joint production, i.e. economies of scope. These studies are discussed in detail under the title "Functional Form".

3.1.2 The Cost in Banking Industry

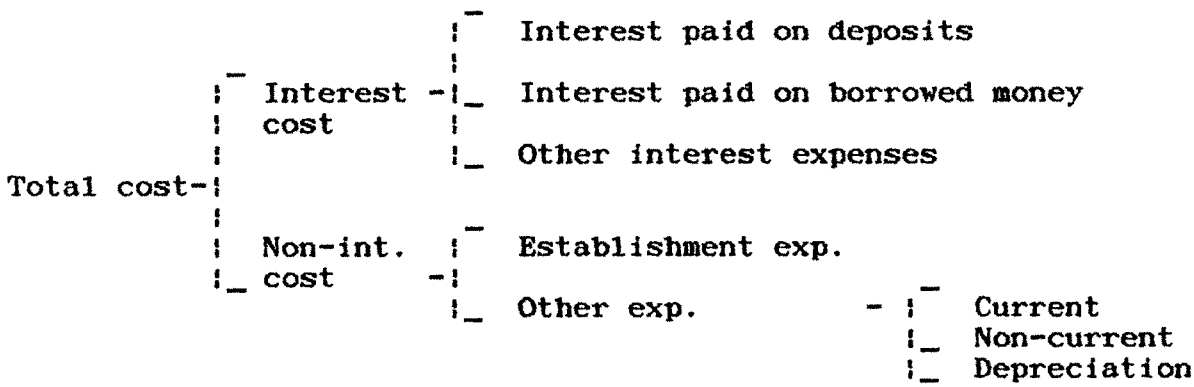
Cost refers to the economic or opportunity cost [private and social] of operating a firm, during a time period at a given rate of output. Measurement of cost usually is complicated by the presence of externalities that are difficult to quantify and by a lack of correspondence between the accounting data recorded by firms and opportunity costs. Fortunately, these problems are much less severe for financial institutions than for almost any other industry. There are relatively low externalities in the production of financial services "No smoke pollutes the air or chemicals the rivers." (Benston, 1972).

To estimate cost function, we are required to know various compositions of cost in commercial banks. The major elements of costs in commercial banks can be classified [Exhibit-I] into two groups :-

a. Interest Cost
b. Operating Cost

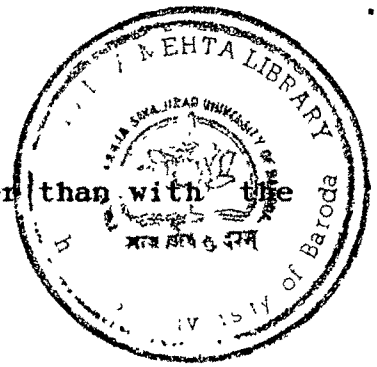
Interest Cost is an indicator of volume of business as it is more or less in proportion to volume of business. Operating cost is the cost incurred in order to provide services to different customers including borrowers and depositors. Operating cost includes wage-salary cost and all other expenses (like printing, postage, rent etc.) excluding interest cost.

Exhibit - I



In majority of the surveyed studies on economies of scale, cost was defined as total operating cost which excluded interest cost from the total cost. Though, in some studies, the interest costs were included in total costs when a more general definition of scale economies was examined but, the inclusion of interest costs resulted in the elimination or much reduced operating cost scale economies.It was argued in Rangarajan and Mampilly (1972) that it was the operating cost which might depict economies of scale in banking industry.

The argument behind omitting interest cost was well explained by Edgar, Hatch and Lewis (1971): "Deposit is not regarded as a factor of production. Banks employ men, materials and capital equipments to produce services for both, lenders and borrowers. In return for the services produced, banks receive income which takes the form of an interest rate differential, foreign exchange differential and miscellaneous commission and service charges. Secondly, the interest rates paid by individual banks over time are associated with the growth of the banking industry as a whole



and the operation of central banking policy rather than with the growth of a particular banking firm."

3.1.3 The Functional Form

Most of the earlier studies focussed on individual bank functions using Cobb-Douglas functional form. However, this method has serveral limitations. Firstly, it requires product specific cost data which is not easily accessible to the researchers. Secondly, it ignores the total cost of banking operations. Thirdly, it does not permit the average cost to be U shaped and thus, fails to estimate the optimum size of a bank. This function has a constant elasticitiy and does not allow the cost curve to turn up.

Used in recent years, the translog cost functional form is considered to be an improvement over the simple logarithmic (Cobb-Douglas) function. It permits the estimation of U shaped cost curves and the measurement of scale economies and branch economies. The first study found in the literature using translog cost function was by Benston, Hanweck and Humphrey (1982). The functional form used was:

$$\begin{aligned} \log TC = {} & a + bq \log Q + bqq\ 1/2\ [\log Q]^2 + cb \log B \\ & + cbb\ [\log B]^2 + dbq \log B \log Q + ea \log A \\ & + eaa\ [\log A]^2 + faq \log A \log Q + gh\ H \\ & + mhb\ H \log B + \sum_j nj\ Pj + \sum_j ojq \log Pj \log Q \\ & + \sum_j \sum_k rjk\ 1/2\ [\log Pj \log Pk] \\ & (j,k = L,K) \end{aligned}$$

where

TC = Operating Cost
Q = Bank Output
B = Number of Branches
A = Average Size of Deposits/Loan Accounts
H = Dummy Variable (affiliated/non affiliated)
Pj = Price of Labour and Capital

The two theoretical restrictions imposed on the cost equation were symmetry and input price homogenity.

The measure of scale economies (SCE), i.e. the percentage change in total operating cost associated with a percentage change in output was :-

$$SCE = \frac{\log C}{\log Q} = bq + bqq \log Q + dbq \log B + faq \log A + \sum_j ojq \log Pj$$

The economies of scale estimates from a translog cost function are the same as those obtained from its dual production function if regulations on duality conditions are met in the data.

The common criticism of the translog cost function is the existence of severe multi-collinearity which results from the estimation of the interactive variables. A common suggested solution to the multi- collinearity problem is to increase the sample size. By jointly estimating the total cost and factor share equations (share of costs accounted for by an input), the multi-collinearity problem is reduced since the joint estimation procedure effectively doubles the number of observations. Nevertheless working out the optimum size remains a complicated problem in using translog cost function.

3.2 Profit Function Studies

3.2.1 Conventional Profit Function

Donald J. Mullineaux (1978) was the first to use profit function for estimating economies of scale in banking industry. His study was inspired by the theoretical foundations developed by Mc Fadden (1966) and Lau (1969), who developed the theory of profit function for competitive and non-competitive firms and examined its relationship with the production function.

Mullineaux (1978) summed up the following properties of the profit function which made it a more desirable approach than cost function for studying economies of scale in banking (pp. 261) -

1. The level of output is not a variable in the profit function.Therefore,we can avoid the difficulties involved with the output definition, usually faced in the cost function.

2. Bank cost studies consider only technical efficiency, whereas, the profit function, since it considers prices, relates to the more complete concepts of economic efficiency.

Mullineaux (1978) tested commercial banking economies of scale and other organizational efficiency using a hybrid profit function which was transcedental logarithmic (quadratic in log) in labour input prices and Cobb-Douglas in the prices of outputs and other inputs and the quantities of fixed factors of production. This study was for 922 American banks drawn from 12 Federal Reserve Districts for two years, 1971 and 1972.

The profit function was :-

$$\log p = ao + \sum_{i=1}^{m} ai \log Pi + \sum_{j=1}^{n} bj \log Qj + \sum_{m=1}^{t} Sm \log Vm$$

$$+ 1/2\left[\sum_{m=1}^{t} \sum_{j=1}^{t} hmj \log Vm \log Vj + \sum_{k=1}^{w} ck\, Zk \right]$$

The dependent variable bank profit (p) was measured as operating revenue minus operating expenses net of occupancy costs. Mullineux estimated profit function using several different measures of profits also but none of the significant conclusions of the study were altered.

The independent variables, output prices (Pi; i=1.....m), included real estate loan rate, consumer instalment loan rate, commercial and agriculture loan rate and safe deposit rental fee. Input prices (Qj and Vm) included officers wage rate, employee wage rate, demand deposit rate, saving-term deposit rate, certificate of deposits rate and computer hardware rental rate. All loan rates were measured as the ratio of annual interest income plus fees to the average volume of loans outstanding. Deposit prices were measured as effective yields ratio of interest payments to deposit volume. This data suffered from the limitation of aggregation, for which, Mullineaux has given a clarification, " because of the balance-sheet constraint, banks can not freely choose every element of the balance-sheet".

The quantities of fixed factors (Zk; k=1.....w) included number of bank offices of different types and a proxy variable for office size. Proxy variable used for size was the ratio of

furniture and equipment expenses to the number of offices, included to control for the impact of office size on bank profits.

The important findings of this study were :a) banking industry is characterised by competition and increasing returns to scale b) the magnitude of scale economies indicated by profit function exceeded that of most cost function estimates.

The theoretical formulation of the profit function assumes that firms are price takers in output and input markets. An application to commercial banks would appear to require that this condition atleast be approximately satisfied for the banking industry (Mullineaux,1978; pp-262). Lau (1969) has suggested that the profit function can be used to test whether a firm is a price taker in a given market. A finding that output prices make no significant contribution to the empirical explanation of bank profit is consistent with the hypothesis that firms are not price takers in any of the market for their products or services. A firm may operate competitively for a subset of their products, in which case, a subset of output prices would appear in the profit function. A finding consistent with price setting behaviours in any output market suggests that variables reflecting the external structure of a bank's market and be included in the profit function.

Using the profit function, an independent research examined whether banks located in the money centres were more profitable than others. He used the data of the largest 200 commercial

banks in the US for the year 1977. The results indicated that larger banks were less economically efficient irrespective of the location. His specifications were simple log-linear regressions between return on assets and each of the different 'size' variables separately viz., total assets, total deposits and total loans, with a dummy variable representing money center bank.

3.2.2 **Risk Adjusted Profit Function**

Warapatr (1983) used the same function as used by Mullineaux (1978) to study the economies of scale in banking after adjusting for the risk factors. He included a few ratios as variables with the price variable as input for measuring the bank's overall business and financial risk. As Warapatr (1983) stated; "in the context of the profit function theory, an individual commercial bank is viewed as an economic unit whose goal is to maximise profits. To achieve this objective banks take deposits and borrow funds and convert them into various types of earning assets. Banks also provide other services, viz., safe deposit vaults, etc. and charge fees. In the regular conduct of their business, banks expose themselves to many kinds of risks. It is useful to examine a typical bank's balance-sheet and to focus attention on the major items which have significant effects on bank risk and profits" (pp. 59).

Bank, being a financial intermediary, Warapatr (1983) gave equal attention to both the sides of the balance-sheet i.e. assets and liabilities. From the assets side, the aggregate loans to assets ratio was used for the risk carried by the bank, based on the

general hypothesis about the relationship between risk and rate of return. It was observed "given the characteristics and distribution of its liabilities, a bank attempts to structure its portfolio of assets in such a manner as to yield the greatest return. The higher the percentage of the bank's total assets held in earning assets, the greater is the return. The higher the percentage of the bank's total assets held in earning assets, the greater is its expected profit, ceteris peribus".

He further stated, "as for the two types of earning assets, namely, investment and loans, the risk of loss is typically greater for loans than investments. Yet a major portion of the bank's earning assets is usually held in the form of loans because they are potentially the greatest source of income to the bank. "In this context, the loans to assets ratio is viewed as a variable cost whose effect will be reflected in the level of the banks actual profits".

From the liabilities side of the bank's balance-sheet, the ratio of equity capital to total assets and the ratio of borrowed funds to total liabilities were taken as two variable risk factors which influence the actual profit. As for equity capital, the ratio of equity capital to total assets, a popular ratio, used by bank regulators as measure of bank soundness, and for borrowed funds, the ratio of borrowed funds to total liabilities, as the bank relies less on traditional deposits and more on borrowed sources from the money market and so the bank is subjected to greater interest risk, were chosen as risk factors. All these three factors were included as risk factors. The risk adjusted

profit function model for commercial banks' estimated by Warapatr (1983) was Cobb-Douglas in character -

$$\log \text{Profit} = ao + \sum_{i=1}^{m} ai \log Pi + \sum_{j=1}^{n} bj \log Qj + \sum_{k=1}^{w} ck \log Zk$$

where

Pi = Bank output prices (i = 1.....m)
Qj = Variable input prices (1.....n) (including risk factors)
Zk = Quantities of fixed factors (k = 1.....w)

3.3 Studies on Indian Banking Industry

There are very few studies on economies of scale in Indian banking industry. Rangarajan and Mampilly (1972) made the first attempt in 1972. They established relationship between operating expenses (net of interest costs) and total deposits. The other explanatory variables were ratios of composition of deposits (output-mix), number of branches and salary ratio. The quadratic functional form was used to estimate the relationship. The cross-sectional data of the top 30 banks for the years 1967 and 1968 was provided by the Banking Commission. The deposit size of sample banks varied from Rs. 11 crores to Rs. 830 crores. Rangarajan and Mampilly (1972) first tried to introduce all the bank activities separately but did not get satisfactory results. Hence, deposits were selected as a proxy for total output. They estimated the least cost bank to have deposit size of Rs.300 crores.

The limitations of this study were : First by total deposit was not an appropriate measure of total output, and secondly, Two, instead of Three deposit ratios (output-mix variables) could have been used to avail the degrees of freedom.

In other study on economies of scale in Indian banking by the Study Group on Banking Costs (1971), a weighted composite index of output and Cobb-Douglas functional form were used. The study concluded that there was indefinite scope for expansion of the output, as elasticity coefficients were less than one. Though the output measure was superior than that of the previous study, it could be used only with accessibility to detail data at the branch level. Further, the study could not suggest the optimum bank size.

The third study (Banks Since Nationalisation, 1981) was carried out by the Economic Research Division of Birla Institute of Scientific Research, New Delhi. Three different measures of output were taken, viz. total income, total deposits and working funds. It was a cross-sectional study of three bank groups - 14 nationalised banks, 13 selected private sector banks and both groups combined together with the State Bank of India group. The Cobb-Douglas functional form was used. The empirical results suggested that irrespective of the output measure, there were not significant economies or diseconomies of scale for any of the years except for the combined group, where in certain years there were significant economies of scale. Thus, it was recommended that there is scope for the smaller banks to grow

atleast till they reached the size of the larger banks, without affecting the cost efficiency adversely.

We have not come across any study having a profit function approach for examining the economies of scale in banking industry in India.

3.4 REFERENCES

1. Alhadeff,D.J. (1954), Monopoly and Competition in Banking,University of California, Barkeley,1954.

2. Bamoul, W.J., Panzar, J.C. and Willing, R.D. (1982), Contestable Markets and the Theory of Industry Structure, New York, USA,1982.

3. Banks Since Nationalisation (1981), Economic Research Division, Birla Institute of Scientific Research, New Delhi, 1981, pp. 79-104.

4. Bell, F.W. and Murphy, N.B. (1968), "Cost in Commercial Banking: A Quantitative Analysis of Bank Behaviour and its Implications to Bank Regulations", Research Paper No. 41, Federal Reserve Bank of Boston, USA.

5. Benston, George J. (1965), "Branch Banking and Economies of Scale", Journal of Finance, XX, May 1965, pp. 312-32.

6. Benston, G.J. (1972), "Economies of Scale of Financial Institutions", Journal of Money, Credit and Banking, 4 May 1972, pp.312-41.

7. Benston G.J., Hanweck, G.A. and Humphrey D.B. (1982), "Scale Economies in Banking: A Restructuring and Reassessment", Journal of Money, Credit and Banking, Vol. 14, No. 14, Nov.1982, Part II, pp. 435-456.

8. Caves, D.W., Critensen, L.R. and Diewert, W.E. (1982), "Multilateral Comparision of Output, Input and Productivity Using Superlative Index Numbers", Economic Journal, 92, March 1982, pp. 73-86.

9. Clark, J.A.(1984), "Estimation of Economies of Scale in Banking using Generalised Functional Form", Journal of Money, Credit and Banking, Vol.16, No. 1, Feb. 1984, pp.53-68.

10. Edgar, R.J., Hatch, J.H. and Lewis, M.K.(1971), "Economies of Scale in Australian Banking: 1947-68", The Economic Record, March 1971, pp.17-37.

11. Gilligan, T.W. and Smirlock, M.L. (1984), "An Empirical Study of Joint Production and Scale Economies in Commercial Banking", Journal of Banking and Finance, 8, March 1984, pp. 67-77.

12. Gilligan, T.W., Smirlock, M.L. and Marshall, W. (1984), "Scale and Scope Economies in the Multiproduct Banking Firm", Journal of Monetary Economic+s, 13, May 1984, pp. 393-405.

13. Grebler, L. and Brigham, E.F. (1963), "Savings and Mortage Markets in California", Pasadena: California Savings and Loan League, 1963.

14. Greenbaum,S.I. (1967), "Competititon and Efficiency in the Banking System: Empirical Research and Its Implications", Journal of Political Economy, 75, August 1967, pp. 461-481.

15. Horvitz, P.M. (1963), "Economies of Scale in Banking", in Private Financial Institutions, Commission on Money and Credit, Pranctice Hall, New Jersey, pp. 1-54.

16. Lau, L.J. (1969), "Some Applications of Profit Function", Memorandum No. 86-A, Research Center in Economic Growth, Stanford University, USA, 1969.

17. McFadden, D. (1966), "Cost, Revenue and Profit Functions: A Cursory Review", Working Paper No. 86, Institute of Business and Economic Research, Univerisity of California, 1966.

18. Master, L.J. (1987), "A Multiproduct Cost Study of Savings and Loans", The Journal of Finance, Vol.XLII, No. 2, June 1987, pp. 423-445.

19. Mullineaux, D.J. (1978), "Economies of Scale and Organisational Efficiency in Banking: A Profit Function Approach", Journal of Finance, Vol. XXXIII, No. 1, March 1978, pp. 259-280.

20. Rangarajan C. and Mampilly, P. (1972), "Economies of Scale in Banking", Technical Studies prepared for the Banking Commission, Reserve Bank of India, Vol. II, 1972.

21. Report of the Study Group on Banking Costs (1971), Banking Commission, Government of India, Bombay, 1971.

22. Schweiger, I and McGee, J.S. (1961), "Chicago Banking", Journal of Business, University of Chicago, XXXIV, July 1961, pp. 203-216.

23. Warapatr,T.(1983),"The Risk Adjusted Profit Function : Measurement of Economies of Scale and Efficiency in Commercial Banks", Ph.D Thesis, University of Illinois , USA,1983.