CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The objective of this chapter is to analyse the different concept, methods and variables that is being employed to analyse efficiency in banks with the aim to justify the use of Data Envelopment Analysis (DEA) model for this study and to obtain the most appropriate variables that is to be used in this study.

The flow of the literature review will is as presented in Figure 2.1 below.

Figure: 2.1
Literature Review Framework

Efficiency Measurement Concepts 1. Input Orientation Measure 2. Output Orientation Measure
Efficiency Types 1. Technical and Allocative Efficiency 2. Profit Efficiency 3. Cost Efficiency 4. X- efficiency
Measuring Techniques 1. Parametric Approach a. Stochastic Frontier Approach b. Thick Frontier Approach c. Distribution Free Approach 2. Non Parametric Approach a. Free Disposable Hull Analysis b. Data Envelopment Analysis
Input and Output Variables 1. Production Approach 2. Intermediation Approach 3. Bank Specific Environmental Variable

2.2 Efficiency Measurement Concepts

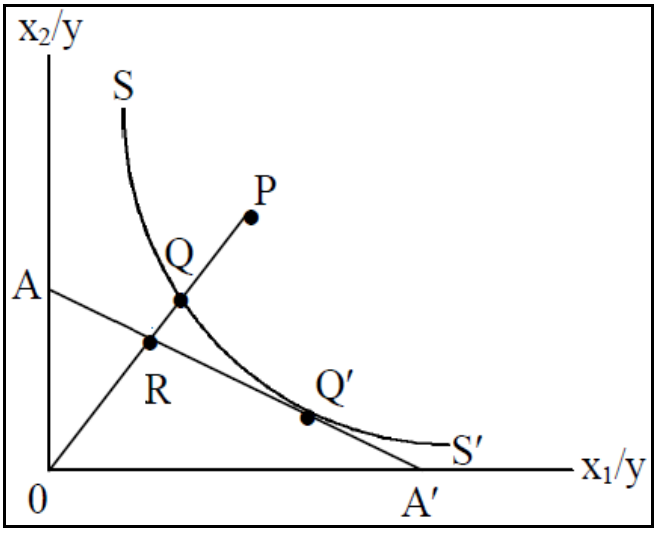
The modern efficiency measurement was being introduced by Farrell (1957) which he had defined a simple measure of firm efficiency that is able to account for multiple inputs. He proposed that the efficiency of a firm consists of two components consist of technical efficiency and allocative efficiency. Both of this combined measure will provide the measure of total economic efficiency. This will be discussed further in the types of efficiency section below.

2.2.1 Input Orientation Measure

The input orientation is to obtain the input quantities that are able to be proportionally reduced without changing the output quantities produced. Farrell (1957) had introduced the input oriented measure and based on Figure 2.2, it provides an illustration of the technical efficiency, cost efficiency and allocative efficiency with the consideration of the firm uses two factors of input to produce a single output.

The input orientation can also be termed as the input quantities that can be proportionally reduced without changing the output quantities produced. The two inputs are represented by (X1 and X2) which is used to produce output Y, with constant return to scale and (SS') represents the isoquant of a fully efficient firm.

Figure: 2.2
Input Orientation - Technical and Allocative Efficiency

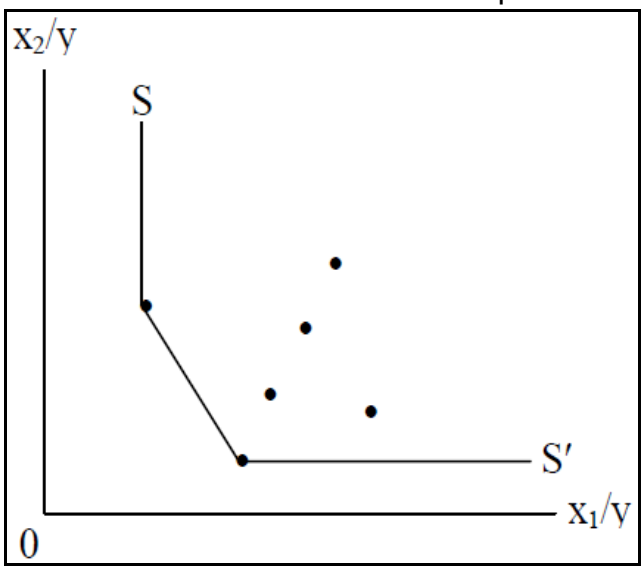


Based on Figure 2.2, the technical efficiency (TE) is measured by the ratio of TE = 0Q/0P which is equals to 1 – 0Q/0P. It will produce a value between zero and one which will provide an indicator of the degree of technical inefficiency of the firm. The point Q is technically efficient as it lies on the isoquant represented by (SS').

The line (AA') represents the allocative efficient line whereby the allocative efficiency (AE) of the firm operating at P is defined as the ratio of AE = 0R/0Q. The point RQ represents the reduction in the production cost that will happen if the production were to occur at the point Q'. The total economic efficiency (EE) is defined as the ratio of EE = 0R/0P where the distance of RP can be interpreted in term of cost reduction. The product of technical and allocative efficiency provides the overall economic efficiency which is represented by TE X AE = (0Q/0P) X (0R/0Q) = (0R/0P) = EE. All the three measurement, TE, AE and EE are bounded by zero to one.

Farrell (1957) study on the 48 continental states of U.S. on the agriculture data have indicated that the efficient isoquant must be estimated from the sample data by using a non-parametric piecewise-linear convex isoquant constructed so that no observed point is located at the left or bottom as illustrated Figure 2.3 or a parametric function such as Cobb-Douglas form whereby the sample data having no observed point is located to the left or below it.

Figure 2.3
Piecewise Linear Convex Isoquant

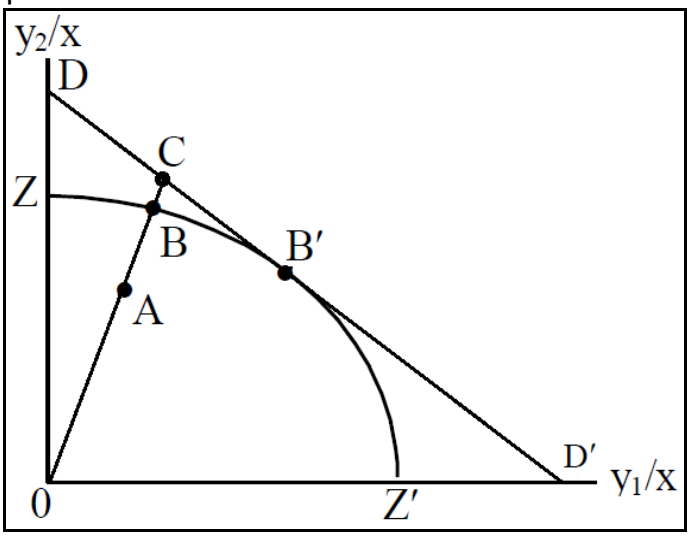


Jaffry et al. (2007) had examined the changes in productivity and technical efficiency levels for banking sectors in India, Pakistan and Bangladesh for the period of 1993 to 2001 and it is indicated that the output orientation will be more suitable to be used given the objectives of developing country banking. Output orientation refers to the emphasis on the equi-proportionate augmentation of outputs, within the context of a given level of inputs.

2.2.2 Output Orientation Measure

As opposed to input orientation, the output orientation can also be termed as the output quantities that can be proportionally expanded without changing the input quantities used. Figure 2.4 illustrates the technical and allocative efficiencies from an output orientation.

Figure 2.4
Output Orientation - Technical and Allocative Efficiency



Based on Farrell (1957) the measure technical efficiency ratio can be derived as TE = 0A/0B. The allocative efficiency can be derived based on the isorevenue line (DD') which is based on the price information whereby the allocative efficiency ratio is AE = 0B/0C.

The total economic efficiency is the combination of the technical and allocative efficiency which is derived as EE = (0A/0C) = (0A/0B) X (0B/0C) = TE X AE. These three measures are again bounded by zero to one.

2.3 Types of Efficiency

Bank efficiency studies ranges from technical and allocative efficiency, profit efficiency, cost efficiency and X-efficiency. Berger and Humphrey (1997) had survey 130 studies that apply frontier efficiency analysis to financial institutions in 21 countries. Out of the 130 studies, only 14 of the studies use efficiency of revenue and/or profit. Frederic Taylor (1911) had defined efficiency as putting an end to the unnecessary of forests, waterpower, coal, soil and human effort. In other words, efficiency can be termed as optimal usage of resources by maximising its usefulness.

Traditionally, efficiency of an organisation performance is relied on ratios and index analysis (Akhigbe & McNulty, 2003). Ratios provide a single dimensional perspective and fail to provide a specific reason for the performance result (Sherman & Gold, 1985).

The subtopics below will provide an overview of the different types of efficiency.

2.3.1 Technical and Allocative Efficiency

Farrell (1957) had termed technical efficiency as the ability of a firm to obtain maximum output from a given set of inputs.

The technical efficiency can be divided into two categories namely the pure technical efficiency and allocative efficiency. Pure technical efficiency is

defined as the excess usage of input level at a given output level due to the management of the operations in a firm. Thus, pure technical efficiency is more towards the ability of the management to maximize the production of output at a given level of input or to minimize the input used to produce a given level of output.

Gilbert and Wilson (1998) studied on the technical progress in terms of the productivity changes of Korean commercial banks from 1980 to 1994 and based on the findings, it suggest that bank positively react to privatization and deregulation with large increase in productivity and also substantial changes in the mix of inputs and outputs.

Chen and Yeah (2000) had studied on technical efficiency and its relation with private ownership using Data Envelopment Analysis on 34 commercial banks in Taiwan and it was found that there is a significant lower level of technical efficiency in public owned banks compared to privately owned banks.

Farrell (1957) has termed allocative efficiency as the ability of a firm to use the inputs in optimum proportions given their respective prices. Allocative efficiency can also be termed resource allocation to suit best in producing an output.

Fukuyama and Weber (2002) had developed a new measure of output allocative efficiency where the ratios of direct and indirect input quasi distance functions are obtained as the measure of output allocative efficiency. Using

this method, the study on Japanese banks in the period of 1992 to 1996, the study findings is that Japanese banks experienced productivity decline averaging two percent per year. Besides that, they have found that by using the revenue maximizing output mix, the banks could have used only around seventy eight to ninety three percent of actual inputs to achieve the same result.

2.3.2 Profit Efficiency

Profit efficiency measurement takes into consideration for both cost and revenue efficiency (Avkiran, 2004) whereby the profit is obtained from revenue generated deducting the cost of production. The profit efficiency relates to the banks maximizing the profit for a combination of outputs.

Berger and DeYoung (2001) had studied on the effects of geographical expansion on bank efficiency on US banks from the period of 1993 to 1998 in terms of cost and profit. Their findings is that there is both positive and negative linkages between geographic scope and bank efficiency where it is highlighted that the parent organization is able to transfer the skills, policies and practices to further improve the efficiency of its affiliates. Besides that, the study found that smaller banks appear to be relatively inefficient as compared to the larger banks.

Akhigbe and McNully (2003) had studied on the profit efficiency of small US commercial banks by using the single frontier approach. Factors taken into their study are like the market structure, size and the environmental factors.

The study found that smaller banks appear to be more profit efficient than larger banks. However, it was also noted that smaller banks in metropolitan areas tend to be least efficient as compared to banks located in other area.

Williams and Nguyen (2005) had perform a analysis on the cost and profit efficiency of commercial banks in Indonesia, Korea, Malaysia, Philippines and Thailand for the period of 1990 to 2003 where the study takes into account the technical change, productivity and ownership structure. The findings shows that bank privatisation associate with higher profit efficiency performance and will lead to an increase in the overall efficiency and productivity of the banking system for the countries in its study.

2.3.3 Cost Efficiency

Cost efficiency focuses on minimizing the cost of production by the banks while still producing the same amount and combination of outputs.

Berger et al. (1987) had made a study by using functional cost analysis on bank data. The findings shows that large bank experience diseconomies of scope and there are no substantial scope economies were found in savings and loans.

Berger and Humphrey (1991) had examined banks in U.S. in 1984. The findings show that the cost differential exists due to firm specific inefficiencies however it does not affect the analysis of scale and product mix of the study.

Battese and Coelli (1993) by using the intermediation approach and estimating the stochastic cost frontier by adopting the maximum likelihood procedure had found that the average sized banks exhibit constant returns to scales however study done by Tsionas et al. (2003) had found that smaller bank are in their full level of efficiency.

Wheelock and Wilson (2001) had studied on the cost efficiency of U.S. commercial banks by using the intermediation approach for 1985, 1989 and 1994 and found that the banks gain from potential economies of scale up to the level of $300 to $500 million of asset since 1985.

2.3.4 X-efficiency

The definition of X-efficiency is how efficiently the inputs are being used by the management (Mester, 1996).

Goldberg and Rai (1996) studied on the relationship between X-efficiency, concentration and performance of European banks from the period of 1988 to 1991 and based on their findings, there are no positive relationship between concentration and profitability.

Berger et al. (1997) had analyzed the X-efficiency of a 760 branches of US commercial banks in the period of 1989 to 1991 where the study examines the effect of merger and acquisition on bank efficiency as well as to compare between the distributional free approach and thick frontier approach in

investigating X-efficiency. The study found that there is a positive relationship between number of branches and cost efficiency but not profit efficiency.

Garden and Ralston (1999) studied on the efficiency effects of Australian credit union merger from the period of 1992 to 1997 on sixteen credit union mergers that happen in 1993 to 1994. Based on their study, on average the mergers do not result in an increase in x-efficiency or allocative efficiency relative to other credit unions.

Cavallo and Rossi (2001) had made a study on banks in six European countries from the period of 1992 to 1997 on the output efficiency in terms of scale, scope and X-inefficiency. Their findings are that most banks exhibit economies of scale regardless of production scale and types of bank organization however the scope economies are found across all outputs is more prominent to larger banks.

Kwan (2006) had investigated on cost efficiency of Hong Kong banks where the findings are that the X-efficiency has been declining over time indicating the banks were operating closer to the cost frontier than previous years. The study found that the average large banks was found to be less efficient than the average small banks however it was indicated that the size effect appears to be related to differences in portfolio characteristics among different size banks.

2.4 Measuring Techniques

In terms of the approach used to evaluate the bank efficiency can be categorised into parametric approach and non parametric approach. The more well known parametric technique consists of Stochastic Frontier Approach (SFA), Thick Frontier Approach (TFA) and the Distribution Free Approach (DFA) while the non-parametric techniques consist of Free Disposal Hull Analysis (FDH) and Data Envelopment Analysis (DEA).

Most literature relating to efficiency measurement is either based on parametric approach or non-parametric approach. Berger and Humphrey (1997) had survey 130 studies that apply frontier efficiency analysis to financial institutions in 21 countries and based on the studies surveyed, 69 employs non-parametric approach and 60 employs parametric approach. The result based on the survey also show that the parametric and non-parametric approach produces an average efficiency score of about 80% for financial institutions relative to best practice banks with the non-parametric technique generally giving a slightly lower efficiency but greater dispersed estimates. The authors also found that the two methods often rank the same set of institutions very differently however it was pointed out that the assumptions made and used attribute to this inconsistency.

The sections below provide a brief overview on the different measurement approach.

2.4.1 Parametric Approach

Measuring efficiency by way of parametric approach focuses on the production approach. Berger and Mester (1997) introduce a decomposition of total cost changes into a portion due to a change in business condition and a portion to changes in productivity whereby the productivity change is then decomposed into the change in best practice and change in inefficiency components.

Among the well known methods used under the parametric approach are Stochastic Frontier Approach, Thick Frontier Approach and Distribution Free Approach which will be briefly discussed in the sub sections below.

Stochastic Frontier Approach

Stochastic Frontier Approach (SFA) origins from the concept of cost frontier analysis whereby this technique uses the cost frontier to the maximum amount of cost that the bank is able to reduce in order to produce the same amount of output. This method allows estimation of standard errors and tests of hypotheses. The SFA was introduced by Aigner et al. (1977), and Meeusen and Van de Broeck (1977).

This approach expresses a functional form for economic decision such as profit, cost or production relationship based on the different inputs, outputs and environmental factors (Berger & Humphrey, 1997).

The stochastic frontier approach employs the measurement error and statistical noise and allows the separation of the deviation into inefficiency and noise components (Bauer, 1990).

The main criticism on this method is that is not a priori justification for selection of any particular distribution. Despite of the problem measuring the levels of efficiency, one positive aspect of this approach is that it always ranks the efficiencies of the firms in the same order as their cost function residuals no matter which specific distributional assumptions are imposed (Bauer et al., 1998).

Thick Frontier Approach

Thick Frontier Approach (TFA) uses the same functional form for the frontier cost functional as SFA but is based on a regression which is estimated using only ostensibly best performers in the data set as opposed to those in the lowest average-cost quartile for their size class (Bauer et al., 1998). The TFA was introduced by Berger and Humphrey (1991) whereby the use of this approach is to use a set of parameter estimates and based on the parameters, to obtain the best-practice cost for all of the firms in the data set.

The TFA gives an estimate of efficiency differences between the best and worst quartile to indicate the general level of overall efficiency but it doesn't provide point estimates of efficiency for all individual firms (Bauer et al., 1998).

Jondrow et al. (1982) had proposed a method of seperating the error term of the stochastic frontier model into 2 components whereby this method allows one to estimate the level of technical efficiency for each observation the the sample and largely removes the deterministic frontier for which technical inefficiency is reaily meaasure for each observation.

Distribution Free Approach

The Distribution Free Approach (DFA) is introduced by Berger et al. (1993) whereby the DFA assumes a functional form for the cost or profit frontier but without any specific distribution on inefficiency is required.

Bauer et al. (1982) stated that DFA separates inefficiencies from random error in a different way. In this context, the inefficiency for each firm is being measured by the difference between its average residual from the estimated cost or profit function and to the firm on the efficiency frontier. As the random terms tends to cancel off over time, the explanatory variables are allowed to vary from based on the changes in period to reflect the changes in the environment.

The difference between DFA and the other frontier approach is that it doesn't impose a specific shape on the distribution efficiency as how SFA requires. Besides that it doesn't not impose that deviation within one group of firm are all random error and deviation between groups are all inefficiencies as how TFA does. However, the DFA is similar to SFA and TFA approach and the

efficiency ranks is expected to highly correlated with SFA and TFA ranks (Bauer et al., 1998).

2.4.2 Non-Parametric Approach

The non-parametric approach is an input orientated model which is widely used to measure technical efficiency. The non-parametric approach does not require a priori function specification of unknown technology (Favero & Papi, 1995)

A major drawback of non-parametric approach is that it is assumed to have no measurement error in constructing the frontier, no luck that temporarily gives a decision making unit better measured performance one year from the next and no inaccuracies created by accounting rules that would make measure output and inputs deviate from economic outputs and inputs (Berger & Humphrey, 1997).

The common non-parametric approach techniques are the Free Disposal Hull and the Data Envelopment Analysis. The Free Disposal Hull will be briefly discussed below. As this study is towards Data Envelopment Analysis, a more detailed review will be discussed in the sub section below.

<u>Free Disposal Hull Analysis</u>

The Free Disposal Hull (FDH) model was introduced by Deprins et al. (1984) based on a representation of the production technology given by observed

production plan imposing strong disposability of inputs and outputs but without convexity. It can be defined as the smallest free disposal set containing all observation in a sample production unit (Park et al., 2000).

Tulkens (1993) study based on the monthly data of urban transit firm over a period of 12 years, the FDH methodology is an extended to a sequential treatment of time series that supplements the efficiency estimation with a measure of technical progress.

Based on the FDH framework, it is possible to rank the efficiency of producers by comparing each individual performance with a production possibility frontier. The highest possible level of output for a given level of input can also be derived by observing the production possibility frontier.

Data Envelopment Analysis

Data Envelopment Analysis (DEA) is a linear programming based technique for evaluating the performance in productive units whereby this method is able to handle multiple inputs and multiple outputs as opposed to the other techniques such as regression or ratio analysis. As of January 2002, the DEA bibliography database consisted of 3,203 publications written by 2,152 authors (Tavares, 2002).

In the context of DEA, the firm or organization that is under study is known as decision making unit (DMU).The DEA is able to identify a subset of efficient ‘best practice’ DMUs and as for the remaining DMUs, it provides a magnitude

of the inefficiency in comparison to a frontier constructed from the 'best practice' DMU. Farrell (1957), a pioneer in measuring efficiency had first introduced measurement of efficiency by using a single input and a single output of a firm. Based on Farrell's work, the DEA was then extended by (Charnes et al., 1978).

The original DEA that was introduced by Charnes et al. (1978) assumed constant returns to scale (CRS) whereby it is only appropriate when all DMU operates at an optimal scale. However, based on Coelli et al. (1998), the imperfect competition or limited financial resources may cause the DMU not operating at an optimal scale.

Favero and Papi (1995) had indicated that DEA does not explicitly make any assumptions regarding the functional form of the frontier but empirically builds a 'best practice' functions from the inputs and outputs used.

The DEA methodology has also criticism whereby it assumes absence of measurement error and statistical noise. Besides that, Herrero & Pascoe (2002) based on their study have stated that the inefficiency scores may be biased if the production process is largely characterised by stochastic elements.

The DEA approach in analysing financial institution in terms of deriving the summary measure of the efficiency of each unit has been increasingly used by researchers (Bauer et al, 1998).

The concept of measurement in DEA is further elaborated in Chapter 3 - Selection of Measures which will also provide details implementing DEA in the context of this study.

2.5 Input and Output Variables

The input and output variable used for the estimation can be breakdown based on Production Approach, Intermediation Approach and Environmental Variables which are further categorised by regulatory specific variables and bank specific variables.

2.5.1 Production and Intermediation Approach

The production approach treats loans and deposits as outputs and measures in terms of the number of accounts and transaction standpoint. However, the intermediation approach views banks as financial intermediaries whereby the role of banks is to collect funds from units in surplus and then to loan the surplus funds out. The production approach limitation is that it considers only the operating cost however it excludes the interest expense that is also part of the bank's expense.

Berger and Humphrey (1997) indicated that neither of these two approaches is suitable to define the inputs and outputs as either of this is able to capture the dual role of banks as producers of services and being a financial intermediary. However, the study suggests that study on branch level efficiency would be more appropriate to adopt the production approach while

study on the bank itself would be more appropriate to adopt the intermediation approach.

2.5.2 Bank Specific Environmental Variable - Size

Various studies have been done whereby assessment against the size of the bank based on the size (Berger and Mester, 1997) however little studies that have been done in terms of cross country comparison for ASEAN countries.

Theoretically, the larger size based banks will be more efficient as they are able to take advantage of the economies of scale. Previous studies on the relationship between bank size and efficiency is ambiguous where some studies reports to have significantly positive relationship (Berger et al. 1993; Miller and Noulas, 1996) and some studies findings with significantly negative relationship (DeYoung and Noulle, 1996).

Fukuyama (1996) research on Japanese banks finds that the major factor contributing to the overall technical inefficiency is pure technical inefficiency and not scale inefficiency which suggest that size is not an important factor for Japanese banks to perform efficiently.

Drake and Hall (2003) however had later study on the Japanese banks and the result is that size efficiency relationships are established for both technical and scale efficiency which explains the logic of the merger of Japanese banks.

Based on the input minimization model, larger banks are on average more efficient than smaller banks which are consistent with Kwan (2006) and Drake et al. (2006).

Mehdian et al. (2007) had examine the effects of globalization and deregulation on the efficiency and the productivity growth of small and large banks in the U.S. between 1990 and 2003 where the evidence shows that large banks are generally more efficient than small banks for most efficiency indices in both years.

Ray S. (2007) had evaluated the size efficiency of Indian banks by using data from 1997 to 2003. Based on the study made, one point that was highlighted is that there is a need to consider size efficiency as opposed to scale and scope efficiencies.