AN INVESTIGATION INTO THE DETERMINANTS OF COST EFFICIENCY IN THE ZAMBIAN BANKING SECTOR

Anthony Musonda*
School of Economics
University of Cape Town

Paper Presented at the 13th Annual African Econometrics Society (AES) Conference
9 – 11 July 2008
University of Pretoria
South Africa

PRELIMINARY, PLEASE DO NOT CITE

Abstract

Research interest in banking cost efficiency in developing countries has increased over the past years as researchers seek to evaluate the efficacy of reforms implemented in these countries. However, for sub-Saharan Africa, such research remains thin and inconclusive. This study attempts to fill this void by investigating cost efficiency in the Zambian banking sector since the banking crisis of the mid 1990s. Using a single stage maximum likelihood estimation procedure applied to a stochastic frontier cost function, we observe that Zambian banks are on average inefficient in the order of 11.4 %. Furthermore, foreign banks are more efficient than domestic banks, especially the state bank. However, domestic private banks have been closing the efficiency gap over the past few years. Contrary to popular thinking in the Zambian banking circles, regulatory intensity does not exacerbate inefficiency. Instead, bank specific factors and macroeconomic uncertainty have contributed significantly to the relatively low efficiency bank performance. In setting the agenda for future improvement in cost efficiency, attention must be focussed at improving risk management techniques and reducing credit to government in order to unlock the potential in the domestic banking sector. In order to achieve this, the institutional framework must be strengthened to revitalise the sector.

Keywords: banking, regulation, cost efficiency, stochastic frontier analysis

JEL CLASSIFICATION: **D24, G21, G28, L25**

---

*Contact details: School of Economics, University of Cape Town, P/B Rondebosch, 7701, Cape Town.
E-mail: 3kids.wkm@gmail.com

AN INVESTIGATION INTO THE DETERMINANTS OF COST EFFICIENCY IN THE ZAMBIAN BANKING SECTOR

1.0 Introduction

Research on banking efficiency in developing countries has received little attention despite rapid growth in this literature over the years. This is rather unfortunate given the dominance of the banking sector in the financial system in these countries. Banks play an invaluable role in transforming risk, size and maturity of financial instruments. In doing so, they convert deposits into productive investment. Indeed, economies with a sophisticated financial system are more capable of evaluating assets whose real value is difficult to determine (Thankor, 1996). Accordingly an efficient banking sector is essential for vibrant economic performance.

Sweeping economic reforms that engulfed a number of transition and developing economies in the late 1980s to early 1990s injected fresh air in the functioning of markets, including the banking industry. For the Zambian banking sector, financial liberalisation was implemented following many years of government intervention.[1] The process of financial liberalisation was particularly unique because it was embraced wholesomely and became an important component of the overall economic reform agenda. Since the first round of structural reforms instituted in the early 1990s, significant progress has been made in the Zambian banking sector. However, the banking crisis of the mid 1990s also provided valuable lessons for regulatory policy and general bank management.

Thus motivated by concerns of solvency and stability of the banking industry, the central bank adopted tighter supervisory and regulatory policies and implemented new banking legislation to prevent future crises. In this streak, bank restructuring took centre stage and comprised of promotion of investment that aimed at strengthening capital power. The corollary of this in efficiency terms is however unclear. On one hand, expansion in bank capital may be efficiency enhancing since equity serves as an alternative source of funding. On the other hand strengthening bank capital may not always translate into higher efficiency performance and could end in wasteful investment in some cases. To this end, the effect of bank regulatory capital on efficiency may be ambiguous. Nonetheless, the general consensus is that tighter regulations could strangle banking efficiency performance.

---

[1] For a comprehensive review of banking sector reforms in Zambia, see Brownbridge (1998) and Brownbridge and Gayi (1999).

In the Zambian banking industry, changes to institutional and regulatory policy framework significantly shaped the financial landscape and impacted greatly on banks' profitability, cost structure and efficiency. Evidently, reform measures benefited some individual banks and the banking sector as a whole. However, the reforms also posed great challenges for general bank conduct.

Although there had been attempts to address the issue of regulatory intensity and associated costs in the estimation of banking efficiency, such studies have not incorporated financial capital cost and/or opportunity costs induced by required reserves in the efficiency model (Estrada & Osorio, 2004). Since banks operate in an imperfect market, authorities impose capital requirements, required reserves and other restrictions in order to influence banks' conduct. In turn banks incur additional costs in complying with these regulations. In the absence of regulations however, banks would optimally choose their exposures to banking activities so that they remain on the efficient frontier. This presupposes that banks posses full information about the nature and magnitude of risk and return characteristics. Thus, it would be postulated that due to the presence of market imperfections, regulatory requirements attenuate the information problem by limiting risk taking.

It turns out that these forms of regulatory requirements may affect the level of efficiency because they act as fixed inputs in the production process. Therefore in determining bank cost behaviour it is important to incorporate regulatory factors (Elliehausen, 1998).[2] Estrada and Osorio (2004) offer one option by explicitly allowing regulatory cost factors to determine the level of efficiency. This specification directly captures the regulatory burden borne by the banks and helps understand the dynamics of hidden costs which may be difficult to observe in practice. To illustrate, tighter capital regulatory requirements meant to safeguard the banking industry may produce some unintended adverse consequences on bank behaviour, including efficient allocation of resources. Research by Fare *et al.* (2004) found that risk-based capital standards had a significant impact on profit efficiency. Hughes and Mester

[2] Some studies have estimated direct and compliance costs of regulation see for instance, Franks, *et al.* (1998) and Elliehausen (1998) for a survey. These estimates show that costs of regulation range between 9 % – 13 % of non-interest expenses. Studies also show that there are economies of scale associated with regulatory costs; large banks incur lower per unit regulatory costs than smaller banks. Our analysis reveals that the income statement from Zambian banks does not report costs associated with regulatory requirements. Therefore, it is very difficult to estimate the extent of regulatory costs and apply them to banking efficiency. To circumvent this handicap in our empirical framework below, we use reasonable proxies to capture regulatory intensity on and associated costs as measures of regulatory distortions.

(1993) also argue that an increase in reserve requirements raises the opportunity cost of funds and acts as a tax on the price of deposits, thereby impairing banks' decision making.

Another important consideration relates to the macroeconomic environment under which banks operate. Significant changes in the business cycle and general economic performance and the policy environment induced by economic reforms also have implications for bank conduct. For instance, in Zambia, persistent government fiscal deficits, high inflation and rapid currency depreciation created uncertainty in the economy. Collectively, these factors may have affected resource allocation in the banking sector.

In view of the foregoing, this study seeks to address itself to the following research questions. What has been the level of efficiency of Zambian banks in the post-reform period? Are cost efficiency and bank ownership structure related, and if so, which bank categories are more efficient than others? Do loan quality and other factors intensify banking inefficiency? Furthermore, is regulatory burden a major source of cost inefficiency in the Zambian banking industry? These are important policy questions which would shed light on the nature and dynamics of efficiency performance of Zambian banks since the banking crisis of the 1990s. In the remainder of the paper, we outline the approach used to answer these questions. The next section discusses the theoretical basis for economic efficiency.

2.0 Theoretical foundation of economic efficiency

The concept of economic (productive) efficiency is rooted in neoclassical microeconomic theory, which focuses on resource allocation and utilisation. It advocates for non-wastage of resources by emphasising cost reduction while producing the maximum possible level of output for a given technology and available inputs. Thus, a firm that is economically efficient may possess competitive advantage over rival firms producing less efficiently in the same industry. The main driving force behind economic efficiency is value creation. Accordingly, in the process of transforming inputs into some output value, a change that increases value is an efficient change and one that decreases value is an inefficient change. For purposes of policy intervention efficiency has often been used to evaluate the effectiveness of policy alternatives.

A related concept of economic efficiency is Pareto optimality, which has foundations in welfare economics. Pareto efficiency occurs when an allocation of resources from one economic agent to the other makes one agent better off without compromising the welfare of another economic agent (that is without making the other individual worse off). Therefore

Pareto efficiency has important implications for public policy, especially in the redistribution of income. Although theoretically plausible, Pareto efficiency is difficult to measure is practice.

Economic efficiency is better explained by profit maximisation (or analogously, cost minimisation) but is often associated with perfectly competitive markets than with monopoly because of deadweight loss associated with monopoly pricing and output restrictions. For firms operating in a competitive industry, efficiency gains accrue when firms earn only normal profits in the long-run and respond to changes in consumer preferences by increasing output. Whether this output in sold at the same, higher or lower price depends in large measure on the position of the cost curves in the long-run (Griffiths & Wall, 2000). In general however, efficiency is associated with welfare improvements.

Economic efficiency also encompasses allocative efficiency, which occurs when a firm's inputs are allocated in such a way as to maximise its benefits (profits, revenue and output) depending on the firm's objective function. Allocative efficiency is thus concerned with informing resource allocation decisions by taking into account both productive efficiency as well as Pareto efficiency. However, it is still possible to achieve Pareto efficiency without allocative efficiency. At firm level, allocative efficient outcomes occur when price is equal to marginal costs in a perfectly competitive market. Allocative efficiency also addresses the issue of the right mix of inputs and quality of output produced.

Finally, X-efficiency, introduced by Leibenstein (1966) refers to efficiency in production by linking inputs to outputs. It is an economic expression for the effectiveness with which an organisation uses its given set of inputs to produce outputs. Specifically, it refers to the internal organisation of firms and its response to external factors. Under such circumstances, both motivational factors (e.g., moral and bureaucratic inertia and human errors) and competitive pressures may affect X-efficiency. In many of his writings, Leibenstein repeatedly argued that X-efficiency was superior to allocative efficiency, implying that the latter effect was trivial.

Evidently, the concept of efficiency is broad and has been assigned different interpretations. The concept of X-efficiency is especially controversial. Since Leibenstein first proposed the term, analysts have sought to evaluate its meaning and in the process, different interpretations have merged. It is therefore important to be clear when the term X-efficiency is used in analysing an economic problem. In contrast to Leibenstein's construct of the concept, Peel (1974) contends that X-efficiency can actually be attained even with a lazy or hardworking work force only if effort and efficiency dimensions of inputs that translate into maximum output are taken into consideration. Furthermore, Peel (1974) posits that cost reduction can be observed in a movement from monopoly to perfect competition, not due to X-efficiency but because managers are forced to give up goals inconsistent with cost minimisation as the environment becomes more competitive.

Like many other critics of the theory of X-efficiency, Stigler (1976) cautions that failure to recognise that higher than minimum cost levels are mainly due to rationally calculated utility maximising strategies (including leisure) by workers rather than by X-efficiency. This criticism may imply that firms are always producing at the production frontier in which only allocative efficiency would then be relevant.

The above efficiency definitions refer to static efficiency. One form of efficiency with time dimension is dynamic efficiency (Jameson, 1972). Dynamic efficiency takes into account changes in technology as a main source of productivity, so that over time, the production frontier shifts due to technological advancements. One would associate dynamic efficiency with long-run economic growth arising from productivity change through capital accumulation (Abel, Mankiw, Summers, & Zeckhauser, 1989).

Scale economies and economies of scope provide another perspective for analysing firm efficiency. Economies of scale occur when more units of a good or service can be produced on a larger scale, yet with less input costs. Therefore, economies of scale are associated with size of the firm, implying that larger firms enjoy economies of scale due to larger production technology. For a long time, economic growth has been explained by the theory of economies of scale. At firm level, Adam Smith showed that specialisation and division of labour are the two most important pillars behind productivity growth and increase in efficiency. Marshall (1961) distinguished between internal and external economies of scale, attributing the former to lower costs which lead to higher production and the latter as emanating from external forces such as improvements in transportation network. External economies of scale may thus benefit all firms in the industry by lowering costs and stimulating firm growth.

In contrast, economies of scope relate to a reduction in costs resulting from joint production. Thus, a firm which decreases its average costs because of changes in production of different products (related or unrelated) is said to be enjoying economies of scope. Economies of scope also provide firms with means to generate operational efficiencies, especially when these are driven by diversification. In the case of banking, for instance, it may be economical for a bank to diversify into different areas of financial services such as investment banking, commercial banking, leasing and life insurance rather than pursue traditional commercial banking alone. Such a strategy may be driven by synergies of knowledge of commercial banking and investment banking of particular corporate customers. For larger banks, diversification may be an important business strategy aimed at reducing portfolio risk and hence gain efficiency. However, empirical evidence for the United States does not support this hypothesis (Demsetz and Stratan, 1997).

Although scale economies may be important, bigger may not necessarily be better in the sense that when firms expand in size, the chain of command also becomes cumbersome. Therefore decisions taken at the top may take longer and information distorted by the time it reaches the bottom ladder of the firm. When this happens, the firm may not be realising cost reductions as predicted by the theory. Accordingly, diseconomies of scale may set in. Thus, smaller firms may be more efficient than larger firms in this regard. A similar interpretation may be offered for economies of scope.

While economies of scope often provide an incentive to expand product lines, the creation of new products is often less efficient than expected, resulting in diseconomies of scope. This is because the introduction of new product brands may entail additional managerial expertise or personnel, higher raw material costs, a reduction in competitive focus, and the need for additional facilities, which collectively could result in an increase of a firm's per-unit costs. Therefore, economies of scope may not be obvious. For the banking example above, managers may find it challenging to manage insurance business while also trying to provide commercial banking services to a wide range of customers. For this reason, the envisaged synergies may actually lead to huge cost increases.

Different factors may explain efficiency levels in a firm. Some of the factors may be inherent in the internal organisational structure of the firm, including managerial expertise, experience of workers and skills levels. Internal sources of (in) efficiency typically include laxity, human mistakes, disruption of production technology or insufficient capacity to respond to changing incentives. Other factors may be external to the firm. These factors, also called environment factors, may include luck, regulatory constraints, macroeconomic shocks, real business cycles, strikes and labour disputes and structure of the market in which the firm is operating. Taken together, these factors may account for a substantial amount of variability and differences across firms' performance levels. Internal factors are firm specific and therefore within the control of the firm, environmental factors are outside the control of the firm. Therefore, in assessing the efficiency of firms, care must be exercised in differentiating between internal and external factors.

3.0 Measurement and application of economic efficiency in banking

The concept of productive efficiency was first mooted in Farrell's (1957) seminal work in which he showed that overall efficiency can be decomposed into "price efficiency" and "technical efficiency". Farrell's (1957) definition of technical efficiency led to the evolution of different methods for estimating relative efficiencies of firms. The common feature of these approaches is that information on relative efficiency is obtained from extreme observations of available data to determine the best practice production frontier (Lewin & Lovell, 1990). Using this information, relative efficiency of each individual decision making unit can be estimated and comparisons made among them. Although frontier approaches posses some similarities, there are significant methodological differences among them. Therefore, the approaches used for estimating economic efficiency can be categorised broadly as being either parametric (stochastic) or non-parametric (linear programming) depending on the specification and estimation of the efficiency frontier and assumption made about the distribution of the error component.

The stochastic frontier approach (SFA) was first developed by Aigner *et al.* (1977) and Meesen and van den Broeck (1977) who estimated efficiencies using cross-sectional data. Subsequently, Ferrier and Lovell (1990) applied the methodology to banks. The SFA specifies a particular form for the production/cost function allowing for a composite error term. Thus, the methodology involves parameterising the relationship between the level of inputs and the technically efficient level of output. Stochastic frontier models use econometric modelling. However, an often cited criticism of the stochastic frontier approaches is that when the specification of the efficiency function and stochastic term are assumed a priori, it may not be clear whether or not the efficiency measure is contaminated by the misspecification of the estimated econometric model.

Another variant of frontier estimation techniques in founded in the so-called non-parametric approaches based on data envelopment analysis (DEA) or linear programming techniques following the seminal work of Charnes *et al.* (1978). The approach by Charnes *et al.* (1978) uses Farrell's (1957) concept of efficiency under constant returns to scale (CRS). Later reformulation of the DEA model by Banker *et al.* (1984) showed that overall efficiency can be divided into 'pure technical' and 'scale' efficiency and suggested that firms may in fact be characterised by variable returns to scale (VRS). The DEA also decomposes overall efficiency into technical and allocative efficiency. Data envelopment analysis does not explicitly make any assumptions regarding the functional form of the frontier but empirically builds a best-practice function from observed (actual) inputs and outputs (Favero & Papi, 1995). However, a major criticism levelled against the DEA methodology is that it assumes absence of measurement error and statistical noise. Accordingly, errors are taken as measures of inefficiency. However, as Herrero & Pascoe (2002) have observed these inefficiency scores may be biased if the production process is largely characterised by stochastic elements.

3.1 Empirical Literature

This section reviews and discusses some of the related empirical literature on banking efficiency. Over the years, models of bank efficiency have evolved along the two distinct frontier estimation techniques discussed above. Both the stochastic frontier and DEA models have received substantial application in the banking efficiency literature. Traditionally, technical efficiency in banking was measured using the production function capturing scale and scope efficiencies. However, technical efficiency is only a component of overall economic efficiency. Although extensive in use, scale and scope efficiency measures are of little economic significance for financial institutions (Kwan & Eisenbeis, 1996). This view is supported by Berger *et al.* (1993) and Berger and Humphrey (1991) who argue that scale and scope inefficiencies are less important than X-inefficiency in the banking industry. Similarly, Bauer *et al.* (1998) argue that for policy purposes, economic efficiency is a much broader concept than technical efficiency in the sense that the former encompasses the latter and involves an optimal choice of inputs and/or outputs based on the reactions to market prices.

Early models of the banking efficiency literature were mainly for industrial countries, particularly the United States and Europe (Berger & Humphrey, 1991; Aly, Grabowski, Pasurka, & Rangan, 1990; Ferrier & Lovell, 1990; Rangan, Grabowski, Aly, & Pasurka, 1988; Sherman & Gold, 1985). In recent years, there have been a sizable number of studies undertaken in developing countries and countries in transition. Berger and Humphrey (1997) and Bikker & Haaf (2003) provide a survey of some of the early works in developed countries for banking efficiency. For an international comparison of efficiency studies, see Berger (2007). Especially influential in this genre of literature is the Berger and Humphrey's (1991) study. Berger and Humphrey (1991) argued that X-inefficiencies appeared superior to scale and product mix in banking and were in the order of 25 % or more of overall cost inefficiencies. Scale inefficiencies accounted for only 5 % or less. Earlier, Rangan, *et al.* (1988) had found that US banks were about 30 % inefficient largely due to pure technical inefficiency. Estimates for scale inefficiencies were found to be very small with almost all banks operating with constant returns to scale. However, recent evidence by Valverde, *et al.* (2007) shows that correcting for banks' external business environment and productivity indicators at industry level significantly reduces inefficiencies fivefold. Thus, based on this new research, inefficiencies from scale operations and X-inefficiencies are ranked equally. Evidently, one would think of industry productivity growth as emanating from firm level scale expansion and therefore, this would translate into lower than expected inefficiencies for individual banks.

For studies conducted in industrial countries there is unequivocal evidence that deregulation fosters banking efficiency (Sturma & Williams, 2004; Kaparakis, Miller, & Noulas, 1994; Mester, 1997; Mester, 1993; Berger & Mester, 1997; Berger, Hunter, & Timme, 1993). However, other scholars have also found that globalisation of financial markets manifested in increased foreign bank penetration has tended to dwarf banking efficiency although this result must be interpreted in conjunction with the quality of institutions in host countries (Lensink, Meesters, & Naaborg, 2008).

Given the accelerated pace of economic reforms and globalisation of financial markets since the late 1980s, there has been a growing body of research assessing efficiency in banking and productivity performance in developing countries and emerging markets. This interest has been necessitated by availability of data in these countries. Generally, results of banking efficiency during liberalisation period for less developed countries (LDCs) have produced mixed results for a number of countries (Chen, Skully, & Kym, 2005). This thinking tallies with Berger and Humphrey (1997) argument that deregulation might not always improve efficiency and productivity due to other intervening incentives especially in the early years of reform.

To illustrate, for some studies in India banks' overall efficiency was found to be higher for public sector banks than for foreign and domestic private banks (Ataullah & Le, 2006; Sensarma, 2006; Ataullah, Cockerel, & Le, 2004; Shanmugan & Das, 2004). However, the results of di Patti and Hardy (2005) for Pakistan show that efficiency gains were high for all banks during the early years of privatisation but subsequently, the level of efficiency for privatised banks decreased. New private domestic banks were the shining example of better performance while foreign banks lagged behind.

Emerging markets and former communist countries have provided a new 'laboratory' for testing banking efficiency. For the majority of studies undertaken in these countries, evidence is consistent with prior expectations that liberalisation stimulates efficiency of private banks giving impetus for continued reforms and restructuring of public banks (Christopoulos & Tsionas, 2001; Zajc, 2006; Fries & Taci, 2005; Rao, 2005; Staikouras, Mamatzakis, & Koutsomanoli-Filippaki, 2007; Weill, 2003; Yildirim & Philippatos, 2007; Denizer, Dinc, & Tarimcilar, 2007). Similar findings have been reported for the Turkish banking sector, where Zaim (1995) and Isşık & Akçaoğlu (2006) showed that liberalization fostered efficiency with a large number of Turkish banks operating at an optimal scale. This set of studies shows that through financial liberalisation, banks unlock their potential and undertake measures aimed at cutting costs in order to increase profitability. However, Denizer *et al.* (2007), Yildirim (2002) and Işık and Hassan (2002) observed that contrary to the findings of Zaim (1995), there was a slow recovery of efficiency by all Turkish banks mainly due to the unstable macroeconomic and financial environment under which these institutions operated.

Another notable focus of research on banking efficiency has been that received from the Asia-Pacific region. In analysing banking efficiency of Korean banks, Hao *et al.* (2001) concluded that the financial liberalisation measures of 1991 did not spur improved cost efficiency in the banking sector. Instead, better performance was driven by asset growth and low expense ratios. A possible explanation for this may be that due to relaxation of entry barriers and other banking activity restrictions, competition intensified, causing many banks to adopt measures to improve productivity and efficiency. In China, Chen *et al.* (2005) study the efficiency of Chinese banks for the pre- and post-liberalisation period using the DEA methodology. Their conclusion is that there was marked improvement is the level of efficiency after the financial deregulation of 1995. They also found that large state-owned banks and smaller banks were more efficient than medium sized Chinese banks.

Similar results are noted in a United Nations (2005) study which assessed the impact of financial liberalisation on the efficiency performance of banks in selected Western Asian countries. The conclusion from the study was that better performing banks were more cost efficient than banks that did not engage in profitable investment activities. The analysis also showed that more efficiency gains could accrue from further development in the financial sector. Other studies in the Asian-Pacific countries generally conclude that productivity and efficiency improved after the reforms with the private banks the major beneficiaries from reform. However, short-term effects arising from the instability in the macroeconomic environment tended to hamper speedy improvement in efficiency of most commercial banks (Lin, 2005; Lim & Randhawa, 2005; Leightner & Lovell, 1998).

For sub-Saharan African (SSA) countries studies of banking efficiency are limited and include Ikhide (2000) and Adongo *et al.* (2005a; 2005b) for Namibia; Hauner and Peiris (2005) and Beck and Hesse (2006) for Uganda and Čihák and Podpiera (2005) for Kenya, Tanzania and Uganda. Ikhide (2000) and Adongo *et al.* (2005a; 2005b) reached contrasting conclusions on the efficiency of Namibian banks with the former positing that banks in Namibia were characterised by inefficiency while the latter studies indicated that Namibian banks compared relatively well with international evidence. These contrasting findings may be due to differences in the approaches used to measure banking efficiency, and the variables included in the models specified. On the other hand, Čihák and Podpiera (2005) and Hauner and Peiris (2005) reported similar results for East African countries, noting that an increase in bank competition was associated with a rise in efficiency. However, this evidence is not shared by Beck & Hesse (2006) who argue that banking spreads have been significantly high in Uganda, indicating inadequate efficiency in the banking industry. It is important to point out that this study inferred efficiency from high spreads rather than rely on more robust techniques to estimate efficiency. It is well acknowledged that efficiency analysis based on ratios and spreads suffers from a number of shortcomings and may not provide reliable estimates of banking efficiency (World Bank, 2006).

Recently, Kablan (2007) and Kirkipatrick *et al.* (2007) have estimated banking efficiency for countries in the West African Monetary Union (WAEMEU) and SSA, respectively. Kablan (2007) observed that banking efficiency was generally higher in the majority of WAEMU states except for Burkina Faso and Cote D'Ivoire while Kirkipatrick *et al.* (2007) found lower profit X-inefficiency than cost X-inefficiency for SSA countries. Both studies observed that inefficiency was sensitive to the quality of loans and bank soundness. Interestingly, Kirkipatrick *et al.* (2007) also found a negative effect of financial liberalisation but submit that foreign bank penetration helped dampen cost X-inefficiency. Although many countries were already implementing financial reforms during this period, it is important to note that the banking industry in a majority of these countries also experienced significant instability. It is not surprising therefore that cost-efficiency was susceptive to risk and solvency factors and the turbulent economic environment.

As noted earlier, government borrowing from the domestic banking sector to finance the fiscal deficits especially for LDCs has had implications on banking performance and efficiency. In a recent paper, Hauner (2008) addresses this issue for a number of developing and developed countries. As would be expected, credit to government is found to limit banking efficiency in developing countries while in advanced countries, it had the reverse effect. This result could mean that growth in credit to government by the domestic banking sector signals macroeconomic instability with potential devastating impact on efficiency performance.

The review of banking efficiency literature above shows that broadly, there is sufficient piece of evidence to suggest that banking efficiency improves with the breadth and depth of liberalisation and deregulation policies. The reported reverse effect in a few studies may be due to the short-term effects of liberalisation such as credit rationing, high spreads and weakening loan quality (Gruben & McComb, 2003; Leightner & Lovell, 1998). These problems tend to be exacerbated under an unstable macroeconomic environment which is often associated with early years of reforms.

Although research for industrial countries and some advanced emerging markets may be illuminating, it provides little insight into efficiency trends in developing countries, especially in SSA where local economic conditions and specific factors may be structurally different. As Denizer *et al*. (2007) have recently observed, a number of developing countries experienced instability in the macroeconomic environment characterised by high inflation, slower economic growth and other economic ills, including banking crises. These factors tend to produce a distorted incentive structure for banks, making resource allocation and utilisation to achieve greater efficiency a much harder task. Hence, of relevance to this study is how the efficiency of banks in a low-income developing economy has been influenced by changes in the financial landscape induced by financial liberalisation and regulatory policy changes. These issues are taken up in subsequent sections.

3.2 Methodological Framework

Estimating cost efficiency assumes a given technology relating observed inputs and prices to a predetermined level of output. Thus, when the actual level of costs is above the minimum possible, we get cost inefficiency. Since there is really no broad consensus on the preferred method to use in benchmarking bank performance, estimation of cost efficiency in banking has employed either of the two frontier techniques discussed above. Some researchers advocate use of both techniques. Taking a cue from this argument, a number of studies have provided evidence of efficiency scores derived from both frontier approaches

(Fiorentino, Karmann, & Koetter, 2006; Beccalli, Casu, & Girardone, 2006; Weill, 2004; Eisenbeis, Gary, & Kwan, 1999; Resti, 1997). This approach is known as methodological 'cross-checking' and allows for robust comparison of results across different methods. However, Resti (1997) observes that there is no significant difference between the two approaches. When differences occur, they can be explained by revisiting the intrinsic features of the models. On the other hand, Eisenbeis *et al.* (1999) argue that efficiency scores derived from linear programming DEA are two to three times larger than those estimated by the SFA. However, the patterns of scores across banks are similar and there is a relatively high correlation between the efficiency scores derived from the two methods.

Against the above background, there is really no loss of generality in using either methodology to analyse efficiency in banking. The choice of the approach adopted is a matter of convenience and is largely dictated by the data used in the analysis and ease of application. Consequently, in this study we prefer the stochastic frontier approach to the DEA. In choosing the SFA, we argue that the model allows for simultaneous estimation of the cost function and the inefficiency model, a feature which the DEA does not support. A further advantage of the SFA over the DEA is that with panel data, statistical noise is better handled in the former in the sense that additional information from multiple time periods is incorporated into the estimation.

3.2.1 Formulating the Stochastic Frontier Cost Function

Estimation of a stochastic cost frontier requires specifying a particular functional form for the production function. However, since banking is a multi-output industry, specification of a production function is not feasible. For this reason, we appeal to Bikker and Bos (2005) and develop the model from first microeconomic principles by applying duality theory based on Jehle and Reny (2001). Duality theory posits that if the objective of a firm is profit maximisation or, equivalently cost minimisation, a given production function can be recovered from a corresponding cost function.

In order to remain relevant with economic theory, we postulate that overall cost efficiency is a precondition for bank performance and for regulatory policy analysis. This is justified by asserting that cost minimisation is a reasonable objective pursued by Zambian banks. Thus, following Bauer *et al.* (1998), we conjecture that Zambian banks minimise costs subject to a specific technology. Due to institutional and regulatory factors, commercial banks also face policy and institutional constraints which impose additional costs on production optimisation decisions, thus affecting efficiency performance. Therefore, these constraints form part of the decision variables for individual banks.

Although departures from cost minimisation may be justified on theoretical grounds, empirically the bank's objective function still remains that of cost minimization, even in the presence of regulatory constraints. Thus, by explicitly incorporating measures of regulatory pressure, empirical estimates of regulatory intensity on cost inefficiency could be captured in a simplified framework. For the purpose of this study the cost function is specified as:

$$TC(\boldsymbol{w}, \boldsymbol{y}) = \min \boldsymbol{w.x} \; s.t. \; T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\psi}) = 0 \tag{5-1}$$

where $TC(\boldsymbol{w}, \boldsymbol{y})$ is the cost minimisation function, TC is total measured economic costs incurred by the bank, $\boldsymbol{y}$ is a vector of bank output/services, $\boldsymbol{w}$ is a vector of input prices, $\boldsymbol{x}$ is a vector of factors of production, $T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\psi})$ is the transformation function and $\boldsymbol{\psi}$ denotes a vector of bank specific factors shifting the bank's cost function.

Equation (5-1) can be used to solve for cost minimisation input demand functions by invoking the Envelop theorem and applying the Shephard's Lemma (Jehle and Reny, 2001). Forming the Lagrangian function of (5-1) and differentiating it with respect to input prices, we can solve for input demand functions. Specifically, Shephard's Lemma states that conditional input demands from the cost function may be derived as follows:

$$\frac{\partial TC(\boldsymbol{w}, \boldsymbol{y})}{\partial w_j} = x_j^*\left(w_j, y_q, \psi_s\right), \; j = 1,2,...,J; \; q = 1,2,...,M; \; s = 1,2,...,S \tag{5-2}$$

where y_q is qth measure of output for bank i, w_j is the jth input price for bank i and ψ_s denotes sth exogenous cost shifters. An asterisk denotes optimal input demands. Note that Equation (5-2) states that conditional input demand is a function of the level of input prices, output and exogenous factors. Substituting (5-2) into (5-1) yields a firm's minimum cost function expressed as

$$TC_i^* = w_j x_j^*\left(w_j, y_q, \psi_s\right) \tag{5-3}$$

The above cost function embodies a long-run equilibrium relationship between costs on one hand and inputs, outputs and bank specific factors on the other, and it possesses all the properties of a cost function. Derivation of the minimum cost function above is useful in formulating the cost efficient frontier along the lines of the methodology proposed by Aigner *et al.* (1977) and Meesen and van den Broeck (1977). Battese and Coelli (1995) posit that a similar approach can be extended to analysis of efficiency using panel data. Hence, an empirical reformulation of (5-3) yields the stochastic frontier cost function expressed in logarithmic general form:

$$\ln\left(TC_{it}\right) = f\left(\ln\left(y_{qit}, w_{jit}, \psi_{sit}\right), \boldsymbol{\beta}\right) + \boldsymbol{\omega_{it}} \qquad (5\text{-}4)$$

where TC, y_q, w_j, ψ_s are as defined before but extended to include time dimension, $\boldsymbol{\beta}$ is a vector of unknown parameters to be estimated, and finally $\boldsymbol{\omega_{it}}$ is a two-component error term for observation i at time t and takes the following form:

$$\omega_{it} = v_{it} + u_{it} \qquad (5\text{-}5)$$

where v_{it} denotes a random uncontrollable factor which depicts measurement error and $u_{it} > 0$ is a one sided controllable component of the disturbance term, capturing cost inefficiency, that is, deviation of actual cost from its minimum possible level given the hypothesised frontier. Put differently, it reflects the inability of bank i at observation t to attain the potential minimum cost defined by the stochastic frontier.

According to Aigner *et al.* (1977) the random uncontrollable error term is assumed to be standard normal, independently and identically distributed with zero mean and constant variance, i.e., i.i.d $N\left(0, \sigma_v^2\right)$. There are no apriori reasons to prefer a specific distribution on the inefficiency term. However, in the literature, the inefficiency term is assumed to be derived from a truncated normal distribution or half normal distribution (Jondrow, Lovell, Materov, & Schmidt, 1982) with zero mean and variance, σ_u^2 i.e., i.i.d $N\left(0, \sigma_u^2\right)$.[3] Further, it

[3] Other inefficiency error distributions identified in the literature are the exponential and gamma distributions. See Greene (1990) for explanations on these forms of inefficiency error distributions.

is assumed that both v_{it} and u_{it} are independently distributed for all periods and across observations and orthogonal to all explanatory variables.

The assumptions made on the one-sided error allows cost inefficiency effects (u_{it}) obtained from estimating (5-4) to be modelled as a function of a set of explanatory variables (bank specific as well as time specific exogenous factors) which determine the level of inefficiency in the banking industry. Battese and Coelli (1995) specify a function for the inefficiency effects for panel data as follows:

$$\boldsymbol{u} = \boldsymbol{z\theta} + \boldsymbol{\varepsilon} \qquad (5\text{-}6)$$

where $\boldsymbol{u}$ is a vector of stochastic inefficiency scores estimated from the frontier function, $\boldsymbol{z}$ is the vector of firm-specific and environmental (including regulatory) factors which may influence the banks' cost inefficiency and $\boldsymbol{\theta}$ is the associated matrix of coefficients depicting the statistical relationship between cost inefficiency and its correlates. Finally, $\boldsymbol{\varepsilon}$ is a matrix of random error terms defined by the truncation of the normal distribution with zero mean and constant variance, that is, $N(0,\sigma_\varepsilon)$.[4] Consistent with Battese and Coelli (1995), the parameters of the cost frontier and the inefficiency model can be estimated in a one-step procedure. The alternative is a two-step procedure as espoused in other studies.[5] Following Jondrow *et al.* (1982) the specific cost inefficiency estimation for bank i at time t is given by the conditional mean distribution of u_{it} denoted as:

$$E(u_{it} \mid \omega_{it}) = \frac{\sigma}{1+\gamma^2}\left[\frac{f(\omega_{it}\gamma/\sigma)}{1-F(\omega_{it}\gamma/\sigma)} + \frac{\omega_{it}\gamma}{\sigma}\right] \qquad (5\text{-}7)$$

where $E(u_{it} \mid \omega_{it})$ is an unbiased estimator of inefficiency effects u_{it}, $\gamma = \sigma_u/\sigma_v$ measures the relative importance of u_{it} and v_{it} to the composite error term, ω_{it} and must lie between 0 and 1. F and f, are respectively the standard normal distribution function and standard

[4] The cost inefficiency model may also include other unobserved systematic differences across banks that have not been fully incorporated into the model.

[5] We argue that it is inappropriate to use the two-step regression procedure because it contradicts the assumption of identically distributed inefficiency effects in the stochastic frontier model and because of possible misspecification of the cost frontier; this could introduce some bias, see Battese and Coelli (1995).

normal density function evaluated at $\omega_{it}\gamma/\sigma$. As $\gamma \to 0$ the effect of the noise error term, v_{it} dominates the stochastic cost inefficiency error term u_{it} in determining the variation of the global residual ω_{it}. When this is the case, a traditional average cost function is the appropriate specification without any stochastic inefficiency effects. Conversely, when $\gamma \to 1$ the cost inefficiency effects are stronger in driving the composite error term, in which case a deterministic cost function would be the correct model. The bank specific cost inefficiency index $\left[exp\left(u_{it}^*\right)\right]$ is calculated by employing results from (5-4) above. Since a cost function is estimated in logarithmic terms, cost inefficiency $\left(CI\right)$ can be shown to be[6]

$$CI_{it} = E\left[exp\left(u_{it}^*\right) | \omega_{\iota\tau}\right] \quad (5\text{-}8)$$

In order to obtain the coefficients of the cost stochastic frontier function and the inefficiency model, the two equations are estimated by the method of maximum likelihood (ML). The econometric counterpart of (5-4) above is given by the transcendental logarithmic (translog) cost function (5-9). The translog function is more general and suitable for frontier estimation. To capture dynamic changes in cost performance over time, a time trend and its quadratic term are included. Accordingly, the estimated translog econometric cost function is given as:

[6] The converse is cost efficiency, which is then given by the inverse of CI, denoted as $CE = E[\exp(-u_{it}) | \omega_{it}]$. In the empirical section below, this is the adopted cost efficiency measure.

$$
\begin{aligned}
\ln\left(\frac{TC}{w_2}\right) &= \beta_0 + \beta_1 \ln\left(\frac{w_1}{w_2}\right) + \beta_3 \ln\left(\frac{w_3}{w_2}\right) + \frac{1}{2}\beta_{11}\left(\ln\left(\frac{w_1}{w_2}\right)\right)^2 + \frac{1}{2}\beta_{33}\left(\ln\left(\frac{w_3}{w_2}\right)\right)^2 \\
&+\beta_{13}\ln\left(\frac{w_1}{w_2}\right)\ln\left(\frac{w_3}{w_2}\right) + \delta_1 \ln(sec) + \delta_2 \ln(loans) + \frac{1}{2}\delta_{11}\left(\ln(sec)\right)^2 \\
&+\frac{1}{2}\delta_{22}\left(\ln(loans)\right)^2 + \gamma_{11}\ln\left(\frac{w_1}{w_2}\right)\ln(sec) + \gamma_{12}\ln\left(\frac{w_1}{w_2}\right)\ln(loans) \\
&+\gamma_{31}\ln\left(\frac{w_3}{w_2}\right)\ln(sec) + \gamma_{32}\ln\left(\frac{w_3}{w_2}\right)\ln(loans) + \delta_{12}\ln(sec)\ln(loans) \\
&+\theta_1 t + \frac{1}{2}\theta_2 t^2 + \theta_3 t\ln(sec) + \theta_4 t\ln(loans) + \theta_5 t\ln\left(\frac{w_1}{w_2}\right) + \theta_6 t\ln\left(\frac{w_3}{w_2}\right) \\
&+\varphi_1 \ln(fundrisk) + \varphi_2 \ln(branch) + \varphi_3 \ln(lossgls) + \omega
\end{aligned}
\tag{5-9}
$$

where the variables are defined below. Note that we have suppressed the time and firm specific subscripts to avoid clutter. Nonetheless, Equation (5-9) is estimated by pooling observations across time. It is common practice in this genre of literature that duality theorem conditions are invoked. Accordingly, the cost function must be linearly homogenous in input prices. For this purpose, homogeneity of degree one in input prices is imposed by normalising total costs and input prices by the input price of funds, w_2 before log transformations are undertaken.[7] According to Berger & Mester (1997) such normalisation also resolves problems of potential heteroscedasticity. Furthermore, it is a normal to impose symmetry restrictions in parameters, i.e., $\delta_{12} = \delta_{21}; \beta_{13} = \beta_{31}$. Although some studies also estimate share equations, these are excluded from the analysis herein because they assume away possible allocative inefficiencies in the sample.

3.2.2 The inefficiency correlates

The inefficiency model is given by (5-10):

$$
\begin{aligned}
u_{it} &= \phi_0 + \phi_1 \ln(lossgls_{it}) + \phi_2 \ln(contass_{it}) + \phi_3 \ln(oppcost_{it}) \\
&+ \phi_4 \ln(inflation_t) + \phi_5 ownership + \phi_6 capdum + \varepsilon_{it}
\end{aligned}
\tag{5-10}
$$

Although Eqs. (5-9) and (5-10) are estimated in one-step by ML, it is important to highlight

[7] See Cebenoyan (1988) and Zardkooh *et al* (1986) and references cited therein for details on appropriate forms of normalisations.

the significance of the inefficiency correlates- relating the inefficiency estimates to possible independent determinants. Battese and Coelli (1995) have argued that both the frontier and inefficiency model could have all or some variables in common as drivers of costs may equally influence banks' inefficiency performance.

3.2.3 Economies of scale and technological change in Zambian banking

In the preceding sections, focus has been devoted to analysing cost efficiency and its determinants in the post reform era. This section provides a methodology for analysing the existence of economies of scale and technological change in Zambian banking over the study period. The importance of economies of scale for regulatory policy is well established. Regulatory authorities use estimates of economies of scale to determine the appropriate regulatory policy in respect of mergers and acquisitions and in predicting future industry structure. It will be recalled that cost inefficiency in Zambian commercial banks depicts overutilization (mismanagement) of input resources in the production process. However, it is not clear whether there is some level of output at which banks are said to be enjoying global economies of scale or indeed if the cost inefficiency could be ascribed to existence of diseconomies of scale and absence of technological advancement. Estimates of economies of scale are obtained from the cost frontier using (5-11) below:

$$ES = \left[\sum_{q=1}^{2} \frac{\partial \ln(TC)}{\partial \ln(y_q)} \right]^{-1} = [\delta_1 + \delta_2 + (\delta_{11} + \delta_{12}) \ln(sec) + (\delta_{22} + \delta_{12}) \ln(loans) + (\gamma_{11} + \gamma_{12}) \ln\left(\frac{w_1}{w_2}\right) + (\gamma_{31} + \gamma_{32}) \ln\left(\frac{w_3}{w_2}\right) + (\theta_3 + \theta_4) t]^{-1} \quad (5\text{-}11)$$

Economies of scale, as denoted by (5-11) measure the relative change in the bank's total operating cost for a given proportional change in the measured outputs. If ES is greater than, less than or equal to one banks are said to be experiencing economies of scale, diseconomies of scale or constant returns to scale, respectively.

Another important concept in banking production is technological progress (TP) which estimates a relative change in operating costs due to passage of time. This might occur when banks' investment in some technologically improving production techniques in turn allows them to capture efficiency gains from reduced production costs. Technological change is therefore an encompassing concept that includes financial innovations, changes in competition intensity among commercial banks and dynamics in loan allocations, vis-à-vis households and firms. Accordingly, technological change may affect how banks provide financial services and the instruments used to provide these services. For example, technological change could be underpinned by the spread of automated teller machines (ATMs) and adoption of better risk-management techniques.

Equation (5-12) below provides an approximation of technological progress in Zambian banking assuming technological change directly affects the banks' cost function. Put differently, we conjecture that banks have been subjected to the same form of technological shocks over time.

$$TP = -\sum \frac{\partial \ln C}{\partial t} = -\left[\theta_1 + \theta_2 t + \theta_3 \ln(sec) + \theta_4 \ln(loans) + \theta_5 \ln\left(\frac{w_1}{w_2}\right) + \theta_6 \ln\left(\frac{w_3}{w_2}\right)\right] \quad (5\text{-}12)$$

Since we are using a cost function, technological progress occurs when $TP < 0$ and when $TP > 0$ technological change is regressive. From (5-12), pure technological change is given by the first two elements on the right hand side. The second two elements denote non-neutral technical change while the last two terms measure scale augmented technological progress. The novelty of the above specification is that it does not require knowledge of the timing of these technological changes, given that it is hard to pinpoint when banks adopt those technologies and at what point effects on the cost structure take hold.

3.2.4 Data description and sources

In order to capture bank-specific cost efficiency, we employ a richly assembled bank level quarterly unbalanced panel data spanning 9 years (1998:1 – 2006:4) for all banks operating in Zambia as at 2006. This period is especially special in the Zambian banking industry because it encompasses years subsequent to the tightening of prudential regulations and enactment of banking legislation in response to the banking crisis of the 1990s. The sample also covers periods of economic turbulence and relative macroeconomic stability and robust economic growth especially since 2002. Accordingly, it is expected that these economic dynamics would have altered the banks' behaviour in a significant manner and should be reflected in efficiency performance.

There is generally no consensus in the literature on what constitutes inputs and outputs in banking. Therefore this issue is of ongoing debate. However, two main approaches have dominated the literature. The first method is the production approach which describes banking activities as the production of services. In this approach, physical quantities of labour and capital are used as inputs to produce physical quantities or number of processed loan and deposit accounts. Thus the production approach treats deposits as outputs. In contrast, the intermediation approach proposed by Sealey and Lindley (1977) recognises the banks' unique characteristics as financial intermediaries. In this approach, banks incur interest and operating expenses as inputs in producing revenue generating services (loans and other investments) as outputs. Of the two measures, the intermediation approach has been commonly used in the conventional banking cost efficiency literature (Shanmugan & Das, 2004).[8] This study takes a similar view in measuring inefficiency in the Zambian banking industry. Banks' total costs (TC) are approximated by the sum of interest and operating expenses derived from the banks' monthly consolidated income statement. The three inputs used are labour, funds and capital with the corresponding prices defined below:

(i) Labour costs $(w_1) = \dfrac{all\ personell\ expenses}{total\ assets}$

(ii) Funding costs $(w_2) = \dfrac{total\ interest\ expenses}{total\ deposits\ \&\ other\ borrowed\ funds}$

(iii) Capital costs $(w_3) = \dfrac{all\ other\ expenses}{fixed\ and\ other\ assets}$

In line with the literature, banks are modelled as multi-output financial intermediaries who collect funds from surplus agents and channel them to the most profitable investment projects at minimum costs. Accordingly, for the purpose of this study, bank output is defined as a sum of net loans, overdrafts and interbank placements $(loans)$ reflecting the gain from the financial intermediation process; investments in government securities (sec) to capture

[8] See also Freixas and Rochet (1997) and Heffernan (1996) for a discussion of the approaches in measuring outputs for the banking sector.

output from investments in risk-free assets.[9] Coefficients on input prices and outputs are expected to be positive, implying that total bank costs rise with increased use of inputs and production. In order to capture the direct effect of the quality of loan output on operating costs and measured inefficiency we use the banks' loan loss provisions as a proportion of gross loans $(lossgls)$ as a proxy for banks' credit risk. We expect a positive impact of credit risk on costs of production.

In order to model banks' heterogeneity emanating from possible scale effects on operating costs we use the number of bank branches, $(branch)$.[10] Often large banks also boast a widespread branch network, with significant implications on cost and efficiency. By incorporating the number of bank branches, we postulate that banking production technology differs in a significant way due to variations in size and other unmeasured factors associated with fixed cost and efficiency effects associated with maintaining these branches. A positive coefficient on the number of branches would indicate an upward shift in the cost curve driven by differences in size. Accordingly, bank heterogeneity would be interpreted as an important part of efficiency performance in Zambian banking.

The flow of deposits into the banking sector determines to a greater degree the amount of loans granted. Furthermore, banks that rely more on deposits to finance assets face a higher funding risk than those that hold a relatively higher proportion of equity capital. These factors have implications on bank cost performance and require a careful treatment. Therefore we include a measure of funding risk $(fundrisk)$ in the cost function to capture this effect. This variable is measured as the proportion of total deposits to gross loans and is expected to yield a positive sign implying that greater reliance on interest bearing deposits (especially) relative to total loans leads to soaring bank expenses.

An important aspect of the single stage estimation is the ability to relate bank-specific and other factors to inefficiency effects. Using this approach, we investigate the effect of regulatory intensity on banks' inefficiency by correlating estimated inefficiency effects to measures of regulatory pressure and other factors, including macroeconomic policy stance.

---

[9] We use net loans to take into account the quality of credit in measuring banking output. Using gross loans assumes that there is no damage to bank credit over the sample period. A separate measure of the quality of bank credit is included in the estimation.

[10] In some studies, the log of assets is used to capture bank size. However, this leads to a misspecification problem given that bank output is measured by the value of net loans and securities. Consequently, using assets introduces problems of multicollinearity.

Specifically for the Zambian banks, it is important to note that changes in regulatory policy were meant to safeguard the banking system from another crisis. However, these regulatory adjustments may have also produced some unintended adverse efficiency outcomes since they were a product of a convolution of forces, especially as pointed out, to protect the banking industry from further damage. In this regard, we surmise that the changes to the regulatory environment have not helped banks in their cost minimisation objective.

To test this conjecture, regulatory distortion factors are proxied by two variables, namely the interest foregone or opportunity costs on mandatory reserve requirements measured as a ratio of interest expenses $(oppcost)$ and a dummy variable capturing the effect of capital adequacy regulatory burden $(capdum)$. Zambian banks are subjected to compulsory reserve requirements which serve a dual purpose. Statutory reserves are used as a monetary policy tool and have a significant regulatory influence in fostering banking safety and soundness. In order to capture the latter effect, a risk-free Treasury bill rate has been used to calculate the interest foregone on statutory reserves. In including the regulatory capital dummy variable, we seek to measure banks' efficiency responses to regulatory burden. Therefore, a threshold value of 10 % for the Tier I capital-asset ratio is employed to set apart banks with lower capital requirements (less than 10 %) from those with high capital compliance ratio (10 %).[11] The dummy variable carries a value of one for the 10 % ratio and zero for capital ratio below 10 %.

Given the banks' reliance on credit to government for continued profitability, we have included a measure of asset concentration $(contass)$ as another control variable in the inefficiency model. This variable captures the intensity of overinvesting in government debt vis-à-vis traditional intermediation activities of the banking sector. Thus, in essence we are evaluating the impact of disintermediation on inefficiency. A positive coefficient associated with $contass$ indicates increasing production costs due to financial disintermediation as defined herein.

Conventional economic wisdom holds that public ownership of banks reduces the bank managers' incentive for cost reduction, thereby aggravating inefficiency. For example, Bhaumik and Dimova (2004) have argued that even when faced with a level playing field in

[11] Zambian banks are largely assessed based on the 10 % capital-asset ratio rather than the 8 % recommended under the Basle Accord. The higher capital ratio is meant to provide cushion to lenders in an event of a bank facing insolvency.

terms of regulatory policy, public managers experience less intense performance pressures than their private peers. Consequently, this creates a problem of moral hazard on the part of public managers since there is often no fear of job losses or erosion of managerial reputation arising from poor performance. This reduces the level of vigilance on efficiency performance by encouraging unproductive activities which undermine cost rationalisation. On the other hand, by analysing banking efficiency based on public choice theory, Sarkar, *et al* (1998) have argued that the existence of such performance pressures compel private bank managers to adopt better strategies that reduce costs, raise productivity and efficiency. However, when competition is intense, ownership structure matters less in performance analysis (Bhaumik & Dimova, 2004). Thus, in a less competitive banking environment such as that obtaining in Zambia, we would expect state ownership of banks to be a significant determinant of efficiency performance.

In line with the above public choice argument, we test the hypothesis that state owned banks are less efficient than privately owned competitors. Accordingly, an ownership dummy $(ownership)$ is included to evaluate the impact of state ownership on banks' cost inefficiency. The binary value of 1 is for local state banks, zero otherwise. We have also included the rate of inflation $(inflation)$ to capture effects of macroeconomic uncertainty and policy stance on banks' inefficiency. Table 1 gives summary statistics of variables used in the estimations.

Table 1: Descriptive statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Total operating expenses (K'mn)	484	3010.94	3851.99	253.67	20839.67
Investment in Securities (K'mn)	484	70243.58	95889.54	0.00	523986.00
Gross loans (K'mn)	484	105777.40	167610.20	55.00	1126077.00
Loan losses (K'mn)	484	9295.57	25302.11	20.33	356227.30
Net Loans (K'mn)	484	96715.22	149912.50	55.00	899164.00
Statutory reserves (K'mn)	462	16320.38	24130.82	5.93	115550.80
Opportunity cost of reserves (K'mn)	462	695.85	1065.37	0.51	5794.02
Unit labour cost (wage rate)	484	0.01	0.00	0.00	0.03
Unit funding price	481	0.01	0.12	0.00	2.67
Unit cost of capital	484	0.11	0.41	0.08	5.76
Funding risk	484	12.24	35.48	0.00	416.78
Number of bank branches	484	12	13	1	43
Asset concentration index	484	2.75	8.40	0.00	152.00
CPI Inflation (%)	484	20.87	5.85	8.20	30.60
Treasury bill rate (% per annum)	484	27.30	12.51	5.50	53.40

Source: Bank of Zambia and author's estimates

4.0 Empirical results of cost efficiency characteristics in Zambian banking

The ensuing discussion reports results obtained from the translog stochastic frontier cost function and the accompanying inefficiency model. First we discuss results of the translog cost function in the next section. Then second we turn to estimates of the inefficiency model before discussing properties of scale economies and technological change in banking. The cost frontier and inefficiency model are estimated in a single step. In correlating inefficiency scores to bank-specific and other exogenous factors we pay particular attention to the impact of regulatory effects and ownership forms and quantify their effects on bank inefficiency. Different hypotheses tests are also conducted to arrive at the suitable model that fits the data well. Consequently, the reported results are based on the chosen specification taking into account the assumptions made about the inefficiency error component. In this vein, a more general unrestrictive truncated normal distribution is assumed. The advantage of the truncated normal distribution is that it allows for a simultaneous estimation of the stochastic frontier function and analysis of the determinants of the inefficiency effects under very general conditions, thus nesting the half-normal distribution as a special case (Fujii, 2001).

4.1.1 Estimation results of the translog stochastic frontier cost function

Empirical results of the cost frontier specification are presented in Table 2 below. All estimations were performed by maximum likelihood function incorporated into Stata 10.0 following the parameterisation by Battese and Coelli (1995). Alternative specifications and assumptions about the distribution of the one-sided error term were also tried. Specifically, a more restrictive half-normal distribution of the inefficiency effects and a specification assuming heteroscedasticity (Bottasso & Sembenelli, 2004; Hadri, 1999) were estimated but results were unreliable due to model instability arising from non-convergence even at a higher number (<5000) of iterations. Accordingly, results of these estimations are not reported. On the other hand, convergence of the single-stage frontier estimation and cost inefficiency model assuming a truncated normal distribution was automatically achieved after only 41 iterations.

Furthermore, to ascertain that the translog cost function adequately captures the data generating process for Zambian banks, a generalised likelihood ratio (LR) test on the technology defined by a restrictive Cobb-Douglas cost function was conducted. We also conducted a test for homotheticity. These tests are asymptotically distributed as a χ^2 random variable with degrees of freedom given by the number of imposed restrictions. Table 3 reports the test results. According to these results, both restrictions are rejected at 1 %, exonerating the generalised translog cost function.

Table 2: Estimates of Stochastic Frontier Cost Function

	Coefficient	Parameters	Standard error	t-statistic	p-value
Intercept	β_0	5.688	0.531	10.720	0.000***
ln(w_1/w_2)	β_1	0.920	0.120	7.670	0.000***
ln(w_3/w_2)	β_3	-0.049	0.105	-0.465	0.642
$1/2$ (ln$(w_1/w_2))^2$	β_{11}	0.154	0.037	4.142	0.000***
$1/2$ (ln$(w_3/w_2))^2$	β_{33}	0.077	0.025	3.107	0.002***
ln(w_1/w_2)ln(w_3/w_2)	β_{13}	-0.097	0.029	-3.327	0.001***
ln(sec)	δ_1	-0.051	0.113	-0.450	0.653
$1/2$ (ln$(sec))^2$	δ_{11}	0.054	0.024	2.284	0.022**
ln(sec)ln(w_1/w_2)	γ_{11}	0.026	0.018	1.461	0.144
ln(sec)ln(w_3/w_2)	γ_{13}	-0.001	0.014	-0.100	0.920
ln$(loans)$	δ_2	0.306	0.068	4.494	0.000***
$1/2$ (ln$(loan))^2$	δ_{22}	0.077	0.014	5.655	0.000***
ln$(loan)$ln(w_1/w_2)	γ_{21}	-0.025	0.014	-1.738	0.082*
ln$(loan)$ln(w_3/w_2)	γ_{23}	0.025	0.011	2.270	0.023**
ln(sec)ln$(loan)$	δ_{12}	-0.042	0.016	-2.577	0.010***
t	θ_1	0.027	0.014	1.884	0.060*
$1/2\ t^2$	θ_2	-0.002	0.000	-5.835	0.000***
t ln(sec)	θ_3	0.004	0.002	2.001	0.045**
t ln$(loans)$	θ_4	-0.003	0.001	-1.927	0.054*
t ln(w_1/w_2)	θ_5	-0.005	0.002	-2.545	0.011**
t ln(w_3/w_2)	θ_6	-0.002	0.002	-0.951	0.342
Control Variables					
ln$(fundrisk)$	φ_1	0.483	0.030	15.973	0.000***
ln$(branch)$	φ_2	0.100	0.015	6.786	0.000***
ln$(lossgls)$	φ_3	0.032	0.011	2.979	0.003***
Diagnostics					
Log likelihood function			198.431		
Wald chi square			50689.42		
σ_u^2			0.224		
σ_v^2			0.013		
σ^2			0.250		
γ			0.943		
LR test of one-sided error			110.012		
Observations			413		

Significance level: * p<0.10, ** p<0.05, *** p<0.01
Source: Author's own calculations

In view of the acceptance of the translog cost function, we then report on the maximum likelihood parameter estimates as presented in Table 3. Generally, the empirical estimates of the translog cost frontier are plausible, consistent with apriori expectations. Diagnostically, the LR test statistic of 110.01 for the one-sided inefficiency error is statistically significant at 1 % evaluated against the critical value of 17.76[12]. Accordingly the hypothesis of ordinary least squares (OLS) estimation which disregards the bank-specific inefficiency component is overwhelmingly rejected.

This result is further reinforced by the $\gamma = 0.94$ statistic which is statistically significant at 99 % confidence level, indicating that the inefficiency term dominates the random error term in the overall error variance. Thus, deviations from the frontier are largely driven by bank-specific inefficiency effects.

Table 3: Likelihood ratio tests for parameters of the stochastic frontier function

Null hypothesis (H_0)	**Log-likelihood**	**χ^2 -statistic**	**Decision**
1.Cobb-Douglas specification	80.68	224.20***	Reject H_0
2.Homotheticity	158.22	69.12***	Reject H_0

Significance level: * p<0.10, ** p<0.05, *** p<0.01

Source: Author's own calculations

Turning to model coefficients, we observe that only the normalised input price for labour (approximate wage rate) is positive and significant, implying that increases in banks' labour costs are directly reflected in higher total operating expenses. The size of the coefficient too is large, depicting an almost one-to-one mapping between labour expenses and total costs. For capital input costs, the parameter estimate is negative and insignificant. These results are generally not at variance with observed evidence in the literature, especially for less mature banking markets, where labour costs constitute a significant part of banks' costs. For example, Fu and Heffernan (2007) found a negative albeit significant parameter estimate on the unit price of physical capital, normalized by the price of labour.

Estimates for output elasticities carry mixed signs as well as level of significance. For total securities, the parameter estimate on the linearly specified term shows a negative but insignificant coefficient. However, a nonlinear estimate on the quadratic term depicts an increasing cost structure associated with incremental investment in securities. Furthermore, production of loans appears to attract higher transactions costs, possibly linked to risk

[12] Appropriate critical values for the LR test for the one-sided error component which follows a mixed chi square distribution are provided in Table I of Kodde and Palm (1986).

assessment of loan applicants (screening) and continued monitoring of repayments. The estimated elasticity on net loans is 0.31 while that on the quadratic term is 0.08. Except for the interaction term between loans and the approximate wage rate; loans and securities, and loans and the time trend, all key variables are well behaved.

In order to capture effects of possible technological progress on operating costs, a time trend was included in the cost frontier. The trend variables also capture the effect of learning by doing on operating costs. From the reported results, it is observed that the trend variable pushed the cost function above the frontier, indicating some evidence of increasing cost inefficiency in the order of some 2 % per quarter over the sample period. However, there are infinitesimal downward nonlinear effects as provided by the estimated coefficient on the quadratic trend term of 0.02 %. A simple message implicit in this result is that over time, banks' total expenses decreased at an increasing rate, probably confirming the hypothesis of presence of learning by doing effects. Although statistically significant, these effects are evidently marginal to exert have any meaningful economic interpretation for policy purposes.

The choice of control variables in the cost frontier specification was aimed at isolating the effect of some of the important bank-specific and environmental factors on operating costs. The results indicate that funding risk, expansion of branch network and provision for loan losses all account for a substantial increase in operating costs. Specifically, banks with a sizeable proportion of loan loss provisions relative to gross loans tend to incur significantly higher costs as depicted by the positive and significant elasticity (0.03) on $\ln(lossgls)$. This highlights a long held view that poor quality of credit increases the banks' costs. In view of this, improvements in the quality of loans should be resolved expeditiously to cut costs, especially those related to monitoring debtors. Strengthening institutions to protect property rights and creating incentives for assessing the credit worthiness of borrowers would in particular resolve this dilemma. Hence the newly legislated law to set up the credit reference bureau (CRB) should be seen in light of tackling these problems.

The coefficient on funding risk is all the more robust, with the estimated elasticity of 0.5. Given the way this variable has been defined we postulate that an increase in the proportion of interest bearing deposits (e.g., time and savings deposits and interbank liabilities) to gross loans, the higher the intermediation costs and by extension, the higher the operating expenses. In order to decrease the costs associated with deposit mobilisation and interbank funds, banks should be encouraged to attract high value, low interest bearing demand deposits.

Earlier it was argued that maintaining a large branch network has significant implications on costs and efficiency performance. We note from the estimates that expanding the branch network by 1 % induces an increase in operating costs in the order of 0.1 percentage points. This result corroborates recent evidence by Giokas (2008) who notes that due to the need for better quality of services, banks expand branches to get closer to customers. However, this also leads to an increase in operating costs of bank branches. Grigorian and Manole (2006) give a similar argument that in highly inflationary economic environments banks tend to increase the number of branches. However, an increase in branching activity tends to weaken banks' efficient performance by raising overhead expenses. Evidently the persistently high rate of inflation in Zambia over the sample period may well explain the reported positive and statistically significant coefficient on the logarithm of bank branches. As shown below, the rate of inflation has had a deleterious effect on cost efficiency. These cost implications may explain Barclays bank's decision in 2005 to close its non-profitable branches across the country in an effort to restructure its operating costs and retain its competitive edge in the market.[13] Therefore our results may reinforce the argument by Berger *et al.* (1997) that although over branching raises revenues from providing extra customer convenience, this comes at a cost which is reflected in disproportionately high X-inefficiencies.

Recent global moves to adopt better technologies could help banks circumvent fixed costs related to over branching by relying less on brick and mortar kind of branches, that is, through a switch to the use of automated teller machines (ATMs) and telephone and internet banking. Although setting up such technological infrastructure attracts high capital outlay initially, these costs tend to diminish over time thereby enabling banks realise significant cost efficiencies.

---

[13] Probably due to lost customer base to competitor banks, Barclays bank has within two years of closing these branches started to open new ones, in some cases in places where they had shut them down. What is paradoxical about this strategy is that on closure, they had sold the buildings to other banks, now they would have to either rent or build new structures.

4.1.2 Empirical results of the conditional mean (CM) inefficiency model

Results of the multivariate conditional mean inefficiency model are presented in Table 4.

Table 4: Parameter estimates of the conditional mean (CM) model

	Coefficient	Parameter	Standard error	t-ratio	p-value
Intercept	ϕ_0	-5.191	2.794	-1.858	0.063*
ln*(lossgls)*	ϕ_1	0.372	0.187	1.995	0.046**
ln(*contass)*	ϕ_2	0.414	0.184	2.247	0.025**
ln*(rescost)*	ϕ_3	-0.838	0.324	-2.584	0.025***
ln(*inflation)*	ϕ_4	2.444	1.024	2.386	0.017**
ownership	ϕ_5	2.554	1.243	2.054	0.040**
capdum	ϕ_6	0.595	0.403	1.477	0.140
LR test of one-sided error			110.012		
Observations			413		

Significance level: * p<0.10, ** p<0.05, *** p<0.01

Source: Author's own calculations

From Table 4, parameter estimates for the two measures of regulatory pressure indicate that neither of the two variables performs according to prior expectations. While the dummy variable $(capdum)$ for capital adequacy regulations has a positive sign, it is statistically insignificant at conventional levels. This finding indicates that although compliance with capital regulation has positive cost implications, these regulatory distortions are small to affect bank efficiency performance. Based on this evidence it is difficult to support the view that regulatory intensity has been the bane of inefficient performance in the Zambian banking sector. Furthermore, we observe a negative and highly significant estimate on $\ln(oppcost)$. Although this result appears counterintuitive, for the Zambian banking sector it can be defended by noting that adjustments to the statutory reserve ratio have been far apart and smaller in magnitude, even as bank deposits have increased substantially. Consequently, the increase in actual statutory reserve placements at the central bank was lower than the growth in deposit liabilities. Simultaneously, the Treasury bill rate declined drastically over the study period.

These forces meant that the opportunity cost (interest forgone) of mandatory reserve requirements decreased, creating downward pressure on cost inefficiency. This finding is buttressed by the fact that Zambian banks boast of large idle excess reserves, emanating mainly from low intermediation. To summarise, we would argue that from the efficiency perspective the current regulatory policy framework is less distortionary. Instead, internal factors related to bank management and organisational behaviour were the main cause of inefficiency in the sector. A volatile macroeconomic has also been responsible for slow improvement in efficiency performance. These issues are highlighted below.

The parameter estimate for $\ln(contass)$ is positive and statistically significant suggesting that a high proportion of securities relative to gross loans poses concern for the banks' inefficiency. In a related study, Hauner (2008) observed that banks credit to government dampens banks' efficiency performance in developing countries than in industrial countries. For Zambian banks, this finding is particularly timely in the wake of recent efforts to reduce the stock of debt in the domestic banking sector. Therefore this move should fuel efficiency gains in the industry.

The hypothesis that state ownership of banks leads to inefficiency cannot be rejected as shown by the statistically significant estimate on the binary dummy variable, *ownership*. The parameter estimate indicates that relative to other forms of bank ownership, state banks are 2.6 times more inefficient. This result gives impetus to privatisation reforms recently concluded in Zambia when the only large state owned bank was sold to Rabobank, a Dutch cooperative bank. The result also corroborates findings of other research which shows that public banks tend to be less efficient than their privately owned peers (Perera, Skully, & Wickramanayake, 2007). The results are also consistent with the public choice theory discussed earlier in the context of the work by Sarkar *et al*. (1998).

Finally, controlling for output quality using $\ln(lossgls)$ we note that risk taking in banking proxied by loan loss provisions reduces banks' efficiency. This is manifested by a positive and statistically significant (at 5 %) of the estimated coefficient. The magnitude of this parameter (0.4) attests to the fact that banks need to heed calls for prudent lending in order to curb growth in past due loans, which affect banks' earnings. This also requires strengthening the institutional and regulatory framework, including setting up an efficient judicial system to deal with loan recovery and/or disposal of collateral. At the bank level, improved screening and monitoring techniques would be the first stop in improving quality of loans and should be reflected in short-term efficiency gains.

For the rate of inflation, the measure of macroeconomic uncertainty and policy stance, we observe an unambiguous positive impact on banks' inefficiency. Clearly, a volatile macroeconomic environment is counterproductive to banks' efficiency improvement. In particular, the magnitude of the estimate on the log of inflation validates calls for macroeconomic stability as a pre-requisite for an efficiently functioning banking system. Indeed the Turkish case studies above showed that uncertainty in the macroeconomic setting slows down the recovery of banks' performance, especially if banks are emerging from a crisis (Denizer, Dinc, & Tarimcilar, 2007).

4.1.3 Analysis of average bank level efficiency scores

Table 5 below presents a ranking of average bank-specific cost efficiency indices estimated from the translog frontier cost function[14]. The efficiency estimates show that for the banking sector as a whole, banks operated below the frontier, with mean cost efficiency of 0.90. The implication is that Zambian banks could improve their performance by saving up to 11.4 % in operating costs per quarter if they were all utilising the best practice technology over the sample period.

Table 5: Mean cost efficiency of Zambian banks

Bank id	Mean Efficiency	Std deviation	Rank
A	0.957	0.016	1
B	0.956	0.015	2
C	0.955	0.011	3
D	0.943	0.016	4
E	0.816	0.140	13
F	0.849	0.101	12
G	0.931	0.030	6
H	0.861	0.126	11
I	0.879	0.081	10
J	0.910	0.061	8
K	0.916	0.051	7
L	0.916	0.064	7
M	0.626	0.241	14
N	0.881	0.099	9
O	0.936	0.033	5

Source: Author's estimates

[14] In the Battese-Coelli specification framework, inefficiency scores range between 1 (depicting the most efficient bank) and infinity (depicting a least efficient bank). To obtain efficiency scores for comparison purposes, we invert the Battese-Coelli inefficiency estimates to get scores in the [0,1] range. Still, the most efficient bank has a score of unity and zero denotes a least efficient bank. Table I in the appendix presents actual inefficiency scores obtained from the estimation of Eq. (5-9).

Individually, the estimated cost efficiency indicators show that relative to the common frontier, the most efficient bank (Bank A) had a score of 0.957, that is, approximately 6.0 % more efficient than the average bank in the sample. This shows that Bank A could cut its costs by about 4.5 % to move to the best practice frontier. Put differently, the level of inefficiency for Bank A is approximately 4.5 %. Conversely, the least performing Bank M with an efficiency score of 0.626 shows input wastage of up to 60 %. Clearly, bank M could have improved its cost efficiency by reducing costs that prevented it from moving closer to the frontier. From Table 5 it can be observed that two banks have the same cost efficiency score, ranking seventh on the scale. Other banks are generally in good stead although none was operating optimally.

The overall picture emerging from these efficiency indicators is that, although Zambian banks are inefficient on average, for most of them the scores are skewed closer to the frontier. This is clearly shown by the trend in the cost efficiency estimates depicted by Figure 1 and Table 6 below. Figure 1 illustrates mean efficiency scores for the quarterly observations; Table 6 reports aggregated annual summary statistics of these efficiency indicators. Both Figure 1 and Table 6 tell a consistent story. In Figure 1, we observe a sustained upward trend in cost efficiency, depicting some reduction in inefficiency over the sample period. It is important to reiterate that the study period coincides with significant positive developments in the economy, beginning with 2002. These developments were marked by robust economic growth and reduction in consumer inflation. Clearly, these changes may have influenced banks' performance responses.

Figure 1:Evolution of mean cost efficiency (1998:1-2006:4)

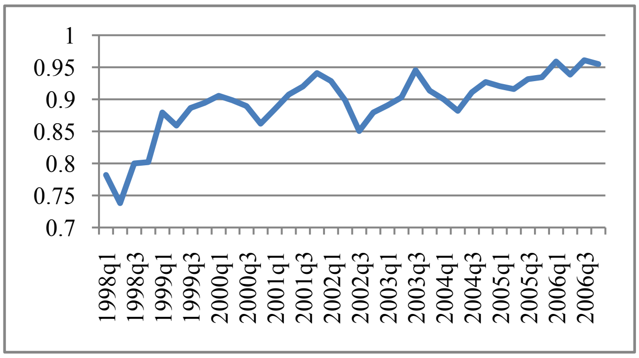


Table 6:Annual summary statistics of cost efficiency

	Mean	Std dev	Min	Max
1998	0.780	0.220	0.201	0.969
1999	0.880	0.103	0.519	0.963
2000	0.889	0.129	0.418	0.974
2001	0.914	0.093	0.471	0.971
2002	0.889	0.097	0.525	0.968
2003	0.913	0.078	0.615	0.975
2004	0.905	0.069	0.650	0.972
2005	0.925	0.071	0.601	0.978
2006	0.953	0.022	0.880	0.982

Source: Author's estimates

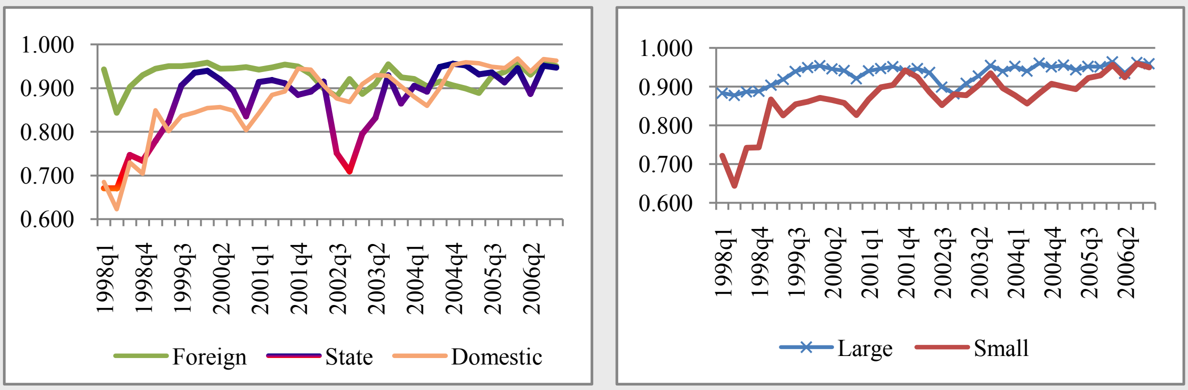
Based on sample estimates between 1998 and 2006, the mean efficiency index rose by 22.3 %. The lowest level of cost efficiency (0.74) occurred in the second quarter of 1998, probably emanating from lagged effects of internal organisational changes in some banks following an earlier crisis and uncertainty in the macroeconomic environment in the wake of the crisis. The highest efficiency rate (0.96) was recorded in the third quarter of 2006. The uptake in efficiency performance is also manifested by the corresponding reduction in the standard deviation as shown in Table 6. Although individual banks did not perform in the same fashion over the sample period, the rise in cost efficiency for all sample banks is remarkable and indicates that since the end of the crisis, the majority of banks have broadly moved closer to the best practice frontier although this shift is light. For star performers, survival strategies aimed at cost reduction and improved management practices partly explain this behavioural change. Hence, the results show that liberalisation policies are slowly bearing fruit, as observed in banks' overall better performance over the years. More importantly, the results represent the banks' strong response to incentives induced by the reforms towards managerial flexibility and functional reorganisation aimed at stemming cost escalation and inappropriate mix of resource inputs.

Figure 2 below depicts cost performance of different class of banks by ownership structure and size. Part (a) of Figure 2 shows that efficiency performance favoured foreign banks relative to domestic and state banks in the early years. Furthermore, domestic private banks also performed relatively well compared with the state owned banks. However, since 2005, efficiency indicators for all three bank categories have been slowly converging, depicting minimal variation in performance. Despite this convergence, there was still some efficiency gap between foreign banks on one hand and domestic private and the domestic state bank on the other, in the order of 5.3 % and 5.8 %, respectively. Conversely, the difference between domestic private and state banks was less than half a percentage point at 0.48 %.

Figure 2: Mean efficiency by bank ownership and size

(a) Cost efficiency and bank ownership structure

(b) Cost efficiency and bank size



Source: Author's estimates

These variations suggest that local banks in general have a lot to learn from their foreign counterparts.

As with efficiency-ownership analysis, the efficiency gap between small and large banks has also been narrowing. How long this would take for complete convergence to occur is unclear. One thing is evident, nonetheless. The level of efficiency among small banks has been increasing while that of large banks has been static for most part of the sample period. Between 1998 and 2006, small banks recorded an improvement in cost performance of a staggering 31.5 % compared with only 8.5 % for large banks. Over the last few years, cost efficiency performance for small banks has been underpinned by the merger in 2004 between two of the smaller banks. One of the two merged banks was the least efficient, posting mean efficiency of 0.77 in the three years preceding the merger, against 0.94 % for its counterpart. For the entire period, the level of efficiency for the least efficient bank of the merged group was 0.65 relative to 0.88 for its merger counterpart. In post merger period, the single bank formed by the two banks was 1.23 times more efficient than the average for the merged banks for the whole sample period preceding the merger. Although the level of assets did not increase following the merger, the improvement in cost performance of the merged bank relative to the average of the two banks corroborates the argument that bank mergers foster cost efficiency, particularly when this is accompanied by product innovation.

4.1.4 Economies of scale and technological change in Zambian banking

The core objective of the paper is the analysis of cost efficiency in banking. Therefore, in the discussions, a significant amount of space has been devoted to address this subject.

Table 7: Economies of scale and technological change in Zambian banking (period average)

	1998	1999	2000	2001	2002	2003	2004	2005	2006	**1998-2006**
					Economies of scale					
Large	2.080	1.685	1.383	1.203	1.061	0.949	0.820	0.745	0.690	**1.180**
Small	2.819	2.176	1.719	1.408	1.201	1.041	0.879	0.788	0.721	**1.394**
All	2.523	1.975	1.587	1.334	1.151	1.008	0.856	0.771	0.709	**1.312**
					Technological change					
Large	0.217	0.231	0.149	0.114	0.060	0.019	-0.131	-0.177	-0.205	**0.031**
Small	0.291	0.238	0.178	0.132	0.102	0.027	-0.108	-0.144	-0.196	**0.058**
All	0.243	0.251	0.179	0.137	0.080	0.033	-0.105	-0.158	-0.190	**0.052**

Source: Author's own estimates

To augment the above analysis, the present discussion focuses on results of the evidence of scale economies and technological change in the banking sector. Indices of global scale economies and technological change are reported in Table 7 above.

The results in Table 7 reveal the existence of economies of scale for all banks although they depict a steady downward trend over the estimation period. The estimated mean value of scale elasticity for all banks was 1.31, indicating that the banks' costs decreased by more than the increase in output. The year 1998 is particularly spectacular in that returns to scale were higher than in all other periods. This visibly contrasts with the impact of the shock of the preceding banking crisis from which many banks were just recovering. The evidence also shows that small banks recorded larger economies of scale (1.39) relative to the large banks (1.18), with the latter experiencing diseconomies of scale faster than the former group of banks in 2003. By 2004, decreasing returns had been entrenched throughout the sector. At the margin, constant returns to scale permeated the sector around 2003, with the average global scale index estimate of 1.00.

Parallel to the results on the returns to scale, estimates for technological change reveal that banks experienced technological recess for the sample period. For all banks, the average growth rate in costs attributed to technological recess was approximately 5.2 %. Relative to size, costs due to technological recess grew less proportionately for large banks than for small banks. For large banks, costs grew by an average of 3.1 % while small banks recorded an increase in costs of about 5.8 %. These results underscore the point that as technological progress takes root, large banks adapt faster to these changes, although the existence of new technologies and diversification in services may not have spurred significant reduction in costs over the sample period. In contrast, small banks responded less enthusiastically to technological incentives, thereby lagging in posting gains. These results reinforce the argument that failure to catch up in adopting better technologies may slow down the small banks' performance in the longer-term.

We also observe from Table 7 that returns to scale and technological change moved in the opposite direction, with shifts occurring at the same time around the year 2004 for most banks. As decreasing returns to scale permeate the banking industry we observe a simultaneous tipping point in technological change. The year 2004 marks the implementation of the Financial Sector Development Plan (FSDP), whose main impetus was to infuse dynamism in the Zambian financial sector. Specifically, the framework emphasised reintegration of banks' operating systems thereby providing an incentive for innovative ways of production. It also stressed the need for improvements in risk valuation and management techniques and upgrading of infrastructure. The reforms to the foreign exchange system of 2003 also accelerated the banks' adoption of better computing facilities, including installation of Reuters dealers machines to improve transparency in foreign exchange trading. Consequently, we detect a clear turnaround in technological progress, a process, if sustained could play a major role in fostering cost efficiency in the longer-term.

However, based on the available data, simple correlation analysis shows that there is a stronger relationship (-0.56) between economies of scale and cost inefficiency, on the one hand, and a much weaker link (-0.38) between technological change and cost inefficiency, on the other. This suggests that returns to scale had a more robust impact on stimulating cost efficiency among Zambian banks than was the case with technological progress. Hence, although technological progress improved from 2004, this did not induce substantial rebound in cost efficiency.

5.0 Summary and conclusions

Since the banking crisis of the mid 1990s, Zambian banks have posted significant gains, breaking the cycle of vulnerability that characterised the industry in the early years of reform. However, doubts about the efficiency of commercial banks reflected in wide interest spreads continue to be a sticking point of public policy debate. This study has evaluated the cost efficiency of Zambian banks using the stochastic frontier approach, incorporating bank-specific and environmental (including regulatory) factors in the analysis. The results show that Zambian banks are on average inefficient in the order of 11.4 %, indicating that mismanagement of resources remains a serious impediment to better cost performance. However, over the years, there has been some improvement in the relative cost efficiency of banks, with small banks displaying spectacular growth in performance. On the other hand, the efficiency gap between domestic banks and foreign banks remains wide.

Evidence of scale economies is also noted, although there has been a decline in recent years. Conversely, technological change permeated the industry almost simultaneously. Most interestingly, large banks appeared to adapt faster to technological change than small banks, even though the scale performance is less firm. Technological change was particularly stimulated by the implementation of the FSDP, which emphasised reintegration of banks' operating systems. Hence, this provided an incentive for innovative ways of production.

At a policy level, risk management and response to regulatory reforms has lagged other changes in the industry. Banks continue to exhibit poor risk assessment techniques as exemplified by higher loan loss provisions. However, regulatory intensity has not adversely impacted banks' performance, contrary to popular belief. On the other hand, macroeconomic uncertainty particularly in the early years of the sample period exacerbated the banks' cost inefficiency. These factors suggest that there should be a stronger policy response to overcome the structural impediments which continue to affect banks' cost performance.

6.0 References

Abel, A. B., Mankiw, N. G., Summers, L. H., & Zeckhauser, R. J. (1989). Assessing dynamic efficiency: Theory and evidence. *The Review of Economic Studies , 56* (1), 1-19.

Adongo, J., Stork, C., & Hasheela, E. (2005a). *Fcators influencing the alternative profit X-efficiency of Namibia's banking sector, NEPRU Working paper no.103*. Retrieved June 2007, from www.nepru.org.na: http://www.nepru.org.na/index.php?id=247&no_cache=1&file=880&uid=424.

Adongo, J., Stork, C., & Hasheela, E. (2005b). *Measuring alternative profit X-efficiency Namibia's banking sector*. Retrieved June 2007, from Namibian Economic Policy Research Unit, Policy Report, No.36: http://www.nepru.org.na/index.php?id=126&no_cache=1&file=461&uid=183.

Aigner, D., Lovell, C., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics , 6*, 21-37.

Aly, H. Y., Grabowski, R., Pasurka, C., & Rangan, N. (1990). Technical, scale, and allocative efficiencies in U.S. banking: An empirical investigation. *Review of Economics and Statistics , 72*, 211-218.

Ataullah, A., & Le, H. (2006). Economic reforms and bank efficiency in developing countries: the case of the Indian banking industry. *Applied Financial Economics , 16*, 653-663.

Ataullah, A., Cockerel, T., & Le, H. (2004). Financial liberalisation and bank efficiency: a comparative analysis of India and Pakistan. *Applied Economics , 36*, 1915-1924.

Banker, R., Charnes, A., & Cooper, W. (1984). Some models for estimating technical and scale inefficiencies in Data Envelopment Analysis. *Management Science , 30* (9), 1078-1092.

Battese, G., & Coelli, T. J. (1995). A model for Technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics , 20*, 325-332.

Battese, G., & Coelli, T. J. (1993). A stochastic frontier production function incorporating a model for technical inefficiency effects. *Working papers in econometrics and applied statistics, No.69* . Armidale: Department of Economics, University of New England.

Bauer, P., Berger, A., Ferrier, G., & Humphrey, D. (1998). Consistency conditions for regulatory analysis of financial institutions: a comparison of frontier efficiency methods. *Journal of Business and Finance , 50*, 85-114.

Beccalli, E., Casu, B., & Girardone, C. (2006). Efficiency and Stock Performance in European Banking. *Journal of Business Finance & Accounting , 33* (1-2), 245-262.

Beck, T., & Hesse, H. (2006). *Bank Efficiency, Ownership, and Market Structure: Why are Interest Spreads so High in Uganda? World Bank Policy Research Working Paper WPS4027*. Retrieved April 2007, from World Bank: http://www-wds.worldbank.org/servlet/WDSContentServer/WDSP/IB/2006/10/03/000016406_20061003154648/Rendered/PDF/wps4027.pdf

Berger, A. N. (2007). International comparisons of banking efficiency. *Financial Markets, Institutions & Instruments , 16* (3), 119-144.

Berger, A. N., & Humphrey, D. (1997). Efficiency of financial institutions: international survey and directions for future research. *European Journal of Operational Research , 98* (2), 175-212.

Berger, A. N., & Humphrey, D. (1991). The dominance of inefficiencies over scale and product mix economies in banking. *Journal of Monetary Economics , 28*, 117-148.

Berger, A. N., & Mester, L. (1997). Inside the black box: What explains differences in the efficiencies of financial institutions? *Journal of Banking & Finance , 21* (7), 895-947.

Berger, A. N., & Mester, L. J. (1997). Inside the black box: what explains the differences in efficiency of financial institutions.

Berger, A. N., Hunter, A., & Timme, S. (1993). The efficiency of financial institutions: a review and preview ofresearch past, present and future. *Journal of Banking and Finance , 17* (2/3), 221-249.

Berger, A. N., Leusner, J. H., & Mingo, J. J. (1997). The efficiency of bank branches. *Journal of Monetary Economics , 40*, 141-162.

Bhaumik, S. K., & Dimova, R. (2004). How important is ownership in a market with level playing field? The Indian banking sector revisited. *Journal of Comparative Economics , 32* (1), 165-180.

Bikker, J., & Haaf, K. (2003). Measures of competition and concentration in the banking industry: a review of the literature. *Economic and Financial Modelling* (Summer).

Bos, J. W., & Kool, C. (2006). Bank efficiency: the role of bank strategy and local market conditions. *Journal of Banking and Finance , 30*, 1953-1974.

Bottasso, A., & Sembenelli, A. (2004). Does ownership affect firms' efficiency: Panel data evidence on Italy. *Empirical Economics , 29*, 769-786.

Brownbridge, M. (1998). Financial policies and banking in Zambia. In M. Brownbridge, & C. Harvey, *Banking in Africa.* New Jersey: Africa World Press, Inc.

Brownbridge, M., & Gayi, S. (1999). Progress, constraints and limitations of financial sector reforms in least developed countries. *Discussion Paper No.7* . Geneva: United Nations Conference on Development and Trade (UNCTAD).

Buchs, T., & Mathisen, J. (2005). Competition and efficiency in banking: behavioural evidence from Ghana. *Working Paper, WP/05/17* . Washington DC: International Monetary Fund.

Cebenoyan, A. S. (1988). Multiproduct cost functions and scale economies in banking. *The Financial Review , 23* (4), 499-512.

Charnes, A., Cooper, W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research , 12*, 429-444.

Chen, X., Skully, M., & Kym, B. (2005). Banking efficiency in China: Application of DEA to pre- and post-deregulation eras: 1993–2000. *China Economic Review , 16* (3), 229-245.

Christopoulos, D., & Tsionas, E. (2001). Banking economic efficiency in the deregulation period: results from heteroscadastic stochastic frontier models. *The Manchester School , 69* (6), 656-676.

Čihák, M., & Podpiera, R. (2005). *Bank Behavior in Developing Countries: Evidence from East Africa. IMF Working Paper No.WP/05/129*. Retrieved April 2007, from International Monetary Fund: http://www.imf.org/external/pubs/ft/wp/2005/wp05129.pdf

Denizer, C. A., Dinc, M., & Tarimcilar, M. (2007). Financial liberalization and banking efficiency: evidence from Turkey. *Journal of Productivity Analysis , 27* (3), 177-195.

di Patti, E. B., & Hardy, D. C. (2005). Financial sector liberalization, bank privatization, and efficiency: Evidence from Pakistan. *Journal of Banking and Finance , 29* (8-9), 2381-2406.

Eisenbeis, R., Gary, D., & Kwan, S. H. (1999). The informativeness of stochastic frontier and programming frontier efficiency scores: cost efficiency and other measures of Bank Holding Company performance. *Working Paper, No.99-33* . Federal Reserve Bank of Atlanta.

Elliehausen, G. (1998). The cost of bank regulation: a review of the evidence. *Staff Studies* . Federal Reserve Board of Governors.

Estrada, D., & Osorio, P. (2004). *Banco de la Republica de Colombia* . Retrieved March 13, 2008, from Banco de la Republica de Colombia web site: http://www.banrep.gov.co/docum/ftp/borra292.pdf

Fare, R., Grosskopf, S., & Weber, W. L. (2004). The effect of risk-based capital requirements on profit efficiency in banking. *Applied Economics , 36*, 1731-1743.

Farell, M. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society , 120*, 252-290.

Favero, C., & Papi, L. (1995). Technical efficiency and scale efficiency in the italian banking sector:a non-parametric approach. *Applied Economics , 27*, 385-395.

Ferrier, G., & Lovell, C. (1990). Measuring cost efficiency in banking: econometric and linear programming evidence. *Journal of Econometrics , 46*, 229-245.

Fiorentino, E., Karmann, A., & Koetter, M. (2006). The cost efficiency of German banks : a comparison of SFA and DEA. *Banking and Financial Studies Discussion Paper Series 2, No, 2006:* . Deutsche Bundesbank.

Franks, J. R., Schaefer, S. M., & Staunton, M. D. (1998). The direct and compliance costs of financial regulation. *Journal of Banking and Finance , 21*, 1547-1572.

Freixas, X., & Rochet, J. (1997). *Microeconomics of banking.* Cambridge: Massachusettes Institute of Technology (MIT) Press.

Fries, S., & Taci, A. (2005). Cost efficiency of banks in transition: evidence from 289 banks in 15 post-communist countries. *Journal of Banking and Finance , 29* (1), 55-81.

Fu, X. (., & Heffernan, S. (2007). Cost X-efficiency in China's banking sector. *China Economic Review , 18* (1), 35-53.

Fujii, A. (2001). Determinants and probability distribution of inefficiency in the stochastic cost frontier of Japanese hospitals. *Applied Economics Letters , 8*, 807-812.

Giokas, D. (2008). Assessing the efficiency in operations of a large Greek bank branch network adopting different economic behaviors. *Economic Modelling , 25*, 559-574.

Griffiths, A., & Wall, S. (2000). *Intermediate Microeconomics: theory and applications.* Harlow, England: Prentice Hall.

Grigorian, D., & Manole, V. (2006). Determinants of commercial bank performance in transition: an application of data envelopment analysis. *Comparative Economic Studies , 48* (3), 497-522.

Gruben, W., & McComb, R. (2003). Privatization, competition, and supercompetition in the Mexican commercial banking system. *Journal of Banking and Finance , 27* (2), 229-249.

Hadri, K. (1999). Estimation of a doubly heteroscedastic stochastic frontier cost function. *Journal of Business and Economic Statistics , 17*, 359-363.

Hao, J., Hunter, W., & Yang, W. (2001). Deregulation and efficiency: the case of Korean banks. *Journal of Economics and Business , 53*, 237-254.

Hauner, D. (2008). Credit to government and banking sector performance. *Journal of Banking and Finance* , doi.10.1016/j.jbankfin.2007.07.012.

Hauner, D., & Peiris, S. J. (2005). Bank efficiency and competition in low-income countries: the case of Uganda. *Working Paper, WP/05/240* . Washington DC: International Monetary Fund.

Hauner, D., & Peiris, S. J. (2005). *Bank Efficiency and Competition in Low-Income: the case of Uganda, IMF Working paper WP/05/240*. Retrieved September 2006, from International Monetary Fund: www.imf.org/external/pubs/ft/wp/2005/wp05240.pdf

Heffernan, G. (1996). *Modern Banking.* London, England: John Wiley & Sons Ltd.

Herrero, I., & Pascoe, S. (2002). Estiamtion of technical efficiency: a review of some of the stochastic frontier and DEA software. *Computers in Higher Education Economics Review , 15* (1).

Hughes, J., & Mester, L. (1993). A qulaity and risk-adjusted cost fucntion for banks: evidence on the "too-big-to-fail" doctrine. *Journal of Productivity Analysis , 27*, 293-315.

Ikhide, S. (2000). *Efficiency of Commercial Banks in Namibia*. Retrieved June 2007, from Bank of Namibia: http://www.bon.com.na/docs/pub/Efficiency%20of%20Commercial%20Banks%20in%20Namibia.pdf.

Işık, I., & Hassan, M. (2002).

Issşık, E., & Akçaoğlu, E. (2006). An empirical analysis of productivity developments in “Traditional Banks”: The Initial post-liberalisation experience. *Central bank of the Republic of Turkey Review , 6* (1), 1-35.

Jameson, K. (1972). Comment on the theory and measurement of dynamic X-efficiency. *The Quarterly Journal of Economics , 86* (2), 313-326.

Jondrow, J., Lovell, K. C., Materov, I. S., & Schmidt, P. (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics , 19*, 233-238.

Kablan, S. (2007). *Measuring banking efficiencyin developing countries: the case of the Western African Monetary Union (WAEMU)*. Retrieved February 2008, from African Development Bank: http://www.afdb.org/pls/portal/docs/PAGE/ADB_ADMIN_PG/DOCUMENTS/AEC/MESURE%20DE%20LA%20PERFORMANCE%20DES%20BANQUES%20UEMOA.PDF

Kaparakis, E., Miller, S., & Noulas, A. (1994). Short-run cost inefficiency of commercial banks: a flexible stochastic frontier approach. *Journal of Money, Credit and Banking , 26*, 873-894.

Kirkipatrick, C., Murinde, V., & Tefula, M. (2007). *The measurement and determinants of X-efficiency in commercial banks in sub-Saharan Africa*. Retrieved February 2008, from World Bank: http://info.worldbank.org/etools/docs/library/239948/Murinde_Reading_CommercialBanksinAfrica.pdf

Kodde, D. A., & Palm, F. C. (1986). Wald criteria for jointly testing equality and inequality restrictions. *Econometrica , 54* (5), 1243-1248.

Kwan, S., & Eisenbeis, R. (1996). An analysis of inefficiency in banking: stochastic frontier approach. *Economic Review , 2*, 16-26.

Leibenstein, H. (1966). Allocative Efficiency vs. "X-Efficiency". *The American Economic Review , 56* (3), 392-415.

Leightner, E., & Lovell, C. (1998). The impact of financial liberalisation on the performance of Thai banks. *Journal of Economic Business , 50*, 115-131.

Leightner, J., & Lovell, C. (1998). The impact of financial liberalisation of the performance of Thai banks. *Journal of Economics and Business , 50*, 115-131.

Lensink, R., Meesters, A., & Naaborg, I. (2008). Bank efficiency and foreign ownership: do good institutions matter? *Journal of Banking and Finance , 32* (5), 834–844.

Lewin, A., & Lovell, C. (1990). Editors introduction. *Journal of Econometrics , 46* (1-2), 3-5.

Lim, G., & Randhawa, D. S. (2005). Competition, Liberalisation and efficiency: evidence from a two-stage banking model on banks in Hong Kong and Singapore. *Managerial Finance , 31* (1), 52-77.

Lin, P.-w. (2005). An empirical analysis of bank mergers and cost efficiency in Taiwan. *Small Business Economics , 25*, 197-206.

Marshall, A. (1961). *Principles of Economics* (9 ed.). London: Macmillan.

McKillop, D., Glass, J., & Ferguson, C. (n.d.).

Meeusen, W., & van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production function with composed error. *Internantional Economic Review , 18* (2), 435-444.

Mester, L. (1993). Efficiency in the savings and loan industry. *Journal of Banking and Finance , 17*.

Mester, L. (1997). Measuring efficiency at US banks: accounting for heterogeneity is important. *European Jouran of Operational Research , 98*, 230-242.

Peel, D. (1974). A note of X-Inefficiency. *The Quarterly Journal of Economics , 88* (4), 687-688.

Perera, S., Skully, M., & Wickramanayake, J. (2007). Cost efficiency in South Asian banking: The impact of bank size, state ownership and stock exchange listings. *International Review of Finance , 7* (1-2), 35-60.

Rangan, N., Grabowski, R., Aly, H., & Pasurka, C. (1988). The technical efficiency of U.S. banks. *Economics Letters , 28*, 169-175.

Rao, A. (2005). Cost frontier efficiency and risk-return analysis in an emerging market. *International Review of Financial Analysis , 14* (3), 283-303.

Resti, A. (1997). Evaluating the cost-efficiency of the Italian banking system: what can be learned from the joint application of parametric and non-parametric techniques. *Journal of Banking and Finance , 21* (2), 221-250.

Sarkar, J., Sarkar, S., & Bhaumik, S. (1998). Does ownership always matter: evidence from the Indian banking industry. *Journal of Comparative Economics , 26*, 262-281.

Sealey, C., & Lindley, J. (1977). Inputs, outputs, and a theory of production and cost at depository financial institutions. *Journal of Finance , 32* (4).

Sensarma, R. (2006). Are foreign banks always the best? Comparison of state-owned, private and foreign banks in India. *Economic Modelling , 23*, 717-735.

Shanmugan, K., & Das, A. (2004). Efficiency of Indian banks during the reform period. *Applied Financial Economics , 14*, 681-686.

Sherman, H., & Gold, F. (1985). Bank branch operation efficiency: evaluation with Data Envelopment Analysis. *Journal of Banking and Finance , 9*, 297-316.

Staikouras, C., Mamatzakis, E., & Koutsomanoli-Filippaki, A. (2007). Cost efficiency of the banking industry in the South Eastern European region. *Journal of International Financial Markets, Institutions and Money (In Press, Corrected Proof)* .

Stigler, G. (1976). "The Xistence of X-Efficiency". *American Economic Review , 66*, 213-216.

Sturma, J.-E., & Williams, B. (2004). Foreign bank entry, deregulation and bank efficiency: lessons from the Australian experience. *Journal of Banking & Finance , 28* (7), 1775-1799 .

Thankor, A. V. (1996). The design of financial systems: an overview. *Journal of Banking and Finance , 20*, 917-948.

United Nations. (2005, October). *Economic trends and impacts: Banking sector lending behaviour and efficiency in Western Asian Countries*. Retrieved May 2007, from Economic and Social Commission for Western Asia (ESCWA): www.escwa.un.org/information/publications/edit/upload/ead-05-7-e.pdf

Valverde, S. C., Humphrey, B. D., & del Paso, R. L. (2007). Opening the black box: Finding the source of cost inefficiency. *Journal of Productivity Analysis , 27* (3), 209-220.

Weill, L. (2003). Banking efficiency in transition economies: the role of foreign ownership. *The Economics of Transition , 11*, 569-592.

Weill, L. (2004). Measuring Cost Efficiency in European Banking: A Comparison of Frontier Techniques. *Journal of Productivity Analysis , 21* (2), 133-152.

World Bank. (2006). Review of World Bank Assistance for Financial Sector Reforms. Washington D.C: Independent Evaluation Group, World Bank.

Yildirim, C. (2002). Evolution of banking efficiency within an unstable macroeconomic environment: the case of Turkish commercial banks. *Applied Economics , 34* (18), 2289-2301.

Yildirim, H. S., & Philippatos, G. (2007). Efficiency of banks: recent evidence from the transition economies of Europe 1993-2000. *The European Journal of Finance , 13*, 123-143.

Zaim, O. (1995). The effect of financial liberalisation on the efficiency of Turkish commercial. *Applied Financial Economics , 5*, 257-264.

Zajc, P. (2006). A comparative study of bank efficiency in Central and Eastern Europe: The role of foreign ownership. *International Finance Review , 6*, 117-156.

Zardkoohi, A., Nanda, R., & James, K. (1986). Homogeneity Restrictions on the Translog cost model: a note. *The Journal of Finance , XLI* (5), 1153-1155.