WORKING PAPERS
RESEARCH DEPARTMENT

WORKING PAPER NO. 03-13
APPLYING EFFICIENCY MEASUREMENT TECHNIQUES TO CENTRAL BANKS

Loretta J. Mester
Federal Reserve Bank of Philadelphia
and
Finance Department, the Wharton School, University of Pennsylvania

July 2003

FEDERAL RESERVE BANK OF PHILADELPHIA
Ten Independence Mall, Philadelphia, PA 19106-1574• (215) 574-6428• www.phil.frb.org

WORKING PAPER NO. 03-13

APPLYING EFFICIENCY MEASUREMENT TECHNIQUES TO CENTRAL BANKS

Loretta J. Mester
Federal Reserve Bank of Philadelphia
and
Finance Department, the Wharton School, University of Pennsylvania

July 2003

This paper was written for and presented at the Workshop on Central Bank Efficiency, Central Bank of Sweden, Stockholm, Sweden, May 23-24, 2003.

Please send correspondence to Loretta J. Mester at Research Department, Federal Reserve Bank of Philadelphia, Ten Independence Mall, Philadelphia, PA 19106-1574, phone: 25-574-3807, fax: 215-574-4303, email: Loretta.Mester@phil.frb.org.

Applying Efficiency Measurement Techniques to Central Banks

Abstract

This paper reviews the standard techniques of efficiency measurement, discusses some of the issues that arise in applying these standard techniques to central banks, and reviews some of the literature that has attempted to apply these techniques to central banking. The uniqueness of some of the activities of central banking, the difficulty in measuring some of the central banking outputs, and the complicated and multiple objectives pursued by central banks makes application of the standard techniques problematic. However, certain central bank activities do lend themselves to efficiency measurement, e.g., payment services provision.

JEL Classification Numbers: E5, E50, E58, E61, G2, D2

Keywords: Central bank, policy objectives, efficiency, cost, profit, production

Applying Efficiency Measurement Techniques to Central Banks

Loretta J. Mester

Federal Reserve Bank of Philadelphia
and
Finance Department, the Wharton School, University of Pennsylvania

July 2003

1. Introduction

This paper discusses some of the issues that arise in applying the standard methods of efficiency measurement to central banks. There is a long literature of the application of such methods to financial institutions, in particular, to commercial banks (see Berger and Mester, 1997, and Berger and Humphrey, 1997 for discussions of the literature). Berger and Humphrey (1997) documented 130 studies of financial institution efficiency, using data from 21 countries, from multiple time periods, and from various types of institutions, including banks, bank branches, savings and loans, credit unions, and insurance companies. These papers fall into the broader literature that attempts to measure the performance of financial institutions. This substantial body of work suggests that progress has been made on this front. Much less progress has been made in explaining the differences in performance (i.e., profitability or efficiency) across institutions. The sources of such differences remain illusive: a study by Berger and Mester (1997) of U.S. banks in 1990-95 found 25 variables explained only about 7 percent of the variance of measured cost efficiency and about 35 percent of the variance of measured profit efficiency.

Many fewer papers have applied efficiency measurement to certain aspects of central banking, most particularly the payment services produced by the U.S. Federal Reserve System, including check processing and currency distribution (examples include Bauer and Hancock, 1993; Bauer and Ferrier, 1996; Bohn, Hancock, and Bauer, 2001; and Gilbert, Wheelock, and Wilson, 2002). Only a handful of papers have attempted to apply standard efficiency methods to the study of monetary policymaking. In

order to understand the issues that arise in applying efficiency measurement techniques to central banking, I will briefly review some concepts of efficiency and the methods used in measuring efficiency. I'll then discuss applications of these methods to central banking.

2. Definition of Efficiency

Efficiency measurement is one aspect of firm performance. Efficiency is measured with respect to an objective; it can be measured with respect to maximization of output, maximization of profits, or minimization of costs. Duality theory can be used to derive the cost function from the production function, and cost is a component of profit; hence, the three concepts are not independent. Scale economies, scope economies, and X-efficiency are different aspects of performance. Scale and scope economies refer to selecting the appropriate outputs, while X-efficiency refers to selecting the appropriate inputs. Typically, scale economies refer to how the firm's scale of operations (its size) is related to cost — i.e., what percentage increase in costs occurs with a 1-percent increase in scale. A firm is operating at constant returns to scale if, for a given mix of products, a proportionate increase in all its outputs would increase its costs by the same proportion; a firm is operating with scale economies if a proportionate increase in scale leads to a less than proportionate increase in cost; a firm is operating with scale diseconomies if a proportionate increase in scale leads to a more than proportionate increase in cost. Scope economies refer to how the firm's choice of multiple product lines is related to cost. A firm producing multiple products enjoys scope economies if it is less costly to produce those products together than it would be to separate production into specialized firms.[1] X-efficiency measures how productive the firm is in its use of inputs to create output. If all firms in an industry are producing the scale and combination of outputs that minimize the average cost of production, then the total cost of producing the industry's output is minimized, and the industry is producing the efficient combination and level of products, provided each firm is using its inputs efficiently. Firms that exhibit X-inefficiency are either wasting some of their inputs (technical inefficiency), or are using the wrong combination of inputs to

produce outputs (allocative inefficiency), or both. Management ability (or lack thereof) may be a source of X-inefficiency, but managerial preferences might be another source, to the extent that managers can pursue objectives that differ from those of stockholders. For example, managers might derive utility from having large staffs or other perquisites, as well as high profits, so that $U=U(\pi,E)$, where π is profits and E is expenditure on labor (or other input). Some studies of commercial banks and savings and loans have found evidence of such "expense-preference" behavior; others found evidence of "empire building," i.e., pursuit of inefficient mergers to gain larger scale and presumably prestige (see Edwards, 1977; Mester, 1989a, 1989b, 1991; and Hughes, Lang, Mester, Moon, and Pagano, 2003). In the following discussion, I will focus on measures of X-efficiency, rather than scale and scope economies.

In commercial banking in the U.S., many studies have found large X-inefficiencies, on the order of 20 percent or more of total banking industry costs, and about half of the industry's potential profits. The estimates often vary substantially across studies according to the data source, as well as the efficiency concepts and measurement methods used in the studies.

As a general definition, efficiency is a measure of the deviation between actual performance and desired performance. Thus, efficiency must be measured relative to an objective function. Much of the literature focuses on simple objective functions, like output maximization, cost minimization, or profit maximization, but other studies acknowledge the fact that the objectives of firm management may differ from these and try to incorporate this into efficiency measurement, or focus on more market-based definitions of efficiency, e.g., operation on a risk-return frontier (see Hughes, Lang, Mester, and Moon, 2000, and Hughes, Mester, and Moon, 2001).

A fundamental decision in measuring financial institution efficiency is which concept to use, and the choice will depend on the question being asked. The concept chosen should be related to economic optimization in reaction to market prices and competition, rather than being based solely on the use of technology. We can ask the question, is the firm maximizing the amount of output it produces given its

[1] Note that I've defined both of these concepts relative to the costs of production, but they could just as well have been defined relative to the profitability of production. That is, what effect does an increase in scale or scope have

inputs or minimizing the amount of inputs it uses to produce a given level of output — i.e., is it operating on its production frontier — but that is a question about technological optimization. This is less interesting from an economic perspective, since it ignores values. It cannot account for allocative inefficiency in misresponding to relative prices in choosing inputs and outputs, and it is difficult to compare firms that tend to specialize in different inputs or outputs, because there is no way to compare one input or output with another without the benefit of relative prices. There is also no way to determine whether the output being produced is optimal without value information on the outputs.

Instead, we would like to investigate questions of economic optimization. For example, is the firm minimizing its costs of production given its choice of inputs, taking input prices as given; is the firm maximizing its profits given its choice of inputs and outputs, taking input and output prices as given. A firm might be operating on its production frontier (i.e., not wasting resources), and so be technically efficient, but could still be allocatively inefficient if it is choosing the wrong mix of inputs given the relative prices of those inputs. Similarly, the firm could be technically and allocatively efficient in producing its chosen level of output, but choosing the wrong level of output in order to maximize profits.

Figure 1 shows a simple two input, one output case of firm production. The figure shows an isoquant — the combinations of input x_1 and x_2 (say labor and capital) it takes to make output level y_0. Firm B is technically efficient, since it is operating on the isoquant. Firm A is inefficient, since it is operating interior to the isoquant. That is, Firm A is using more of inputs x_1 and x_2 to produce y_0. But note that Firm B could do better as well. Firm B could lower its costs of producing y_0 by using a different combination of the inputs, given their prices w_1 and w_2. Namely, Firm B would minimize its cost of producing y_0 by operating at point O — given the prices of the inputs, Firm B should use more x_1 and less of x_2. Since we want to capture such allocative inefficiency, we'll want to focus on the economic concepts of cost-minimization and profit-maximization, which are based on economic optimization in reaction to market prices and competition, rather than based solely on the use of technology.

---

on profit?

2.1 Cost Efficiency. Cost efficiency gives a measure of how close a firm's cost is to what a best-practice firm's cost would be for producing the same output bundle under the same conditions. It is derived from a cost function in which variable costs depend on the prices of variable inputs, the quantities of variable outputs and any fixed inputs or outputs, environmental factors, and random error, as well as efficiency. Such a cost function is often written in logarithmic form:

$$ln\ C_i = ln\ f(y_i, w_i, z_i, h_i) + u_i + v_i, \quad (1)$$

where C measures variable costs, w is the vector of prices of variable inputs, y is the vector of quantities of variable outputs, z indicates the quantities of any fixed netputs (inputs or outputs, like physical plant, which cannot be changed quickly), h is a set of environmental or market variables that may affect performance (e.g., regulatory restrictions), u_i denotes an inefficiency factor that may raise costs above the best-practice level, and v_i denotes the random error that incorporates measurement error and luck that may temporarily give firms high or low costs. The inefficiency factor u_i incorporates both allocative inefficiencies from failing to react optimally to relative prices of inputs, w, and technical inefficiencies from employing too much of the inputs to produce y.

The function f denotes some functional form and represents the best-practice frontier. The term, $u_i + v_i$, is treated as a composite error term: v_i is a two-sided error, since random measurement error or luck can be positive or negative, and u_i is a one-sided (positive) error, since inefficiency means higher costs. The various X-efficiency measurement techniques use different methods to identify the inefficiency term, u_i, as distinct from the random error term, v_i.

The cost inefficiency of any Firm i would be measured relative to the best-practice frontier. Note that the best-practice frontier refers to the best practice observed in the industry and not true minimum cost, which is not observable. Conceptually, the cost inefficiency of Firm i measures the percentage increase in cost of Firm i, adjusted for random error, relative to the estimated cost needed to produce Firm i's output vector if the firm were as efficient as the best-practice firm in the sample facing the same

exogenous variables (w,y,z,v).[2] It can be thought of as the proportion of costs or resources that are used inefficiently or wasted. Figure 2 shows an example. The estimated cost frontier is given by f. Firm j is fully efficient. Its actual cost lies below the frontier due to random error. Firm i is inefficient. The difference in bank i's cost and the frontier value at the same y is due to both random error and inefficiency.

2.2 Standard Profit Efficiency. Cost minimization is not the only goal of the firm. The firm should minimize the cost of producing a given output bundle, but that output bundle should be chosen to maximize profits. Standard profit efficiency measures how close a firm is to producing the maximum possible profit given a particular level of input prices and output prices (and fixed netputs and environmental variables). In contrast to the cost function, the standard profit function specifies variable profits in place of variable costs and takes variable output prices as given, rather than holding all output quantities statistically fixed at their observed, possibly inefficient, levels. That is, the dependent variable in the profit function allows for consideration of revenues that can be earned by varying outputs as well as inputs. Output prices are taken as exogenous, allowing for inefficiencies in the choice of outputs when responding to these prices or to any other arguments of the profit function.

The standard profit function, in log form, is:

$$ln\ (\pi+\theta)_i = ln\ g(p_i, w_{i,} z_i, h_i) - u_{\pi i} + v_{\pi i,} \qquad (2)$$

where π is the variable profits of the firm; θ is a constant added to every firm's profit so that the natural log is taken of a positive number; p is the vector of prices of the variable outputs; $v_{\pi i,}$ represents random error; and $u_{\pi i}$ represents inefficiency that reduces profits. Similar to cost inefficiency, profit inefficiency is defined as that amount of profit that is not being earned compared to the predicted maximum profit that could be earned if the firm was as efficient as the best-practice firm. Thus, it is the percentage of profits that is left on the table, so to speak.

As discussed in Berger and Mester (1997), profit efficiency is a more comprehensive measure of performance than is cost efficiency, since it accounts for errors on the output side as well as those on the

[2] To see this, note that, ignoring random error, $u_i = ln\ C_i - ln\ f(y_i, w_i, z_i, h_i)$.

input side. It is based on the economic goal of profit maximization, which requires that the same amount of managerial attention be paid to raising a marginal dollar of revenue as to reducing a marginal dollar of costs. That is, a firm that spends $1 additional to raise revenues by $2, all else held equal, would appropriately be measured as being more profit efficient but might inappropriately be measured as being less cost efficient. Note that cost efficiency evaluates performance holding output constant at its current level, which generally will not correspond to an optimum. A firm that is relatively cost efficient at its current output may or may not be cost efficient at its optimal output, which typically involves a different scale and mix of outputs. Standard profit efficiency embodies the cost inefficiency deviations from the optimal point, as well as revenue inefficiencies.[3]

2.3 More Complicated Objectives. As we discussed earlier, the standard concepts of cost minimization or profit maximization may not be the only goals being pursued by the firms' managers and some studies have incorporated more complicated objectives.

The goals of cost minimization and profit maximization may not be general enough. Explicitly recognizing the tradeoff between return and risk, where risk is a choice variable of the firm, would seem to be an important consideration for financial institutions (see Hughes, 1999, and Hughes, Lang, Mester, and Moon, 2000, and Hughes, Mester, and Moon, 2001). For example, an increase in a bank's scale of operations may allow it to reduce its exposure to both credit and liquidity risk through diversification. All else equal, this could mean scale economies in risk management costs. But all else is not equal: by reducing the risk attached to any given production plan, better diversification can decrease the marginal cost of risk-taking and lead banks to take on more risk to earn a greater return. Not accounting for risk when specifying the production structure can obscure scale economies, since additional risk-taking is costly in terms of the additional resources needed to manage the risk and the higher risk premium that has

[3] Berger and Mester (1997) discuss another type of profit efficiency, alternative profit efficiency. This concept is based on estimates of the alternative profit function, which substitutes output levels for output prices in the specification of the profit function. This function is estimated to provide additional information when the maintained assumptions underlying the standard profit function do not hold. It may provide useful information if there are unmeasured differences in output qualities across firms; outputs are not completely variable; output markets are not perfectly competitive; or output prices are not accurately measured.

to be paid to attract uninsured funding. (Hughes, Mester, and Moon, 2001, show that accounting for risk can uncover substantial scale economies at U.S. commercial banks.) When exposure to risk is influenced by production decisions, then cost minimization and profit maximization need not coincide to value maximization. Estimates of efficiency that are derived from cost and profit functions may be mismeasured, since they do not penalize suboptimal choices of risk and quality that then affect prices.

If firms take risk as well as profit into account when making production decisions, then the model of production that one evaluates efficiency against would need to include this. Hughes, Lang, Mester, Moon (2000) construct a model of firm production that incorporates the risk-return tradeoff. Managers' most preferred production plan maximizes a utility function that accounts for how the probability distribution of profit depends on the production plan. Duality theory is used to derive the most preferred input and profit demand equations. These demand functions are those that maximize the managers' utility function.

Hughes, Mester, Moon (2001) develop measures of efficiency based on the expected return-risk tradeoff implied by the production model. ER is the firm's predicted profit, as calculated from the estimated profit-share equation from the model, divided by the firm's equity level. RK is the standard error of predicted profit divided by equity. The authors show that ER and RK are systematically related to the market value of equity for the subsample of publicly traded banks, so they can be used to derive market return efficiency measures. A risk-return frontier is then estimated:

$$ER_i = \Gamma_0 + \Gamma_1 RK_i + \Gamma_2 RK_i^2 + v_i - u_i, \qquad (3)$$

where v_i is a two-sided error term representing random error, and u_i is a one-sided error term representing inefficiency. An inefficiency measure based on this frontier would give the increase in expected return that would occur if the firm moved to the frontier, holding risk constant. That is, it identifies lost potential return given the firm's level of return risk. One can identify the group of banks that are most

efficient (say, the quarter of banks with the lowest levels of measured inefficiency) as those that are value-maximizing banks.[4]

Note that even with more complicated objectives, the efficient frontier takes on a general form (see Hughes, Lang, Mester, Moon, 2000). If X_i denotes a measure of the financial performance of firm i, e.g., profit, or the market value of its assets, and G_i denotes a measure defining the peer group used to compare firm i's financial performance, e.g., risk or the replacement cost of assets, the general form of the frontier, which gives the highest potential value of X_i given G_i is:

$$X_i = \alpha_0 + \alpha_1 G_i + \alpha_2 (G_i)^2 + v_i - u_i,$$

where v_i is a two-sided random error term with zero mean, and u_i is a one-sided error term representing inefficiency. (Note that more flexible function forms than the quadratic could be specified.) For example, financial performance, X, might be measured by predicted profit from an estimated model and G might be measured by risk, e.g., the firm's interest-rate beta, or by size, e.g., its equity or asset level. Note that for any G, the optimality of the choice of G is not taken into account when measuring efficiency. That is, if G is risk, then a firm's performance would be compared only to those taking on the same level of risk. The firm would not be penalized for a nonoptimal choice of risk that lowered performance.

3. Efficiency Measurement Methods

Even after the appropriate concept or goal against which efficiency is to be evaluated is chosen, certain issues need to be confronted before the estimates can be obtained. These include estimation technique; specification of the functional form of the frontier; variables to include in the frontier; and data measurement issues.

---

[4] Hughes, Lang, Mester, and Moon (1996) present two other efficiency measures: a measure of the decrease in risk that would occur if the firm moved to the frontier along the ray orthogonal to the frontier relative to risk at the frontier (the orthogonal ray is the shortest path to the frontier); a measure of the increase in expected return that would occur if the firm moved to the frontier on the ray orthogonal to the frontier relative to expected return at the frontier. These measures cannot account for random error's effect on the placement of the bank relative to the frontier.

Hughes, Mester, and Moon (2001) present an additional two efficiency measures. For publicly traded bank holding companies they derive an efficiency measure based on estimating a frontier that relates the market value of assets to the book value of assets, and they derive another efficiency measure based on estimating a frontier that relates the market value of equity to the book value of equity. These measures indicate the bank holding company's lost potential market value of equity or assets based on the book value of equity or assets, respectively.

3.1 Estimation Techniques. Here we'll discuss the methods used for estimating frontiers (i.e., identifying the inefficiency component from the random noise component in frontier estimation).

Common frontier efficiency estimation techniques are data envelopment analysis (DEA), free disposable hull analysis (FDH), the stochastic frontier approach, the thick frontier approach, and the distribution-free approach. The first two of these are nonparametric techniques, and the latter three are parametric methods.

My preference is for the parametric techniques. The nonparametric methods generally ignore prices and can, therefore, account only for technical inefficiency in using too many inputs or producing too few outputs (as discussed above). Another drawback is that they usually do not allow for random error in the data, assuming away measurement error and luck as factors affecting outcomes (although some progress is being made in this regard by using bootstrapping methods). In effect, they disentangle efficiency differences from random error by assuming that random error is zero! To see the effect of measurement error, consider Figure 3. Firm C's reported data imply that it is not on the frontier, and that Firm B is more efficient than Firm C. But its data are measured with error. Once measurement error is taken into account, Bank C is actually more efficient than Bank B.

In the parametric methods, a bank is labeled inefficient if it is behaving less optimally with respect to the specified goal — e.g., costs are higher or profits are lower — than the frontier value. The estimation methods differ in the way u_i is disentangled from the composite error term $u_i + v_i$.

In the stochastic frontier approach, the inefficiency and random error components of the composite error term are disentangled by making explicit assumptions about their distributions. The random error term, v_i, is assumed to be two-sided (usually normally distributed), and the inefficiency term, u_i, is assumed to be one-sided (usually half-normally distributed). The parameters of the two distributions are estimated and can be used to obtain estimates of firm-specific inefficiency. The estimated mean of the conditional distribution of u_i given $u_i + v_i$, i.e., $\hat{u}_i \equiv \hat{E}(u_i \mid (u_i + v_i))$ is usually used to measure inefficiency.

The distributional assumptions of the stochastic frontier approach are fairly arbitrary, and sometimes the residuals are not skewed in the direction predicted by the assumptions of the stochastic frontier approach.

If panel data are available, some of these maintained distributional assumptions can be relaxed, and the distribution-free approach may be used. This method assumes that there is a core efficiency or average efficiency for each firm over time. The core inefficiency is distinguished from random error (including any temporary fluctuations in inefficiency) by assuming that core inefficiency is persistent over time, while random errors tend to average out over time. In particular, a cost or profit function is estimated for each period of a panel data set. The residual in each separate regression is composed of both inefficiency, u_i, and random error, v_i, but the random component, v_i, is assumed to average out over time, so that an estimate of the inefficiency term, $\hat{u} \equiv$ the average of a firm's residuals from all of the regressions = average (u_i+v_i) = average (u_i).[5] The reasonableness of the maintained assumptions about the error term components depends on the length of period studied. If too short a period is chosen, the random errors might not average out, in which case random error would be attributed to inefficiency (although truncation can help). If too long a period is chosen, the firm's core efficiency becomes less meaningful because of changes in management and other events, i.e., it might not be constant over the time period.

3.2 Functional Form, Variable Selection, and Variable Measurement. The next step in the parametric estimation methods is choice of functional form for the frontier, including variable selection and measurement. The most popular form in the literature for cost and profit functions is the translog. The Fourier-flexible functional form augments the translog by including Fourier trigonometric terms, which makes it more flexible than the translog. Berger and Mester (1997) found that there was only a small difference in average efficiency and very little difference in efficiency dispersion or rank between cost or profit efficiency estimates based on the translog functional form and those based on the Fourier-

flexible functional form. While formal statistical tests indicated that the coefficients on the Fourier terms were jointly significant at the 1-percent level, the average improvement in goodness of fit was small and was not significant from an economic point of view.

Once the objective and functional form are selected, the next decision is the variables to include in the function and proxies for those variables. Ideally, the frontier to be estimated should be derived from first principles. For example, if the objective is cost minimization, the cost function should be derived based on the specified production technology. Variables to include in the cost function would be those indicated by the theory of duality ─ output levels, input prices, netputs (factors that the firm cannot vary over the shortrun, which are measured in levels), and environmental variables (to account for differences across the firms' environments or markets, which may affect performance but are not a choice for firm management). For example, Hughes, Lang, Mester, and Moon (2000) derive the profit and input demand functions by applying Shephard's Lemma to the managerial expenditure function, which is dual to the managerial utility maximization problem, in which managers trade off risk and return. These equations include revenue terms, the tax rate, and risk terms, which would not be included in the functions were the managers maximizing. Hence, the coefficients on these terms offer a test of profit maximization vs. utility maximization.[6] Other models would lead to other specifications.

In any of the estimation techniques, X-efficiency is essentially the residual. This means that omitted variables (or extraneous variables) can have large effects on measured efficiency. Specification of included variables is important, since the methodology depends on comparing the firm's cost/profit/other outcome to those of a best-practice firm operating at the same level of the exogenous variables included in the frontier. That is, the exogenous variables determine the reference set for the firm whose efficiency is being measured. If something extraneous is included in the frontier specification, then we may mislabel a firm as efficient because we would not be comparing the firm to the

[5] For banks with very low or very high $\hat{u}$, an adjustment (called truncation) is made to assign less extreme values of $\hat{u}$ to these banks, since extreme values may indicate that random error, v_i, has not been completely purged by averaging.

entire set of relevant firms. For example, if two firms differ only in that one's CEO is blond and one is a brunette ─ which I'm assuming is unrelated to efficiency! ─ then we would want to compare their costs to one another. If we included CEO hair color in the cost function as a dummy variable, we would preclude such a comparison. We might want to include in the specification of the frontier variables that account for differences in the environment in which the firm operates that are exogenous to the firm's decision-making but that may affect performance, e.g., we might want to include variables that account for demand, like income growth in the firm's market, or whether the firm is located in an urban or rural market. Then in measuring efficiency, the urban firms would be compared to other urban firms and the rural firms to rural firms. But note that the manager's potentially inefficient choice of where to set up shop ─ in a rural or an urban market ─ would not be penalized. The alternative is to leave the variable out of the frontier specification, but then determine whether the efficiency estimates are correlated with the variable. One has to use one's judgment about what is the better way to proceed.

3.3 Special Issues in Banking. Judgment also has to be used when applying efficiency techniques to certain industries. The special issues that arise in applying the techniques to the banking industry are suggestive of some of the problems/issues that can arise in efficiency estimation. In banking, a big issue has been how to measure outputs and inputs. There has been some disagreement in the literature over what a commercial bank is actually producing. Two general approaches have been taken: the "production" approach and the "intermediation" approach (also called the "asset" approach). The production approach focuses on the bank's operating costs, i.e., the costs of labor (employees) and physical capital (plant and equipment). The bank's outputs are measured by the number of each type of account, like commercial and industrial loans, mortgages, deposits, etc. because it is thought that most of the operating costs are incurred by processing account documents and debiting and crediting accounts; inputs are labor and physical capital. The "intermediation" approach considers a financial firm's production process to be one of financial intermediation, i.e., the borrowing of funds and the subsequent lending of

[6] Hughes, Lang, Mester, and Moon (2000) reject the hypothesis of profit maximization using 1989-1990 data on U.S. banks that reported at least $1 billion in assets as of the last quarter of 1998.

those funds. Thus, the focus is on total costs, including both interest and operating expenses. Outputs are measured by the dollar volume of each of the bank's different types of loans, and inputs are labor, physical capital, deposits and other borrowed funds, and in some studies, financial capital.[7] The studies on X-efficiency in banking have tended to use the intermediation approach.

Theoretically, to compare one firm's efficiency to another's, we would like to compare each firm's cost of producing the same outputs. For banks, significant characteristics of loans are their quality, which reflects the amount of monitoring the bank does to keep the loan performing, and their riskiness. Unless these characteristics are controlled for, one might conclude a bank was producing in a very efficient manner if it were spending far less to produce a given output level, but its output might be highly risky and of a lower quality than that of another bank. It would be wrong to say a bank was efficient if it were scrimping on the credit evaluation needed to produce sound loans. Thus, recent studies have included quality and nonperforming loans in the specifications of cost and profit functions. Hughes, Lang, Mester, and Moon (2000) derive the risk-return tradeoff explicitly from a utility maximization model rather than just augmenting the cost and profit functions with risk and quality measures. See Hughes (1999) for further discussion.

Unfortunately, there are likely to be unmeasured differences in quality because the banking data do not fully capture the heterogeneity in bank output. The amount of service flow associated with financial products is by necessity usually assumed to be proportionate to the dollar value of the stock of assets or liabilities on the balance sheet, which can result in significant mismeasurement. For example, commercial loans can vary in size, repayment schedule, risk, transparency of information, type of

[7] A slight variation on the intermediation approach, which has been used in some studies, is to distinguish between transactions deposits, which are treated as an output, since they can serve as a measure of the amount of transactions services the bank produces, and purchased or borrowed funds (like federal funds or large CDs purchased from another bank), which are treated as inputs, since the bank does not produce services in obtaining these funds. The strict intermediation approach would consider the transactions services produced by the bank as an intermediate output, something that must be produced along the way toward the bank's final output of earning assets. Hughes and Mester (1993) empirically tested whether deposits should be treated as an input or output and found support that they should be treated as an input in their study.

Another approach that has been taken less often is the "value-added" approach, which considers all liabilities and assets of the bank to have at least some of the characteristics of an output. Still another approach, taken in Mester

collateral, covenants to be enforced, etc. These differences are likely to affect the costs to the bank of loan origination, ongoing monitoring and control, and financing expense. Unmeasured differences in product quality may be incorrectly measured as differences in cost inefficiency.

Another issue raised in recent papers in the bank efficiency literature is the treatment of financial capital. A bank's insolvency risk depends on its financial capital available to absorb portfolio losses, as well as on the portfolio risks themselves. Insolvency risk affects bank costs and profits via the risk premium the bank has to pay for uninsured debt and through the intensity of risk management activities the bank undertakes. For this reason, the financial capital of the bank should be considered when studying efficiency. To some extent, controlling for the interest rates paid on uninsured debt helps account for differences in risk, but these rates are imperfectly measured.

Even apart from risk, a bank's capital level directly affects costs by providing an alternative to deposits as a funding source for loans. Interest paid on debt counts as a cost, but dividends paid do not. On the other hand, raising equity typically involves higher costs than raising deposits. If the first effect dominates, measured costs will be higher for banks using a higher proportion of debt financing; if the second effect dominates, measured costs will be lower for these banks. Large banks depend more on debt financing to finance their portfolios than small banks do, so a failure to control for equity could yield a scale bias.

Studies that have considered financial capital include the level of capital rather than its price. Including the price assumes that banks on the frontier are selecting the cost-minimizing level of capital. This might not be the case because of regulations that set a minimum capital-to-asset ratio or because of risk-aversion on the part of bank managers. See Hughes and Mester (1993) for further discussion.

3.4 Tests of Expense-Preference Behavior. The sections above give an overview of efficiency measurements. Expense-preference is one particular form of X-inefficiency, in which firm managers are assumed to derive utility from choosing a greater than efficient (i.e., cost-minimizing or profit-

---

(1992), is to consider the bank's output to be its loan origination and loan monitoring services, since these outputs are more closely related to the theory of financial intermediation.

maximizing) level of one or more of the firm's inputs, usually labor. That is, the managerial utility function is $U=U(\pi, E)$, where E represents expenditures on the input.

Tests for expense preference are based on estimating input demand functions or cost functions. The functional forms are derived explicitly from the utility function, which depend on the underlying production function of the firm. Edwards (1977) derived the demand for labor equation for a firm using a Cobb-Douglas production function and exhibiting expense preference for labor. Mester (1989b) generalizes expense-preference tests to allow for less restrictive production structures and the presence of expense-preference toward any input, not just labor. Note that the derived tests in both of these studies cannot give firm-specific measures of inefficiency. Rather they are tests of whether a group of firms is showing expense-preference toward any input. Issues of functional form choice, variable choice, and variable measurement discussed above are as relevant in these tests as they are in measuring more general X-inefficiency.

4. Application of Efficiency Measurement Techniques to Central Banks

The review above discusses some of the steps that need to be followed in implementing efficiency measurement. The main steps involve choosing the:

(1) Efficiency concept, i.e., firm objective function (this includes specification of the production function of the firm),

(2) Estimation technique,

(3) Functional form,

(4) Variables and their proxies.

Can this process be applied to the central bank? There are some (narrow) aspects of central banking to which efficiency measurement can be applied. Other aspects of central banking would be more difficult to study with efficiency measurement techniques. If we are interested in the narrow case of whether the central bank is creating its “output” in the most efficient manner in the sense of resource costs, then there are certain outputs where we can apply the standard efficiency measurement techniques. The key is being able to define that output and to specify the central bank's objective function with

respect to that output. Once that is done, the rest of the process is fairly straightforward, although one is likely to confront measurement and data issues (which could be serious). For other bank activities, this would be difficult because measuring the central bank's output is difficult. Moreover, if we are interested in social efficiency, i.e., is the central bank operating to minimize the costs borne by society for a given level of "output," that is a more difficult question than the question of resource costs.

In the U.S., central bank activities include monetary policy, bank supervision and regulation, and payment services, to ensure financial system stability and maximum sustainable economic growth. Of these "products," efficiency measurement techniques can most easily be applied to the question of resource cost efficiency of providing payment services, given that the central bank is a provider of the services.[8]

4.1 Application to Payment Services. The Fed's payment services include check clearing and collection, wire transfers, automated clearinghouse transfers, securities safekeeping, and coin and currency distribution. To the extent that the Fed is mandated to provide some of these services on a competitive basis, it would seem reasonable to assume that the Fed would want to minimize its costs of producing these services, subject to the constraints under which is must operate (e.g, given the areas it is required to service, particular inputs it is required to use, quality of output it is required to provide) and given the choice of production technology (one that does not subject the payment system to excessive risk). An efficient Fed would likely want to minimize its costs of production even for the payments services over which it has a monopoly, e.g., currency distribution, although in this case there would be no competitive forces driving the Fed toward cost minimization. Thus, in the case of payment services, the goal could be specified ― minimize the cost of producing the payment service ― and the output of the payment service could be measured ― e.g., number of checks processed, number of pieces of unfit currency destroyed. So the efficiency measurement techniques could be applied. (There are many issues that would come up in implementing the techniques, e.g., whether the central bank wants to minimize

[8] Of course, there is an issue about whether the central bank should be a provider of these services.

costs each year, or over a longer period, but these types of problems come up in any efficiency study and are not unique to applications to central banking.)

Several papers have estimated cost functions for central bank payment services and estimated scale economies based on the estimated cost function. Fewer papers have applied efficiency measurement techniques to central bank payment services (see, e.g., Bauer and Ferrier, 1996; Bauer and Hancock, 1993; Bohn, Hancock, and Bauer, 2001; and Gilbert, Wheelock, and Wilson, 2002). But these papers show it can be done, under the assumption that the Fed wants to minimize costs of production. As an example consider Bohn, Hancock, and Bauer (2001). They estimate cost functions for Fed currency operations, where the outputs are number of fit notes generated by high-speed currency processing operations; number of notes destroyed either on-line by the high-speed machines or off-line at the reconciliation stations; and total number of transactions with depository institutions (which is number of incoming shipments of currency received + number of outgoing orders for currency filled). Inputs specified are buildings, labor, equipment, and materials, with equipment or equipment and buildings entered as levels (netputs) in some specifications. The translog and a hybrid of the translog cost function are estimated and the stochastic frontier and distribution-free methods are used to derive cost efficiency measures for the 37 Federal Reserve Banks and Branches that do currency operations. They find that the average office operates at more than 80 percent of the efficiency of the best-practice office. Although the cost-efficiency estimates for individual offices vary substantially across the estimated models, there is a consistent grouping for the most efficient and least efficient offices across models. Thus, the study shows that one can apply the standard techniques and get seemingly reasonable estimates of efficiency for a central bank business.

But the discussion above assumed that the choice of technology used in the production of payment services was given. That is, once the technology is chosen, the central bank would want to provide services in the most efficient manner. In choosing the technology with which it provides the service, however, the central bank may have other goals in addition to efficient production. To the extent that the central bank wants to ensure stability of the payment system, it may choose or design a

technology or type of payment service that is not least-cost if that means lower risk with regard to stability. This is similar to the tradeoff facing bank managers with regards to their choice of loan quality that Hughes, Lang, Mester, and Moon (1996 and 2000) study. Berger, Hancock, and Marquardt (1996) discuss this tradeoff and the risk in different forms of payment. The main risk is settlement risk, which includes both credit risk and liquidity risk, and which can develop into systemic problems if the problems of one participant spill over onto others.

There is also the issue of whether the central bank should be providing payment services or whether it should leave the provision to the private sector. Gilbert (1998) studies whether the creation of the Fed and its provision of check-collection services improved the efficiency of the U.S. payments system. He argues that the Fed's service lowered transactions cost of making payments and that banks found the Fed's system for collecting checks across regions more attractive than the old system. Standard efficiency techniques used to investigate resource costs in payments would not capture the social efficiency of technologies and products.

4.2 Application to Banking Supervision and Financial Stability. It is more problematic to apply the efficiency techniques to either banking supervision or monetary policy. The reason is that it is difficult to specify the relevant objective function with respect to these activities. Unless one can clearly specify the objective function, then it is difficult to proceed. I do not know of any papers that have applied the standard methods of efficiency measurement to banking supervision, which is located in some central banks as well as the bank supervisory agencies. The goal of financial market stability is a worthy one, but how do we measure it? One can conceptualize the question of whether the central bank efficiently deploys its bank examiners in a way to achieve the highest level of banking industry soundness. But taking that question to data is difficult because it is difficult to quantify the output. It would have to be more than just counting the number of banks examined and relating that to the cost of the bank supervisory staff. One could relate the number of bank failures over a certain period with the number/cost of supervisory staff involved in examining the banks during the prior period and estimate a frontier across different regulators or over time (holding constant other factors related to failure). But

does a high failure rate mean poor supervision? Is a bank failure a failure of the supervisory process, since the examination process didn't identify the problem in the bank early enough to get management to fix it, or is a bank failure (that doesn't spread) a success because it weeds out bad management?[9]

4.3 Application to Monetary Policy. Similar issues surround the use of the efficiency estimation techniques with regard to monetary policy. Even a narrow application just looking at resource efficiency still requires being able to write down the objective function of the monetary policymaker, measuring the "outputs" and "inputs" of policy. And resource allocation does not seem to be the interesting question regarding monetary policy efficiency. Cosier and Longworth (2003) present a good overview of the Bank of Canada's approach to assessing the efficiency of monetary policy.

Section 2A of the Federal Reserve Act lays out the Fed's monetary policy objectives: "The Board of Governors of the Federal Reserve System and the Federal Open Market Committee shall maintain long run growth of the monetary and credit aggregates commensurate with the economy's long run potential to increase production, so as to promote effectively the goals of maximum employment, stable prices, and moderate long-term interest rates." Thus, the Fed is to conduct monetary policy to promote price stability, sustainable growth, and financial stability. Translating this into a well-specified objective function to which efficiency measurement techniques can be applied is difficult, partly because the parameters of any loss function are not known and may differ across policymakers.

Here I'll describe two different applications of efficiency techniques to monetary policy activites.

4.3.1 Expense-Preference Behavior of Central Banks. Only a handful of papers have looked at the X-efficiency of the monetary policy function at central banks. These papers have investigated expense-preference behavior at the Fed and include Toma (1982), Shughart and Tollison (1983a), Boyes, Mounts, and Sowell (1988), and Mester (1994). Boyes, Mounts, and Sowell (1988) define the Fed's

[9] Note that one can compare the efficiency (cost or profit) of banks examined by the Fed, OCC, and FDIC, holding other traits of the bank constant, as perhaps an indirect measure of supervisory effectiveness. The supervisor, like the market, exerts some control over the bank. But this would not tell one whether the supervisor was efficiently exerting that control. Berger and Mester (1997) found only weak relationships between regulator identity and cost and profit efficiency.

output as the monetary base (reserves + currency outside the banking system), assume the Fed has a Cobb-Douglas production function, and investigate whether the Fed exhibits expense preference toward labor (at the Board and Reserve Banks) by estimating the derived input demand function assuming that the Fed's goal is to maximize profits with regard to money creation. They find evidence of expense preference. Mester (1994) expands the model and tests, showing that one of the maintained assumptions underlying Boyes, Mounts, and Sowell ― that the Fed uses a Cobb-Douglas technology ― is rejected by the data. More general tests of expense preference that allow a less restrictive production technology for the Fed cannot reject the hypothesis that the Fed did not indulge in expense preference toward labor.

In the Boyes, Mounts, and Sowell model the Fed uses a Cobb-Douglas production function to create the money supply, then $M_t = AL_{1t}^{b}L_{2t}^{c}K_{1t}^{d}K_{2t}^{e}$ where M_t = monetary base, L_{1t} is the number of employees at the Board of Governors, L_{2t} is the number of employees at the Federal Reserve Banks, K_{1t} is physical capital at the Board of Governors, and K_{2t} is physical capital at the Reserve Banks.[10] The demand for money function facing the Fed is $R_t = j_0 M_t^{j1} Y_{1t}^{j2} Y_{2t}^{j3}$, where R_t is the appropriate interest rate, Y_{1t} is real GDP, and Y_{2t} is the currency-to-deposit ratio. Y_{1t} and Y_{2t} are shift variables representing all nonprice factors affecting demand. To the extent that the Fed also produces check-processing services, in addition to the monetary base, and the volume of such processing is related to the level of real GDP and currency-to-deposits, these variables might also proxy other outputs.

Mester (1994) shows that if Fed managers act to <u>maximize profits,</u> $\pi = RM - W_1L_1 - W_2L_2 - r_1K_1 - r_2K_2$, (i.e., they do not exhibit expense preference), then the implied demand functions for labor are:

$$ln\ L_{1t} = a_0 + a_1\ ln\ W_{1t} + a_2\ ln\ W_{2t} + a_3\ ln\ r_{1t} + a_4\ ln\ r_{2t} + a_5\ Y_{1t} + a_6 Y_{2t} \quad (4)$$

$$ln\ L_{2t} = b_0 + b_1\ ln\ W_{1t} + b_2\ ln\ W_{2t} + b_3\ ln\ r_{1t} + b_4\ ln\ r_{2t} + b_5 Y_{1t} + b_6 Y_{2t}, \quad (5)$$

where W_{1t} = wage of labor at the Board of Governors, W_{2t} = wage of labor at the Federal Reserve Banks, r_{1t} = cost of capital at the Board of Governors, and r_{2t} = cost of capital at the Reserve Banks.

[10] Actually, Boyes, Mount, and Sowell ignore physical capital; the Mester extension considers it.

If, on the other hand, the Fed exhibits expense preference for labor by maximizing a utility function of the form $U=U(\pi_t,E_{1t},E_{2t})$, where π_t = System profits, E_{1t} = expenditures for Board labor, and E_{2t} = expenditures for Reserve Bank labor, then the labor demand functions are:

$$ln\ L_{1t} = \alpha_0 + \alpha_1\ ln\ W_{1t} + \alpha_1\ ln(1-(U_{E1}/U_\pi))_t + \alpha_2\ ln\ W_{2t} + \alpha_2\ ln(1-(U_{E2}/U_\pi))_t + \alpha_3\ ln\ r_{1t} + \alpha_4\ ln\ r_{2t} + \alpha_5 Y_{1t} + \alpha_6 Y_{2t} \quad (6)$$

$$ln\ L_{2t} = \beta_0 + \beta_1\ ln\ W_{1t} + \beta_1\ ln(1-(U_{E1}/U_\pi))_t + \beta_2\ ln\ W_{2t} + \beta_2\ ln(1-(U_{E2}/U_\pi))_t + \beta_3\ ln\ r_{1t} + \beta_4\ ln\ r_{2t} + \beta_5 Y_{1t} + \beta_6 Y_{2t}, \quad (7)$$

where U_{E1} = marginal utility of expenditures on Board staff, U_{E2} = marginal utility of expenditures on Reserve Banks' staff, and U_π = marginal utility of profits.[11]

Proxying the expense preference term $ln(1-(U_{E1}/U_\pi))_t$ by *ln* [(Board assessments/number of Board employees)/System profits] ≡ $ln\ p_{1t} - ln\ L_{1t}$, and the expense preference term $ln(1-(U_{E2}/U_\pi))_t$ by *ln* [(Federal Reserve Bank expenses/number of Reserve Bank employees)/System profits] ≡ $ln\ p_{2t} - ln\ L_{2t}$, the reduced form of the labor demand functions (6) and (7) are:

$$ln\ L_{1t} = A_0 + A_1\ ln\ W_{1t} + A_1\ ln\ p_{1t} + A_2\ ln\ W_{2t} + A_2\ ln\ p_{2t} + A_3\ ln\ r_{1t} + A_4\ ln\ r_{2t} + A_5 Y_{1t} + A_6 Y_{2t} \quad (8)$$

$$ln\ L_{2t} = B_0 + B_1\ ln\ W_{1t} + B_1\ ln\ p_{1t} + B_2\ ln\ W_{2t} + B_2\ ln\ p_{2t} + B_3\ ln\ r_{1t} + B_4\ ln\ r_{2t} + B_5 Y_{1t} + B_6 Y_{2t}, \quad (9)$$

with the appropriate coefficient definitions and where p_{1t} = Board assessments/System profits, and $ln\ p_{2t} - ln\ L_{2t}$, where p_{2t} = Reserve Bank expenses/System profits.

A joint test of Cobb-Douglas production technology and expense-preference behavior involves estimating the system without coefficient restrictions and then testing whether the coefficient on $ln\ W_{1t}$ = coefficient on $ln\ p_{1t,}$ the coefficient on $ln\ W_{2t}$ = coefficient on $ln\ p_{2t}$ in each input demand function, the coefficient on $ln\ W_{1t} < 0$ in the Board demand function, and the coefficient on $ln\ W_{2t} < 0$ in the Reserve

[11] Boyes, Mount, and Sowell's labor demand functions differ from these for unexplained reasons. The demand for Board labor excludes the term $\alpha_2\ ln\ W_{2t}$ and the restriction of equal coefficients on $ln\ W_{1t}$ and $ln\ (1-(U_{E1}/U_\pi))$ was not imposed. And similarly for the Reserve Bank labor demand equation.

Bank demand function. Mester rejects the joint hypothesis of Cobb-Douglas technology and expense preference based on these tests.

Mester (1994) then generalizes the tests to allow for a more general production technology than Cobb-Douglas and to allow for the possibility that the Fed shows expense preference toward physical capital instead of or together with labor. Assuming that the Fed's underlying cost function ― i.e., the cost function it would have if it did not exhibit expense preference ― is of the translog form, she derives the log cost function and log input expenditure functions were the Fed to exhibit expense preference. This is a nonlinear model and yields coefficient tests for expense preference. Again, the tests reject expense preference on the part of the Fed.

These expense-preference tests show that standard efficiency techniques can be applied to monetary policy activities. However, the tests are based upon a model where an efficient Fed maximizes profits. It is not clear that profit maximization is the correct metric for the central bank's money supply activities.

4.3.2 Other Objective Functions. There is a long literature that looks at monetary policy reaction functions, or Taylor-type rules for monetary policy (see Taylor, 1999, for a survey and Hetzel, 2000, for a critique of the Taylor-rule literature). Such a rule relates the policy instrument to targets for inflation and output gap or the unemployment rate, i.e., it relates the instrument to macroeconomic variables. It also assumes that the economic dynamics imply a tradeoff between inflation and the output gap or unemployment (i.e., it is based on an underlying Philips curve). For example,

$$f_t = r^* + \pi_t + \theta_\pi (\pi_t - \pi^*) + \theta_y y_t, \qquad (10)$$

where f_t is the nominal interest rate, π_t is inflation, y_t is the output gap (the percentage deviation of output from potential output), π^* is the policymaker's inflation target, and r^* is the long-run equilibrium or "natural" real rate of interest. (Sometimes the rule is written so that the output gap is replaced by the difference between the unemployment rate and the natural rate of unemployment, $u_t - u^*$.) Taylor's original specification had $\pi^* = 2$, $r^* = 2$, $\theta_\pi = 1/2$, and $\theta_y = 1/2$. According to Orphanides (1998) and Taylor (1999), Taylor's rule appears to perform well in a variety of models and appears to be robust to

model specification. However, the original Taylor rule specification is not necessarily "efficient" in the sense of stabilizing output and inflation at their targets. That is, there are alternative values of the coefficients that yield better performance. In addition, the rules are hard to implement as written, since the policymaker does not have accurate information on the current values of inflation or the output gap when setting the interest rate. Both inflation and the output gap are estimated with considerable noise.

Such a rule can be derived from a model of the economy in which the central bank's goal is to stabilize output and inflation. Let inflation be determined by

$$\pi_t = \pi_{t-1} + \alpha y_t + e_t, \tag{11}$$

where $\alpha > 0$ and e_t is the disturbance to inflation, which captures supply shocks.

Let output be determined by

$$y_t = \rho y_{t-1} - \xi(r_{t-1} - r^*) + u_t, \tag{12}$$

where $\xi > 0$, $0<\rho<1$, and u_t is the disturbance to output, which captures demand shocks.

Let the central bank's objective function be to minimize a weighted sum of the unconditional variances of inflation and the output gap,

$$\text{minimize } L = \omega \text{Var}(\pi_t - \pi_t^*) + (1-\omega)\text{Var}(y_t),\ 0<\omega<1. \tag{13}$$

Orphanides (1998) shows that given the processes for output and inflation, in the absence of noise, the optimal policy relative to this objective function is

$$f_t = r^* + \pi_t + \theta_\pi^N (\pi_t - \pi^*) + \theta_y^N y_t, \tag{14}$$

$$\text{where } \theta_\pi^N = \frac{-\alpha\omega + \sqrt{4(1-\omega)\omega + (\alpha\omega)^2}}{2(1-\omega)\xi} \text{ and } \theta_y^N = \frac{\rho}{\xi}. \tag{15}$$

Given this model, one can trace out the inflation-output gap variance frontier as a function of ω, the weight on inflation in policymaker's loss function. That is, for a given ω, if the policymaker follows the optimal policy, it produces a particular inflation variance and output gap variance. Each ω yields a single point on the frontier. The frontier gives the values of the variance of inflation and the variance of the output gap as ω varies between 0 and 1. See Figure 4. An efficient policy will be one that places us on

the frontier ─ which point on the frontier depends on the policymaker's loss function, i.e., his/her preferences for inflation and output stability.[12]

One can then examine where actual experience has been relative to this frontier. If we knew the weight, ω, in the policymaker's loss function, then we would want to compare the actual point to the point on the frontier corresponding to ω. Inefficiency could be measured as the distance to this point, but this would weight the deviation of actual inflation variability from optimal inflation variability and the deviation of actual output variability from optimal output variability equally. Instead, it might make sense to measure inefficiency using the weights in the loss function, i.e.,

$$\begin{aligned}\text{Inefficiency} = {} & \omega\,[\text{Actual inflation variability} - \text{Optimal inflation variability given } \omega] \\ & + (1-\omega)[\text{Actual output variability} - \text{Optimal output variability given } \omega]. \qquad (16)\end{aligned}$$

This essentially uses the policymaker's loss function to weight the deviations of actual output gap and inflation variance from optimal variances. However, in general, we will not know the policymaker's loss function parameter, so we will not know which point on the frontier is the proper reference point as the optimal point. There are several different measures of distance from the frontier that could be used to measure inefficiency ─ horizontal distance, which would measure the increase in inflation variance for a given level of output gap variance; vertical distance, which would measure the increase in output variance for a given level of inflation variance; the minimum distance to the frontier (which is the distance to the frontier along a ray orthogonal to the frontier); and the distance to the frontier along a ray through the origin.

Cecchetti, Flores-Lagunes, and Krause (2001), and Cecchetti and Krause (2002) studied central bank efficiency across different countries using this type of efficiency frontier. They estimated a two-equation model of the economy, similar to the model described above, for 24 countries using data from 1991Q1 to 1998Q4, and trace out an efficient frontier for each country. They assume that the inflation target is 2 percent for all countries. Their output measure is industrial production and they use trend growth in industrial production as a measure of potential growth. Then, for a given ω, they measure the

[12] There is a whole other question about whether the policymaker's preferences reflect those of society, but we'll set

loss associated with actual performance of the country using the loss function in equation (13). Part of this loss is due to inefficiency, the fact that the actual performance point is interior to the frontier. They measure the inefficiency of actual performance relative to optimal performance using equation (16). To determine each country's ω they use the ω that corresponds to the point at the intersection of the frontier and a ray through the origin and the point of actual performance. The optimal values of inflation variability and output variability are those on the frontier at this intersection. The authors find that there is high variation in both performance and policy efficiency across countries. They also find that central bank credibility ― the difference between the policymaker's plans regarding inflation and the public's belief about those plans ― is the main factor from among central bank independence, transparency, accountability, and credibility that explains most of the cross-country variation in macroeconomic outcomes. Independence, transparency, and accountability explain little of the variation. The authors are able to look at the shift in each country's efficient frontier over time by estimating the models over two different time periods. Thus, they have estimates of both "technological change" and changes in efficiency (i.e., changes in the dispersion from the frontier) over time.[13]

The work of Cecchetti, Flores-Lagunes, and Krause (2001) and Cecchetti and Krause (2002) is an example of an application of the ideas from the efficiency literature to monetary policy. But the actual implementation is tricky because it depends on a number of embedded assumptions. The efficiency measures are based on a particular value of ω, but the policymaker's ω is not known. The authors choose a value based upon the location of actual performance.[14] The policymaker's inflation and potential growth targets, which are important components of the objective function, are not necessarily explicitly stated by the central bank. One has to believe the economy's dynamics imply a tradeoff between inflation

---

that issue aside.

[13] Berger and Mester (2002) develop a framework for decomposing changes in bank performance into three components: change in the best-practice technology, change in efficiency or dispersion from best practice, and change in business conditions or economic factors exogenous to the firm. The first two components ― change in best practice and changes in inefficiency ― together form the more traditional notion of change in productivity.

[14] Cecchetti and Ehrmann (2002) provide a methodology for estimating ω for a country based on the slope of the country's aggregate supply curve and the variances of output and inflation, and assuming policymakers are acting efficiently, i.e., they are on the frontier. They find that for most of the 23 countries they study, ω, is quite large over

and the output gap. And nontrivial choices have to be made about how to measure inflation and the output gap (which inflation measure, which indicator of output, time period, etc.) Moreover, if policy is efficient given the policymaker's ω, it could still be inefficient from society's viewpoint. (E.g., putting no weight on inflation variability would likely lead to very poor economic outcomes. This is saying that the loss function of policymakers and the loss function of society may differ.) In addition, even if ω were known, there are significant measurement problems with implementing these efficiency estimates. Orphanides (1998) shows how the frontier shifts if the policymaker takes into account measurement error in the data, which, of course, will change the measures of inefficiency.

A key insight from the efficiency measurement literature is that you would also want to explicitly recognize that at the time of evaluation, the data on inflation and output are measured with error. That is, there are differences in the real-time data that are available when the policymaker has to make his/her decisions and the ultimate values of these variables. Thus, one would want to account for this kind of measurement error when evaluating efficiency frontier. Using data on revisions to inflation and output gap measures, one can derive confidence intervals around the estimates of actual inflation and output gap variability. And then measure inefficiency based on the distance from the edge of the confidence boundary to the confidence interval around the frontier ─ see Figure 5.

5. Conclusions

This paper has reviewed the standard techniques of efficiency measurement, discussed some of the issues that arise in applying these standard techniques to central banks, and reviewed some of the literature that has attempted to apply these techniques to central banking. The uniqueness of some of the activities of central banking, the difficulty in measuring some of the central banking outputs, and the complicated and multiple objectives pursued by central banks makes application of the standard techniques problematic. Certain central bank activities do lend themselves to efficiency measurement, e.g., payment services provision. It is much more difficult to apply the techniques to bank supervision

---

the 1984-97 period, with average value across countries of 0.73 (assuming an inflation target of 2 percent). Once one relaxes the assumption that policymakers are acting efficiently, then ω cannot be estimated from the data.

and monetary policy; however, insights from the efficiency literature can help improve the work that has been done to date on measuring the efficiency of central banks.

Figure 1

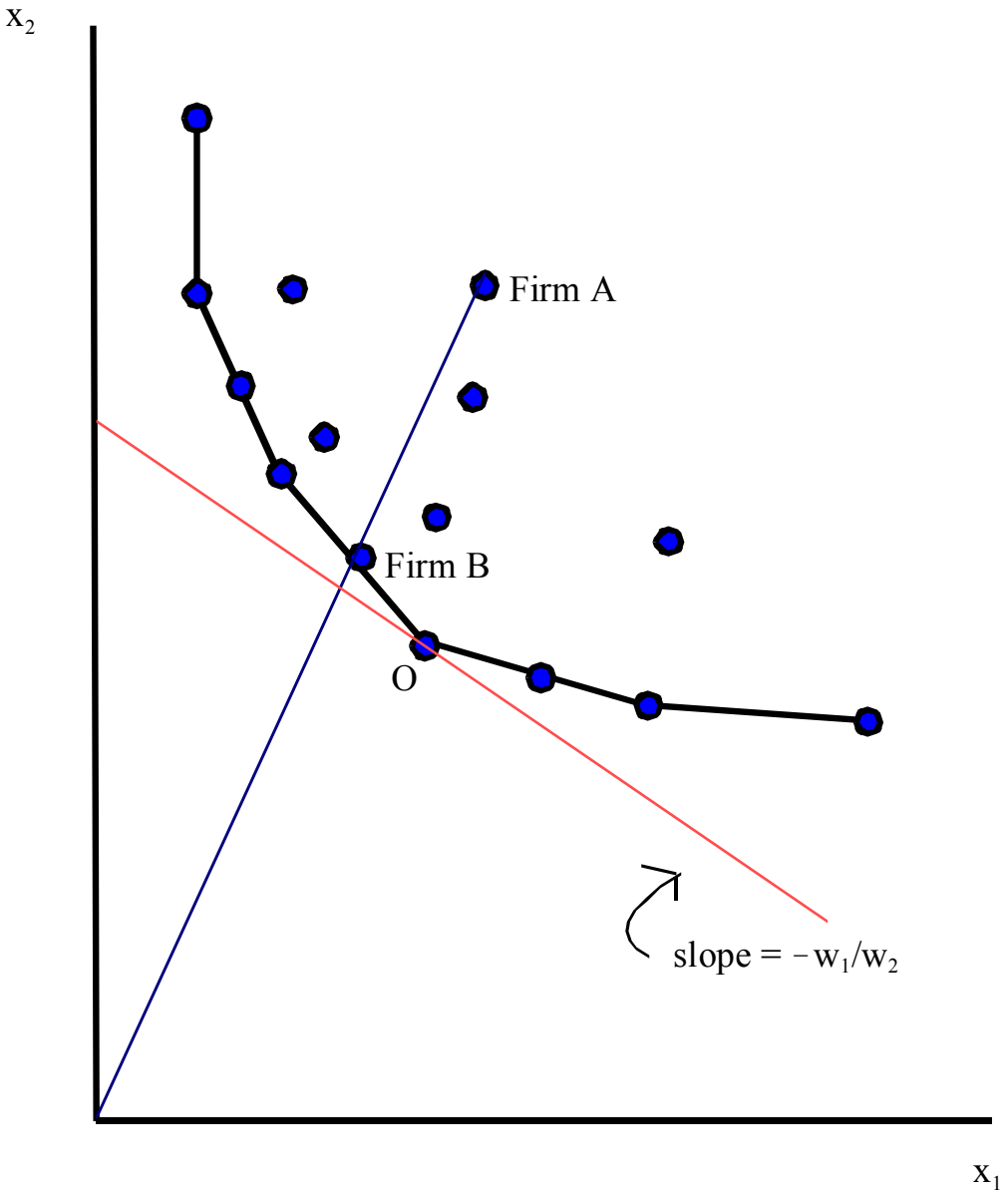


Figure 2

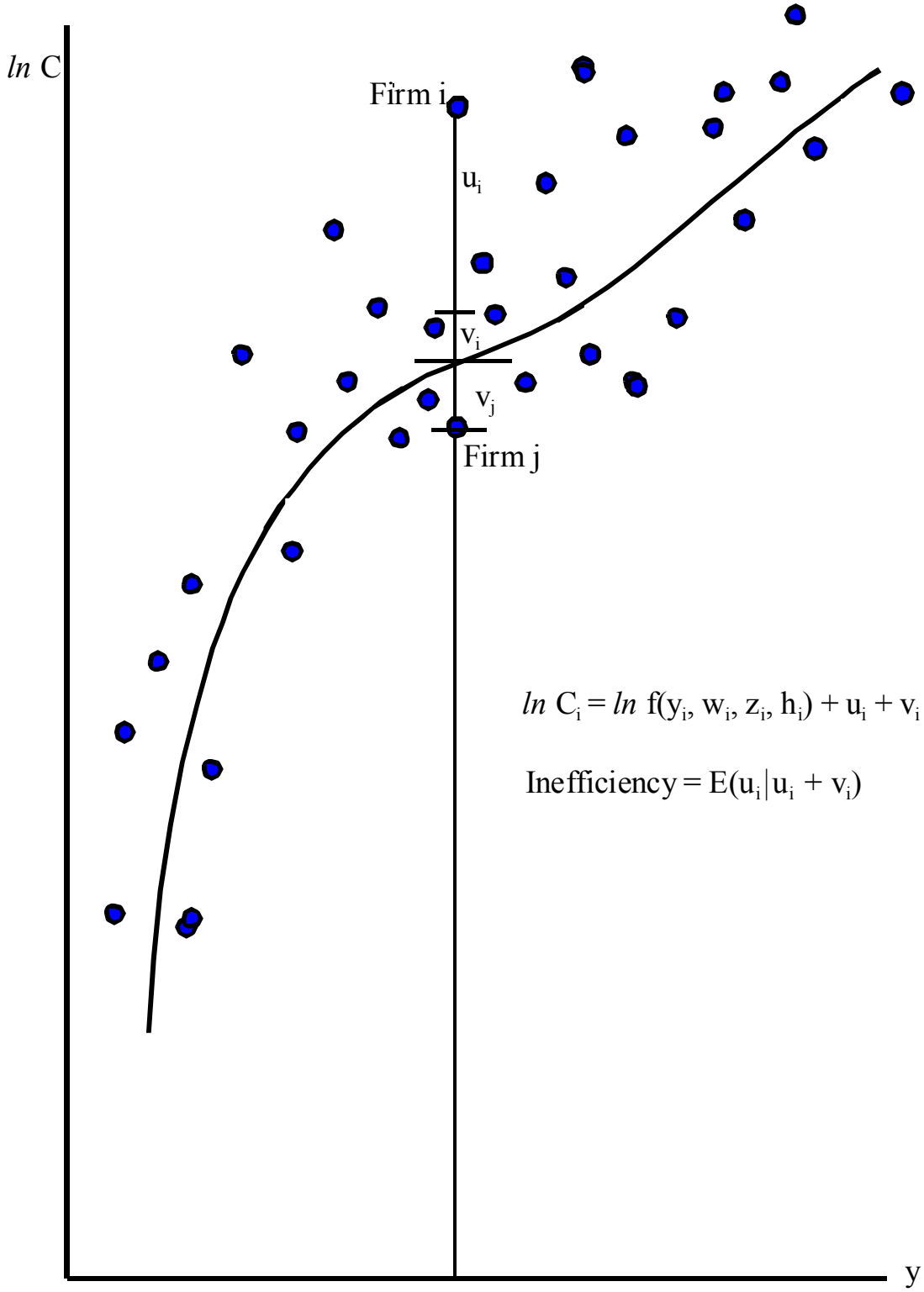

ln C
Firm i
u_i
v_i
v_j
Firm j
ln C_i = ln f(y_i, w_i, z_i, h_i) + u_i + v_i
Inefficiency = E(u_i | u_i + v_i)
y

Figure 3

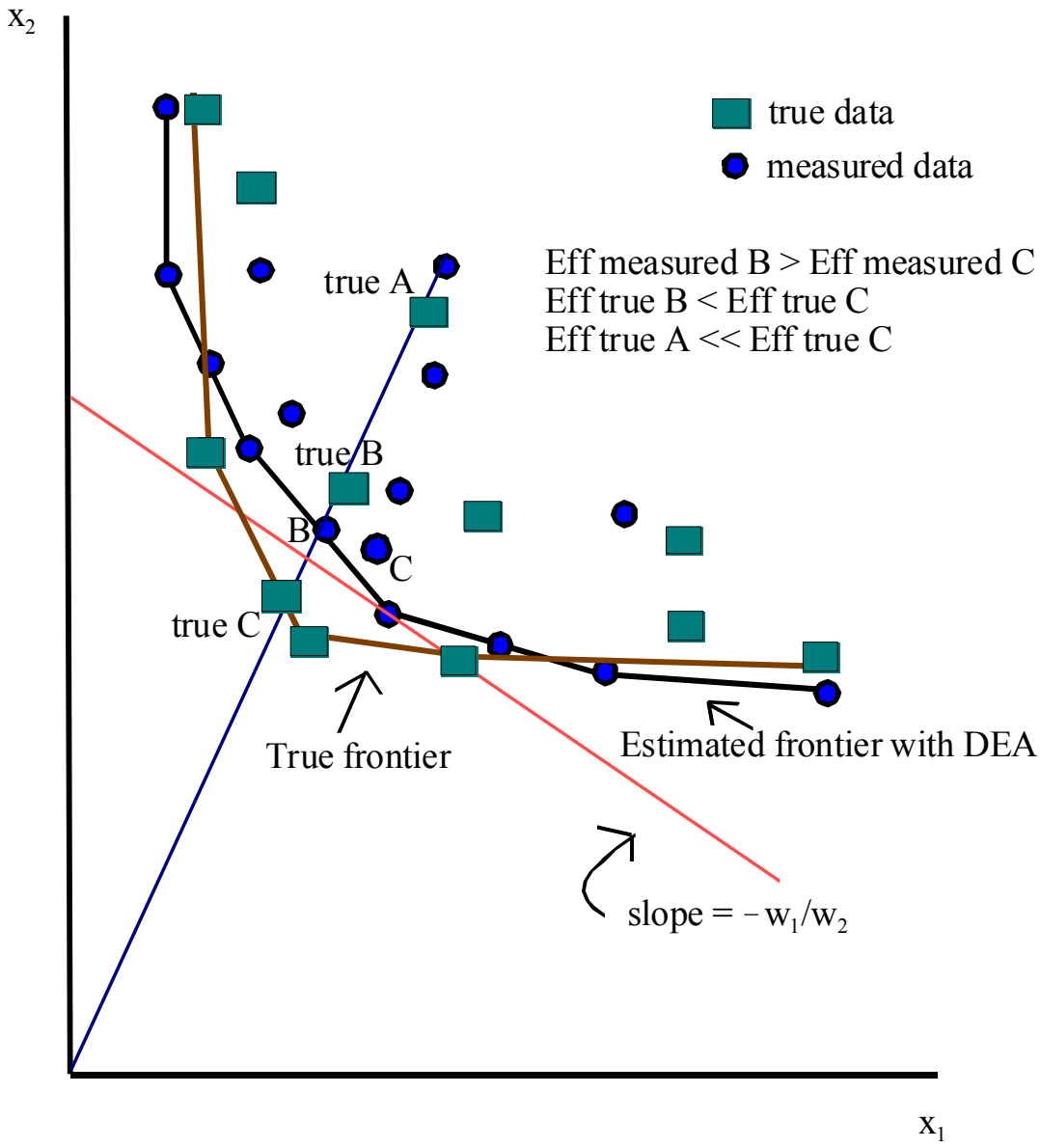
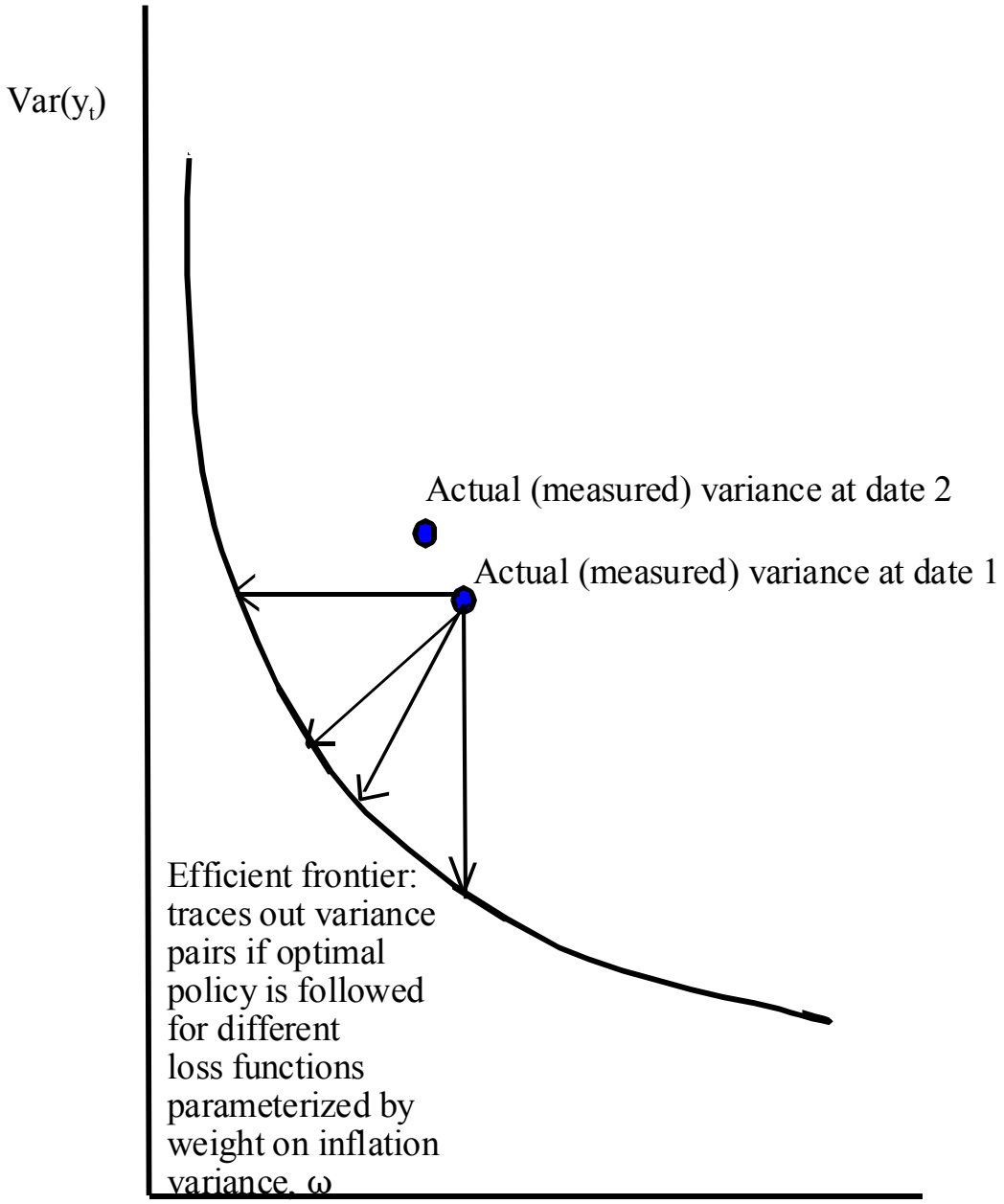
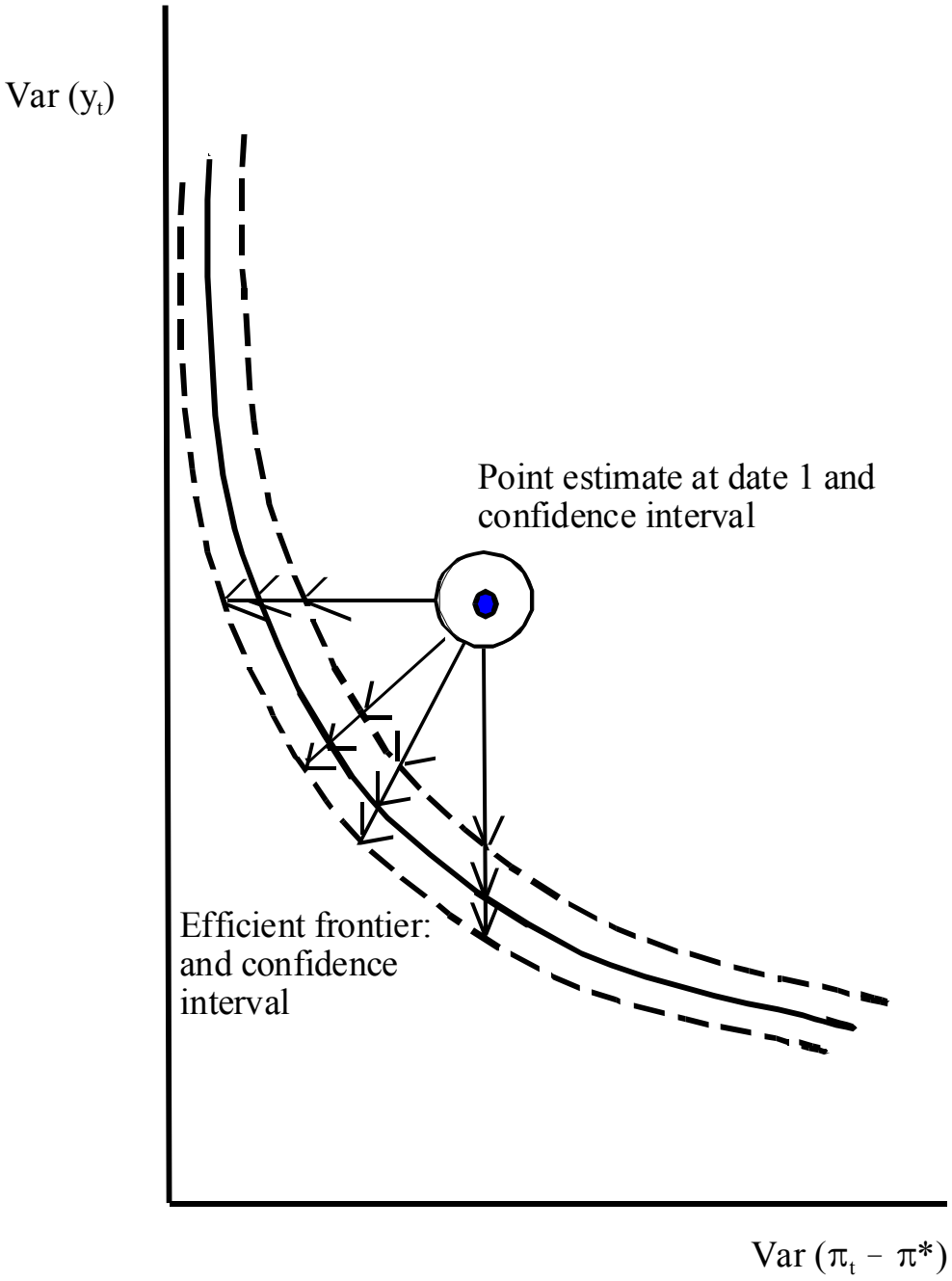

x2
true data
measured data
Eff measured B > Eff measured C
Eff true B < Eff true C
Eff true A << Eff true C
true A
true B
B
C
true C
True frontier
Estimated frontier with DEA
slope = – w1/w2
x1

Figure 4



$$\text{Loss} = \omega \text{ Var } (\pi_t - \pi^*) + (1-\omega) \text{ Var } (y_t)$$

Figure 5



$$\text{Loss} = \omega \text{ Var } (\pi_t - \pi^*) + (1-\omega) \text{ Var } (y_t)$$

References

Bauer, P.W., and G.D. Ferrier, “Scale Economies, Cost Efficiencies, and Technological Change in Federal Reserve Payments Processing," *Journal of Money, Credit, and Banking*, part 2, 28 (November 1996), pp. 1004-1039.

Bauer, P.W., and D. Hancock, “The Efficiency of the Federal Reserve in Providing Check Processing Services," *Journal of Banking and Finance*, 17 (April 1993), pp. 287-311.

Berger, A.N., D. Hancock, J.C. Marquardt, “A Framework for Analyzing Efficiency, Risk, Costs, and Innovations in the Payments System,” *Journal of Money, Credit, and Banking*, part 2, 28 (November 1996), pp. 696-732.

Berger, A.N., and D.B. Humphrey, “Efficiency of Financial Institutions: International Survey and Directions for Future Research,” European Journal of Operational Research (1997), pp. 175-212.

Berger, A.N., and L.J. Mester, “Inside the Black Box: What Explains Differences in the Efficiencies of Financial Institutions?” *Journal of Banking and Finance* 21 (July 1997), pp. 895-947.

Berger, A.N., and L.J. Mester, “Explaining the Dramatic Changes in Performance of U.S. Banks: Technological Change, Deregulation, and Dynamic Changes in Competition," *Journal of Financial Intermediation*, 12 (2003), pp. 57-95.

Bohn, J., D. Hancock, and P. Bauer, “Estimates of Scale and Cost Efficiency for Federal Reserve Currency Operations," *Economic Review*, Federal Reserve Bank of Cleveland, 37 (Quarter 4, 2001), pp. 2-26.

Boyd, J.H., “The Use of Inputs by the Federal Reserve System: Comment,” *American Economic Review*, 74 (December 1984), pp. 1114-1117.

Boyes, W.J., W.S. Mounts, and C. Sowell, “The Federal Reserve as a Bureaucracy: An Examination of Expense-Preference Behavior,” *Journal of Money, Credit, and Banking*, 20 (May 1988), pp. 181-190.

Cecchetti, S.G. and M. Ehrmann, "Does Inflation Targeting Increase Output Volatility? An International Comparison of Policymakers' Preferences and Outcomes," in N. Loayza and K. Schmidt-Hebbel,

eds., *Monetary Policy: Rules and Transmission Mechanisms*, No. 4 in the Series on Central Banking, Analysis, and Economic Policies, Santiago, Chile: Central Bank of Chile (2001), pp. 247-274.

Cecchetti, S.G., A. Flores-Lagunes, and S. Krause, "Has Monetary Policy Become More Efficient? A Cross-Country Analysis," manuscript, Ohio State University (May 2001).

Cecchetti, S.G., and S. Krause, "Central Bank Structure, Policy Efficiency, and Macroeconomic Performance: Exploring Empirical Relationships," *Review*, Federal Reserve Bank of St. Louis (July/August 2002), pp. 47-60.

Cosier, J. and D. Longworth, "Efficiency in Monetary Policy ─ Some Approaches at the Bank of Canada," manuscript, Bank of Canada (April 2003).

Edwards, F.R., "Managerial Objectives in Regulated Industries: Expense-Preference Behavior in Banking," *Journal of Political Economy*, 85 (February 1977), pp. 147-162.

Gilbert, R.A., "Did the Fed's Founding Improve the Efficiency of the U.S. Payments System?" *Review*, Federal Reserve Bank of St. Louis (May/June 1998), pp. 121-142.

Gilbert, R.A., D.C. Wheelock, and P.W. Wilson, "New Evidence on the Fed's Productivity in Providing Payments Services," Federal Reserve Bank of St. Louis Working Paper 2002-020A, September 2002.

Hetzel, R.L., "The Taylor Rule: Is it a Useful Guide to Understanding Monetary Policy?" *Economic Quarterly*, Federal Reserve Bank of Richmond, 86 (Spring 2000), pp. 1-33.

Hughes, J.P., "Incorporating Risk Into the Analysis of Production," Presidential Address to the Atlantic Economic Society, *Atlantic Economic Journal* 27, (1999), pp. 1-23.

Hughes, J.P., W. Lang, L.J. Mester, and C.-G. Moon, "Efficient Banking Under Interstate Branching," *Journal of Money, Credit, and Banking*, 28 (November 1996), pp. 1043-1071.

Hughes, J.P., W. Lang, L.J. Mester, and C.-G. Moon, "Recovering Risky Technologies Using the Almost Ideal Demand System: An Application to U.S. Banks," *Journal of Financial Services Research*, 18 (October 2000), pp. 5-27.

Hughes, J.P., W. Lang, L.J. Mester, C.-G. Moon, and M. Pagano, "Do Banks Sacrifice Value to Build Empires? Managerial Incentives, Industry Consolidation, and Financial Performance," *Journal of Banking and Finance*, 23 (2003), pp. 417-447.

Hughes, J.P., and L.J. Mester, "A Quality and Risk-Adjusted Cost Function for Banks: Evidence on the 'Too-Big-To-Fail' Doctrine," *Journal of Productivity Analysis*, 4 (September 1993), pp. 293-315.

Hughes, J.P., L.J. Mester, and C.-G. Moon, "Are Scale Economies in Banking Elusive or Illusive? Evidence Obtained by Incorporating Capital Structure and Risk-Taking into Models of Bank Production," *Journal of Banking and Finance*, 25 (December 2001), pp. 2169-2208.

Mester, L.J., "Owners versus Managers: Who Controls the Bank?" *Business Review*, Federal Reserve Bank of Philadelphia (May/June 1989a), pp. 13-23.

Mester, L.J., "Testing for Expense Preference Behavior: Mutual versus Stock Savings and Loans," *RAND Journal of Economics*, 20 (Winter 1989b), pp. 483-498.

Mester, L.J., "Agency Costs Among Savings and Loans," *Journal of Financial Intermediation*, 1 (June 1991), pp. 257-278.

Mester, L.J., "Traditional and Nontraditional Banking: An Information-Theoretic Approach," *Journal of Banking and Finance*, 16 (June 1992), pp. 545-566.

Mester, L.J., "Further Evidence Concerning Expense Preference and the Fed," *Journal of Money, Credit, and Banking*, 26 (February 1994), pp. 125-145.

Orphanides, A., "Monetary Policy Evaluation With Noisy Information," Finance and Economics Discussion Series Paper #1998-50, Federal Reserve Board (October 1998).

Shughart II, W.F., and R.D. Tollison, "Preliminary Evidence on the Use of Inputs by the Federal Reserve System," *American Economic Review*, 73 (June 1983a), pp. 291-304.

Shughart II, W.F., and R.D. Tollison, "The Use of Inputs by the Federal Reserve System: Reply," *American Economic Review*, 74 (December 1983b), pp. 1121-1123.

Strong, J.S., "The Use of Inputs by the Federal Reserve System: Comment," *American Economic Review*, 74 (December 1984), pp. 1118-1120.

Taylor, J.B., "Discretion versus Policy Rules in Practice," *Carnegie-Rochester Conference Series on Public Policy*, 39 (December 1993), pp. 195-214.

Taylor, J.B., "The Robustness and Efficiency of Monetary Policy Rules as Guidelines for Interest Rate Setting by the European Central Bank," *Journal of Monetary Economics*, 43 (June 1999), pp. 655-679.

Toma, M., "Inflationary Bias of the Federal Reserve System," *Journal of Monetary Economics*, 10 (September 1982), pp. 163-190.

Toma, M., "The Demise of the Public-Interest Model of the Federal Reserve System: A Review Essay," *Journal of Monetary Economics*, 27 (1991), pp. 157-163.