Guide to the Measurement of Government Productivity

Andrew Hughes*
New South Wales Treasury

Analyzing the performance of general government agencies benefits from the support of rigorous quantitative techniques. Drawing on the major conceptual and practical contributions of production economics, this article introduces the major methods for measuring productivity and efficiency in a non-technical way, and sets out the basic steps for their use. Tracking the productivity of a single service delivery unit over time or benchmarking the relative efficiency of a group of units at a point in time can provide a useful source of management information about performance.

This article is directed at both government policy analysts, and government officials accountable for the delivery of public services in a cost-effective manner. Its objective is to stimulate thinking about how more detailed and rigorous analysis of efficiency and productivity can assist in improving the delivery of public services to the community.

The underlying rationale for economic measures of performance stems from the major structural characteristics of most general government activity. In particular, the monopoly position that many agencies hold and the absence of any threat of takeover mean there are relatively weak organizational incentives to improve productivity and cost efficiency.

This situation contrasts with the typical position of private sector firms, where the discipline from product market competition and capital market scrutiny provide incentives for ongoing cost efficiency. Given the absence of market pressures for public sector entities, economic performance measures can be used to provide a set of surrogate incentives to spur performance improvement. The techniques canvassed in this article are summarized in Table 1.

The analysis of general government entities is generally focused on three broad areas:

- budget compliance — ensuring that agencies keep within annual spending limits;
- financial performance — assessing the financial health of an agency over time; and
- value for money — evaluating the efficiency and effectiveness of government service delivery.

Performance monitoring is principally concerned with evaluating effectiveness, efficiency and economy. Effectiveness is the extent to which an agency's programs and services (outputs) achieve the government's desired outcomes. Efficiency is defined as the extent to which an agency maximizes the outputs produced from a given set of inputs or minimizes the input cost of producing a given set of outputs.

Table 1
Summary of Techniques

Type of analysis	**Index number techniques** *Partial factor productivity (PFP) and total factor productivity (TFP) indexes*	**Statistical techniques** *Ordinary least squares (OLS) and stochastic frontier analysis (SFA)*	**Mathematical programming** *Data envelopment analysis (DEA)*
Measurement of *productivity change* through time ***Example*** Measuring TFP growth/decline for a single entity for a period of two or more years	***PFP index*** • Uses time series data — minimum of 4 data points. • Single inputs/outputs — does not require price data (as weights). ***TFP unilateral index (eg, Laspeyres)*** • Uses time series data. • Requires price (or cost/revenue) data to weight changes of multiple inputs/outputs. • Sources of TFP change cannot be identified. • Both index types assume no measurement error.	***OLS*** • Can be applied to measuring productivity change — not covered in Guide. ***SFA combined with Malmquist index*** • Typically uses panel data. • TFP change can be decomposed into changes in technical efficiency, scale and technology. • Allows for measurement error.	***DEA combined with Malmquist index*** • Typically, uses panel data. • TFP change can be decomposed into changes in technical efficiency, scale and technology. • Assumes no measurement error.
Measurement of relative *technical efficiency* levels at a point in time ***Example*** Benchmarking the technical efficiency of a group of service delivery units of an entity for a given year	***PFP index*** • Uses cross-sectional data. • Involves simple comparison of PFP ratios across entities. ***TFP multilateral index*** • Uses panel data. • Entities compared with a hypothetical, average entity in sample.	***OLS*** • Uses cross-sectional data. • Entities compared with average industry/sector performance. • Assumes no measurement error — residual is attributed to inefficiency. ***SFA*** • Uses cross-sectional data. • Comparison against best performing entity. • Residual decomposed into random error (meas. error) and inefficiency parts.	***DEA*** • Uses cross-sectional data. • Entities compared with best performers in sample.

Economy refers to buying inputs in the most economic manner (i.e., obtaining appropriate quality resources at least cost).

The content of this Guide draws upon the contribution of production economics to provide an introduction to the basics of productivity and efficiency measurement. Changes in the efficiency of a single entity can be considered over a period of time (time series analysis), and this is what some performance measurement techniques are designed to do. Alternatively, the efficiency of an entity can be compared at a point in time with other similar entities (cross-sectional analysis), for which other performance measurement techniques are required.

One of the major difficulties faced by any performance measurement exercise is defining robust measures of outputs and inputs. This problem is particularly acute for entities providing unpriced services outside the market mecha-

nism, which by definition is precisely the circumstance where such performance measures are most needed. And it is, of course, the situation for many government service providers.

Unfortunately there are no performance measurement techniques available that bypass this problem. In general, good performance measurement requires a clear definition of outputs and inputs, and data pertaining to those inputs and outputs.

However, some performance measurement techniques do provide scope for "testing" alternative input and output definitions in the following sense. Performance measures for different combinations of inputs and outputs can be derived, allowing the robustness of the measures to changes in definitions to be assessed. Ideally, the performance measures will not be very sensitive to changes in the choice of outputs and input measures.

Performance measures will be more useful where the sources of performance variation across organizations or operational units can be accounted for. In particular, when making judgements about performance and accountability for performance, it is important to distinguish the contribution of those factors that can be controlled by an organization's management from those that cannot. The latter are often characterized as an organization's particular **operating environment**, and include factors such as demographics, climate, and population density. Some of the techniques discussed below are capable of making such allowances in a consistent and robust way.

Public sector entities vary in the nature of their operations and the source of their funding. The international *Government Finance Statistics* reporting system defines the general government sector as consisting of those public sector entities that provide goods and services outside the market mechanism, and facilitate the transfer of income for public policy purposes.

General government service delivery is characterized by a number of factors that present challenges to the objective of achieving value for money. These include:

- complexities of the political process;
- tension between short term imperatives and the design of long term program and service delivery strategies;
- uncertainty about the most appropriate delivery strategies;
- unequal information among stakeholders pertaining to community needs, as well as program and service possibilities;
- different incentives for different stakeholders;
- absence of price signals to guide decisions on service provision and consumption; and
- lack of competition in service provision.

Conceptual Framework for Economic Performance Measurement

The concept of **productivity** is widely accepted as a key performance benchmark for entities. Rising productivity is related to increased profitability, lower costs and sustained competitiveness.

Productivity is defined as the ratio of outputs to inputs. Productivity can be analyzed at various levels — economy-wide, industry, firm/agency and operational unit. The focus of this Guide is at a more disaggregated level: how well the service delivery units of general government agencies convert inputs of labour, materials and capital into outputs of services.

For a simplified example we assume that labour is the only resource required to treat patients in a hospital.[1] Labour is measured as hours worked. The output is the number of treatments produced. If the hospital uses 500 hours of labour to produce 1,000 treatments, its productivity is 2 treatments per labour hour. If there is another hospital that uses 400 hours of

labour to produce 1,000 treatments its productivity is 2.5 treatments per labour hour. Since the second hospital can produce more treatments per labour hour than the first hospital, the second one is more productive.

We can use the concept of the **production function** to understand more about productivity. A production function describes the relationship between the output (number of treatments) and the input (number of labour hours). The nature of the output-input relationship depends on the particular **production technology** that is used to convert the inputs into the outputs. In this context, the term "technology" is intended to capture the skills of the labour input.

Chart 1
Simple Production Function (single output/input)

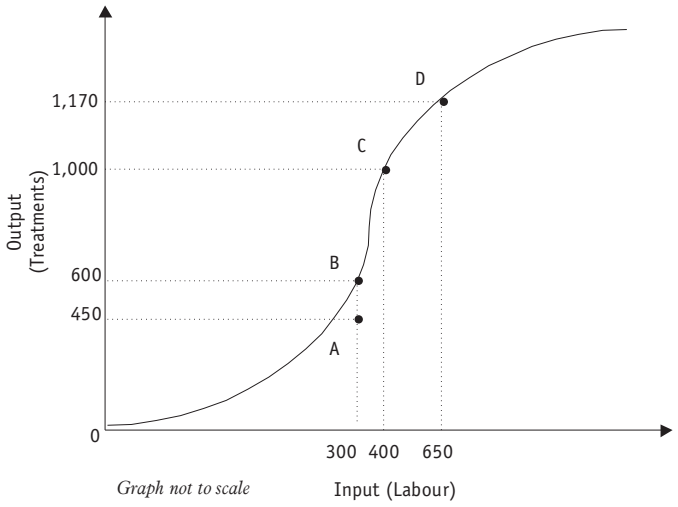


Chart 1 shows the production function for patient treatments. Points B, C and D represent the maximum number of treatments that can be produced at different input levels (or the minimum amount of labour hours required to produce a given number of treatments). The *shape* of the curve will depend on the particular production technology of the hospital.

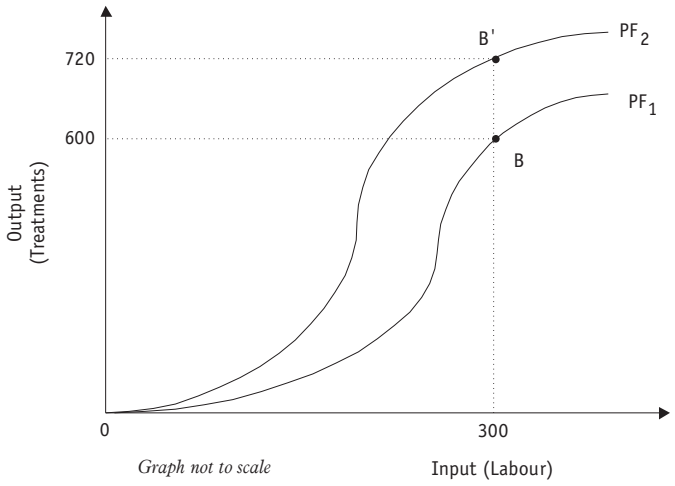
The area from the curve to the horizontal axis comprises different possible combinations of the output and input (input-output combinations above the curve are not feasible unless a new technology is introduced). For example, at Point A it is possible to produce 450 treatments using 300 hours of labour. The productivity ratio, however, of this combination is not at the maximum possible since it is technically possible to produce more treatments using the same amount of labour hours. At Point B, the hospital can produce 600 treatments using the 300 labour hours; the given production technology does not allow more treatments to be produced using this amount of labour hours. Point B is a point of **technical (or productive) efficiency**.[2] Put differently, Point A is technically *in*efficient compared to Point B.

Points on the upper bound or "curve" of the production function (i.e., B, C and D) are all technically efficient. These points represent the maximum output that can be produced for a given level of input. If the hospital were able to move from Point A to Point B through improved labour skills, its productivity would increase from 1.5 treatments per labour hour to 2.0, representing a 33 per cent increase in productivity. Hence, one way for a hospital to increase its productivity is to improve its technical efficiency.

Next, we turn to the relationship between productivity and the hospital's **scale of operations**. Suppose that the hospital can readily adjust its scale or size of operations. When the hospital increases its scale by moving from B to C, its labour hours increase from 300 to 400 hours or 33 per cent, but the number of treatments produced increases from 600 to 1,000 units or 67 per cent. Since the proportionate increase in labour hours is less than the proportionate increase in treatments, the productivity ratio increases from 2.0 to 2.5. This illustrates the concept of **scale economies**. Hence, higher productivity can also result from exploiting economies of scale.

In general, increasing the employment of labour and capital resources allows managers to subdivide tasks so that inputs can be specialized and productivity improved. In a health care context, an increased scale of operations may

Chart 2
Productivity and Technological Improvement (single output/input)



improve inventory management. A hospital's inventory of medical supplies may not need to increase at the same rate as output growth due to probabilistic considerations.

From C to D, labour hours increase by 250 or 63 per cent but the number of treatments produced increases by only 170 units or 17 per cent. In this case, the input increases by a greater proportion than the increase in output. This results in a reduction in the productivity ratio from 2.5 to 1.8. At this size, the hospital faces **scale *dis*economies**.

Scale diseconomies typically arise from coordination problems beyond a certain point. Restructuring a service delivery unit into two or more smaller entities is a possible solution to scale diseconomies.

Based on our illustration in Chart 1, Point C represents the *optimal* scale for the hospital; that is, the size that gives the highest possible productivity ratio. This illustration shows that the hospital could improve its productivity if it were in a position to adjust its scale of operations until it reaches this optimum point.

Lastly, we turn to the relationship between productivity and **technological change**. A positive technological change is illustrated by an outward shift in the hospital's production function between two periods. In Chart 2, the production function from the previous figure is reproduced to represent Period 1.

Providing staff with additional skills allows the hospital to produce more treatments using the same amount of labour at every level of production. The production function (PF) in Period 2 lies above the production function in Period 1. Point B' is on the production function in Period 2. At Point B', the hospital uses 300 hours of labour to produce 720 treatments, giving a productivity ratio of 2.4. Hence, the hospital's productivity increases from 2.0 at Point B to 2.4 at B' due to the use of advanced skills.

To summarize discussion to this point, improvements in an organization's productivity ratio over time (or productivity growth) can be attributed to three major sources: increases in technical efficiency, exploitation of scale economies, and technological advances.

Productivity and Cost Minimization

So far, all discussion has involved the relationship between the physical units of inputs and ouputs. We have not discussed financial concepts such as costs. If input price information is available, such as the wage rate, we can analyze the relationship between productivity and average costs.

Return to Chart 1 and assume that the price of labour is constant at $350 per hour. Consider Point B. The hospital uses 300 hours of labour. Therefore, the total cost is $105,000. Since the hospital at Point B produces 600 treatments the cost per unit (or average cost) is $175.

Similarly, at Point C the hospital uses 400 hours at $350 per hour. The total cost is $140,000 while the total output is 1,000 units. Therefore, the average cost is $140. Table 2 summarizes the computation of average cost for the four illustrated points.

As the hospital increases its scale of operations from B to C, *total* cost increases but *average* cost decreases due to the impact of scale economies. As the hospital moves from C to D, however, it encounters diseconomies of scale. Total costs increase at a faster rate than the increase in output.

As a result, the average cost increases from $140 to $194. Point C represents the optimal scale for the hospital, where the productivity ratio is at a maximum and average cost is at a minimum. Hence, productivity movements are correlated to changes in an organization's cost structure.

Multiple Inputs and Outputs

When only a single input and a single output are involved in an organization's production process the calculation of the productivity ratio is easy. For a production process that has more than one input or output, a method for combining these multiple measures into a single, aggregated quantity is needed to calculate the ratio.

When an organization's production process uses multiple inputs the concept of technical efficiency becomes more sophisticated. It raises issues about the relationships among inputs in the production process; for example, whether they can be substituted for each other and if so at what rate.

Where inputs are substitutable, the organization may be able to reduce its cost of production if it chooses an input mix that minimizes its total costs at prevailing input prices.

When an organization chooses an appropriate mix of inputs and/or outputs that takes into account given market prices, it attains **allocative efficiency**. The combination of allocative and technical (or productive) efficiency gives a measure of **cost efficiency**.[3]

Table 2
Calculation of Average Cost

Point	Number of labour hours x	Labour price ($) w	Cost ($) (x×w)	Number of treatments y	Average cost ($) (x×w)/y
A	300	350	105,000	450	233
B	300	350	105,000	600	175
C	400	350	140,000	1,000	140
D	650	350	227,500	1,170	194

Overview of Economic Performance Measurement Techniques

In this section we provide an overview of available economic performance techniques for measuring productivity and technical efficiency using a series of simple numerical examples to illustrate the main concepts. Three broad approaches to economic performance measurement will be discussed — index numbers, statistical, and mathematical programming.

Measuring Productivity

In this section, we consider a simplified model of a hospital, called XYZ. Measuring productivity performance first involves defining and then collecting data on the outputs and inputs used in the production process. Outputs are associated with a receipt (for commercial activities) while inputs are linked with a payment.

For our example, we will assume that the hospital is part of a tax-financed, public health system, and provides treatments that are free of charge to users (patients).

The hospital produces two outputs:

- *inpatient treatments* measured as number of cases; and
- *outpatient visits* measured as number of consultations.

Table 3
Output and Input Data of XYZ Hospital, Years 1 to 4

Year	Inpatient treatments (no. of cases) y_1	Outpatient visits (no. of consultn.) y_2	Labour (no. of FTE staff) x_1	Contractors (no. of hours worked) x_2	Capital (no. of beds) x_3	Other inputs (imputed quantity) x
1	800	45,000	1,300	250,000	300	200,000
2	820	47,500	1,250	275,000	305	210,000
3	850	50,000	1,200	300,000	310	215,000
4	875	52,000	1,175	330,000	314	230,000

These outputs are produced using four inputs:

- *labour* measured as number of full time equivalent medical (e.g., doctors) and administration staff;
- *contractors* measured as hours worked by visiting medical officers and other contractors (e.g., cleaners);
- *capital* measured by a proxy, number of beds; and
- *other inputs* measured on an imputed quantity basis.

The measurement of the other inputs series requires some explanation. This series comprises a mixture of residual inputs that share no common physical measure; for example, pharmaceuticals and patient meals. A physical estimate can be "imputed" or measured by deflating the expenditure on these items by an appropriate price index.

This can be simply explained from the simple accounting identity where:

$$\text{Cost} = \text{Price} \times \text{Quantity}$$

Rearranging the terms of the identity gives:

$$\text{Quantity} = \text{Cost}/\text{Price}$$

If a price index is used, and given the cost measure, a notional quantity measure can be defined.

The hypothetical data for XYZ Hospital is presented in Table 3.

A distinction is usually made between the performance of individual factor inputs (partial factor productivity) and overall productivity (total factor productivity). Both measures can be applied at any operational level where sufficient data is available.

Partial factor productivity (PFP) ratios are the most common form of productivity measurement. The PFP measure for XYZ hospital can be calculated as the ratio of each of the outputs to each of the four inputs. For example, the labour partial factor productivity ratio using the inpatient treatment output is:

$$\text{PFP} = \frac{\text{Inpatient treatments (no. of cases)}}{\text{Labour (FTE workforce)}}$$

Although the hypothetical data comprises four periods, a single period is sufficient for the calculation of a partial ratio and a minimum of two periods is required for the computation of a productivity change. XYZ hospital's four series of PFP ratios using the inpatient treatment measure of output are presented in Table 4a:

Each series of PFP ratios can be converted into an index by following three steps:

- selecting a base value for *each* PFP series, say their values in Year 1;
- dividing the base value by 100; and
- dividing every ratio in the particular series by the resulting amount.

The index series for each PFP ratio are presented in Table 4b.

PFP percentage changes can be easily calculated from the either the ratios (Table 4a) or index numbers (Table 4b). For example, labour productivity increases by 21.1 per cent between

Year 1 and Year 4: ratio calculation [0.745/0.615 - 1] or index calculation [121.1/100.0 - 1]. Annual percentage changes for the four PFP measures are presented in Table 4c.

From Year 1 to Year 2, the labour productivity ratio increases by 6.7 per cent. Contractor productivity declines by 6.3 per cent over the same period. Meanwhile, capital productivity increases by 0.8 per cent. The other inputs' productivity measure falls by 2.5 per cent.

Partial factor productivity measures are conceptually simple and easy to calculate. These measures, however, need to be interpreted with care, as they do not provide a complete picture of performance. Notably, they can move in different directions, as is the case in our example. XYZ hospital may have improved its labour productivity figures by merely substituting contractors for permanent staff.

Total factor productivity measurement addresses this shortcoming by providing an overall indicator of performance. The general definition for a total factor productivity (TFP) measure is:

$$\text{TFP} = \frac{\text{Combined output quantity index}}{\text{Combined input quantity index}}$$

Table 5 shows TFP indices for the hypothetical data set presented in Table 3. Year 1 is set as the base period for the output and input index series. The output index is used to aggregate or combine inpatient treatments and outpatient visits into a single measure. In the same way, an input index is used to combine the four input quantities into one series.

While creating a combined index series may seem like adding apples and oranges, a standard approach is to use price data as "weights" for combining the different quantities. In this case, since there are no market prices for the outputs (as they are provided free to patients), a proxy price needs to be used to capture the different resource demands of the two outputs.

In this case, based on actual practice we can justifiably assume that the resources required to treat one inpatient case equals, on average, those required to perform 40 outpatient consultations for each of the four years. Hence, if the inpatient treatment "price" is set at $2,400 per case then the price for outpatient visits will be $60 per consultation. The calculation of the input quantity uses

Table 4a
Partial Factor Productivity Ratios (Inpatient Treatment Output Measure), XYZ Hospital

Year	Labour productivity ratio y_1/x_1	Contractor productivity ratio y_1/x_2	Capital productivity ratio y_1/x_3	Other inputs productivity ratio y_1/x_4
1	0.615	0.0032	2.667	0.0040
2	0.656	0.0030	2.689	0.0039
3	0.708	0.0028	2.742	0.0040
4	0.745	0.0027	2.787	0.0038

Table 4b
Partial Factor Productivity Index Series (Inpatient Treatment Output Measure), XYZ Hospital, Year 1 =100.0

Year	Labour productivity ratio y_1/x_1	Contractor productivity ratio y_1/x_2	Capital productivity ratio y_1/x_3	Other inputs productivity ratio y_1/x_4
1	100.0	100.0	100.0	100.0
2	106.7	93.8	100.8	97.5
3	115.1	87.5	102.8	100.0
4	121.1	84.4	104.5	95.0

Table 4c
Partial Factor Productivity Annual Percentage Changes (Inpatient Treatment Output Measure), XYZ Hospital

Year	Labour productivity ratio y_1/x_1	Contractor productivity ratio y_1/x_2	Capital productivity ratio y_1/x_3	Other inputs productivity ratio y_1/x_4
1	-	-	-	-
2	6.7	-6.3	0.8	-2.5
3	7.9	-6.7	2.0	2.6
4	5.2	-3.6	1.6	-6.0

Table 5
Total Factor Productivity Analysis, XYZ Hospital

Year	Combined output index Yr 1 = 100	Combined input index Yr 1 = 100	Total factor productivity ratio Yr 1 = 100	TFP change on previous year (%)
1	100.0	100.0	100.0	-
2	102.9	99.8	103.1	3.1
3	106.8	99.2	107.7	4.4
4	110.1	101.1	108.9	1.1

the same approach, though actual market input price data are used as "weights" for aggregating or combining the different components.[4]

The calculation of the output and input indexes is based on a commonly used index number formula known as the Laspeyres index. The computation of this index requires data on prices of all outputs and inputs for Years 1, 2 and 3.

The Laspeyres index number formula uses the "chain" form. This means that it uses a flexible or "moving" set of base-period prices to weight changes in outputs and inputs instead of a fixed set of base-period prices (say for Year 1). For example, using the chain form, the change in the combined input index between Year 3 and Year 4 (the current period) is calculated using Year 3 prices to weight the quantity changes.

Note that the partial factor productivity measures in Table 4c show positive changes for the labour and capital factors, and negative changes for contractors and other inputs (except for one year). In contrast, TFP performance improves consistently during the period. The TFP measure weights the individual contributions of the four inputs. Note also that total factor productivity techniques need more data than partial factor productivity measures since they require price information for construction of the index weights.

Where price data is either incomplete or distorted, alternative techniques can be used to produce a measure of total factor productivity change.[5] Typically, this is the case with the provision of public services by general government agencies, which do not assign market prices to their multiple outputs.

There is one major limitation of using the index numbers to measure total factor productivity. Total factor productivity change reflects the impact of all three sources of change — changes in technical efficiency, scale and technology. Index number techniques, however, do not provide any means for dissecting TFP change into these three components in practice.

Measuring Efficiency

Index Number Techniques

In principle, the scope for measuring technical efficiency can be assessed by simply comparing a set of productivity ratios across a group of organizations at a point in time. We present an approach to using index numbers to measuring efficiency by extending our previous example.

The focus shifts from analyzing the productivity performance of a single hospital over a period of time (time series analysis) to comparing performance across a group of hospitals at a single point in time (cross-sectional analysis). The empirical measurement of technical efficiency is framed in terms of a benchmarking exercise.

Technical efficiency for each hospital can be viewed in terms of performance *relative* to its peers. Technical efficiency is defined in relation to an individual firm's production technology. In practice, an organization's production technology cannot be easily observed.

Typically, we only have *observable* output and input data (e.g., number of FTE staff) to work with. Our approach to measuring efficiency implicitly assumes that the best performers in a group are using their (common) production

Table 6a
Output and Input Data for Group of Hospitals

Year	Inpatient Treatments (no. of cases) y_1	Outpatient Visits (no. of consultn.) y_2	Labour (no. of FTE staff) x_1	Contractors (no. of hours worked) x_2	Capital (no. of beds) x_3	Other Inputs (imputed quantity) x_4
XYZ	800	45,000	1,300	250,000	300	200,000
ABC	900	50,000	1,400	300,000	320	220,000
LMN	700	48,000	1,200	250,000	280	160,000
HIJ	1,000	60,000	1,500	340,000	420	240,000

Table 6b
Comparative Partial Factor Productivity Ratios and Rankings (Inpatient Treatment Output Measure)

Hospital	Ratio 1 y_1/x_1	Ratio 2 y_1/x_2	Ratio 3 y_1/x_3	Ratio 4 y_1/x_4
XYZ	0.615 (3)	0.0032 (1)	2.667 (2)	0.0400 (4)
ABC	0.643 (2)	0.0030 (2)	2.813 (1)	0.0041 (3)
LMN	0.583 (4)	0.0028 (4)	2.500 (3)	0.0044 (1)
HIJ	0.667 (1)	0.0029 (3)	2.381 (4)	0.0042 (2)

technology in an optimal manner; that is, they are operating at best practice "on the frontier".

Table 6a shows the inpatient output measure and full set of input data for the four hospitals, including XYZ in Year 1. Table 6b shows the set of partial factor productivity ratios and their rankings across the four hospitals.

Using simple PFP ratios, there is no clear way of determining technical efficiency. For example, HIJ hospital is the top-ranking performer for Ratio 1, the third ranking performer for Ratio 2, and the worst performer for Ratio 3. The picture would become even less clear if we considered the partial ratios incorporating the other output measure, outpatient visits.

This example illustrates that there is no clear way, even for a small group of only four organizations, to assess efficiency using partial productivity ratios because different ratios produce different performance rankings. Moreover, it is not possible to identify which organization is inefficient and the magnitude for potential improvement.[6]

Statistical and Mathematical Techniques

We now turn to the statistical techniques of ordinary least squares regression and stochastic frontier estimation, and the mathematical programming technique of data envelopment analysis. Unlike index number techniques, statistical and mathematical programming techniques do not require price information to calculate technical efficiency where organizations have multiple inputs and/or multiple outputs. These techniques, however, typically require data for a larger number of entities than the index numbers approach.

The statistical approach requires explicit specification of a production function (i.e., the mathematical relationship between inputs and

Chart 3
OLS Regression Approach to Measuring Efficiency

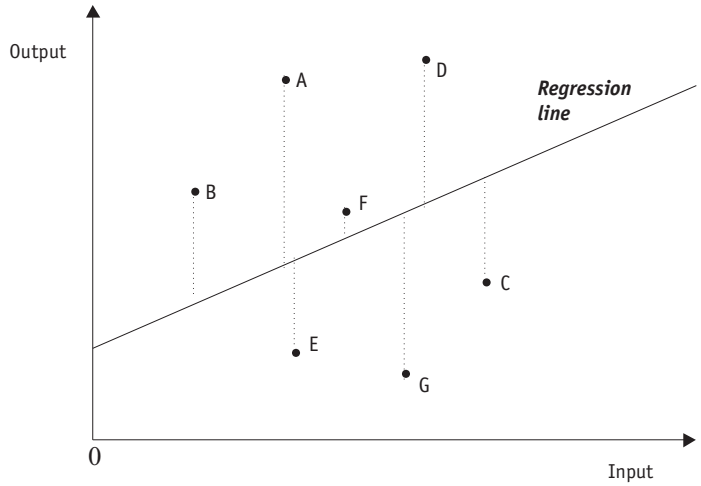
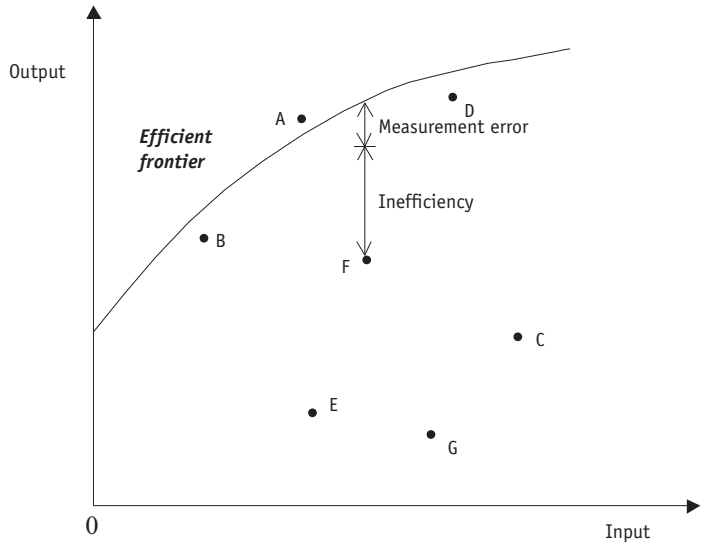


Chart 4
Stochastic Frontier Analysis



outputs) but assumes that the relationship between inputs and outputs is inexact due to measurement error and other factors. This property is captured by the inclusion of an error term that has well-defined probabilistic properties.

By comparison, the mathematical programming approach does not presuppose a particular functional form but "allows" the output and input data to determine the shape of the efficiency frontier. Moreover, the programming approach assumes an exact or deterministic relationship between inputs and outputs, which makes it sensitive to measurement error.[7]

The application of **ordinary least squares (OLS) regression** (see Chart 3) to estimate an industry production function produces a measure of efficiency that is influenced by average practice rather than best practice.

The OLS technique identifies a "line of best fit" through a data set of output/input ratios for a group of organizations. There will, in general, be a discrepancy between the output implied by the regression line for a given level of input (which represents average practice) and observed output at that level. This difference will, by assumption, be attributed entirely to systemic efficiency differences (rather than to random factors). The efficiency of organizations is be ranked according to these differences. The most efficient organization will, by definition, be that with the largest positive difference (see Point A in Chart 3).

Stochastic[8] **frontier analysis** (SFA) is a more advanced statistical technique as it assumes that the gap between predicted and observed performance can be dissected into components for inefficiency and random noise (which is mainly measurement error). A simple stochastic frontier is illustrated in Chart 4 for a group of organizations that each produces a single output using a single input.

The stochastic frontier identifies the predicted performance for the best organization (A), allowing for measurement error. The other organizations (e.g., B) are below this frontier and are therefore relatively inefficient compared to the best. For these organizations SFA assumes that some of the gap between actual and predicted best performance will be measurement error. Empirical work using SFA requires the use of specialist computer software, such as FRONTIER (Coelli, 1996).

Data envelopment analysis (DEA) uses mathematical programming to construct a production frontier comprising a set of linear segments. The

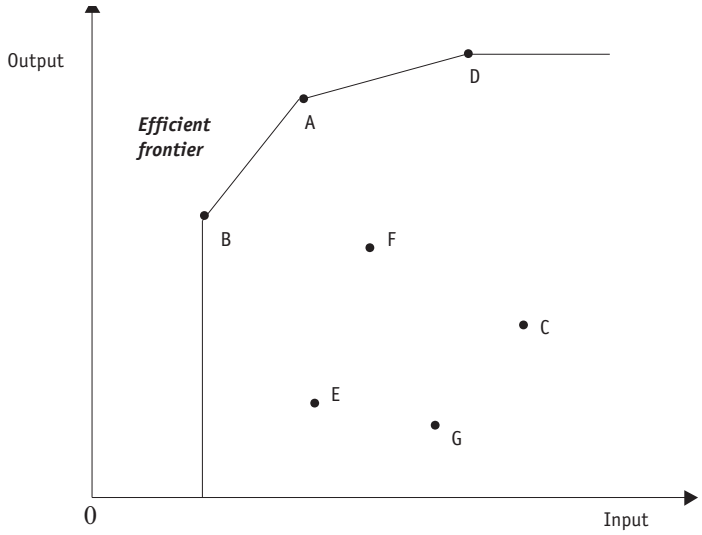
frontier relates to best performance at a point in time. The points separating the segments are from the best practice organizations within a sample. A simple example is illustrated in Chart 5 for a group of organizations that produce a single output using a single input. The frontier "envelopes" the entities with the best output/input ratios. In comparison, a stochastic frontier is estimated using a regression approach from the most efficient organization within a sample.

The distance of an inefficient organization from the frontier is the measure of its inefficiency. A number of approaches can be applied to SFA and DEA efficiency scores to adjust for the influence of the operating environment (i.e., factors beyond management control) such as climate and population density. Like stochastic frontier analysis, DEA empirical applications require the use of specialist computer software.

In general, frontier techniques need more data than total and partial productivity index measures as they are benchmarking techniques requiring a data set of organizations/service delivery units for comparison. DEA results can be sensitive to the number of variables included (i.e., both inputs and outputs) and the sample size. Reducing the sample size will tend to inflate the average efficiency score as it creates fewer comparable organizations and improves the likelihood of any entity being placed on the frontier "by default".

As a non-statistical technique, DEA is sensitive to outliers in the sample, which are often due to measurement errors and/or random events such as climate. By contrast, SFA is less susceptible to outliers as it allows for random noise in measuring inefficiency. In addition, a mathematical functional form (e.g., the shape of the curve for two dimensions or the shape of the plane for three dimensions) representing the underlying production technology has to be assumed *before* the stochastic frontier is estimated. By contrast, DEA does not presuppose a particular functional form for the frontier but allows the data to determine the shape of the frontier (e.g., in effect as a set of linear segments in two dimensions, or as flat triangles in three dimensions).

Chart 5
Data Envelopment Analysis



For each organization inside the frontier that is found by DEA to be inefficient, the technique identifies at least one organization on the production frontier that is a "peer" or role model to the inefficient organization. The technique assigns a weight to each peer, reflecting the relevance of that peer to the inefficient organization. DEA can determine whether an organization's technical inefficiency is primarily related to waste (i.e., the use of too many inputs to produce a given level of outputs), or to the particular scale of operations.

If input and output data for a set of similar organizations is available over a period of time (panel data), then total factor productivity can be measured in a more sophisticated way than the "pure" index number approaches discussed earlier in the Guide. This is achieved by combining a frontier technique (DEA or SFA) with the Malmquist index number formula. This approach allows a change in productivity to be dissected into the three sources: changes in technical efficiency, scale and technology.

Conclusion[9]

A significant share of resources in most industrialized countries is allocated to providing public services. The amount of spending in these taxpayer-funded areas and their important links to the rest of the economy provide a strong in-principle case for the use of robust quantitative performance assessment techniques.

In practice, performance assessment within the general government sector is a challenging exercise for three major reasons. First, it is often difficult to define and measure general government sector outputs (and therefore productivity or efficiency), particularly in view of their unpriced nature. Second, public sector information systems have traditionally been framed around budget compliance requirements rather than service delivery performance (namely the efficiency and effectiveness of service delivery and their contribution to broader government outcomes). Third, the incentive structure implicit in the budget decision-making process can hinder efforts by an entity to improve its efficiency due to the risk of future budget cuts.

In response to these challenges, governments need to be pragmatic in their approach to applying the techniques outlined in this article. No single technique can provide a complete picture of performance; each technique has its own particular strengths and weaknesses. Also, tradeoffs in specifying models are inevitable; analysts should focus on the major outputs and inputs, and environmental variables that shape performance. Lastly, central government policy entities ought to strive for better information management systems, and to be advocates of better incentive structures in the delivery of public services.

Notes

* New South Wales Treasury. The author would like to thank Mr. Richard Cox, Director Fiscal Strategy, NSW Treasury, Professor Suthathip Yaiaswarng (Union College, New York, United States) and Professor Tim Coelli (University of Queensland, Australia) for much-needed advice on the manuscript. This article is an abridged version of the document *Guide to Economic Performance Measurement for General Government Sector Agencies* published by the Office of Financial Management of the New South Wales Treasury in August 2001. The unabridged version is posted at www.csls.ca under the *International Productivity Monitor* and at www.treasury.nsw.gov.au/indexes/tppdex.html

1 The unabridged version of the Guide provides a detailed discussion of definitions of hospital outputs, inputs, and sources of variation in hospital performance.

2 It is worth noting at this point that the terms "productivity" and "efficiency" are typically used interchangeably in business and government circles, which may be a source of confusion for a reader not familiar with economic techniques.

3 For a detailed discussion of allocative and cost efficiency concepts and applications the reader is referred to Coelli, Rao and Battese (1998).

4 In principle, a combined output index could be calculated in a similar way if the output was sold; the output prices (or relative revenue shares) would provide weights.

5 Where output price data is not available, statistical (stochastic frontier analysis) and mathematical (data envelopment analysis) techniques can be combined with the Malmquist index number formula to measure total factor productivity change at the organizational/service delivery level. These techniques are discussed in the context of their application to measuring efficiency in section 4.3.2. in the unabridged version.

6 To calculate an unambiguous measure of relative performance using an index number approach, a comparison of total factor productivity ratios is necessary. A "multilateral" index can be used to facilitate a comparison of TFP performance across a group of organizations over time. A multilateral index compares each organization in an industry to a hypothetical representative entity. The representative organization is "constructed" from average output and input data derived from all data in a given panel. Input price (or cost) data and output price (or revenue) data is required for weighting the quantity changes of the individual entities and the representative organization (see Caves, Christensen and Diewert, 1982).

7 In most practical applications measurement is inexact and therefore is subject to error.

8 Stochastic means "a problem involving probabilities, as opposed to a deterministic problem based on certainties" (The Economist, 1991).

9 Recommended publications for more in-depth information on this topic, see Blank (2000), Coelli, Prasada Rao, and Battese (1998), and Fried, Knox Lovell and Schmitt (1993) and Steering Committee for the Review of Commonwealth/State Service Provision (1997). For computer software, the Centre for Efficiency and Productivity Analysis at the University of New England (Internet: http://www.une.edu.au/febl/EconStud/emet/cepa.htm) has three "freeware" products available for estimating stochastic frontier models to measure technical and cost efficiencies; constructing data envelopment analysis frontiers to measure technical and cost efficiencies as well as Malmquist total factor productivity indices; for measuring total factor productivity using Tornqvist and Fisher index number formulae. Other software is Frontier Analyst, a Windows-based DEA software supplied by Banxia Software (www.banxia.com) and SHAZAM, a widely used general econometrics package that can be applied to efficiency and productivity analysis(http://shazam.econ.ubc.ca).

References

Blank, J. L. T. editor (2000) *Public Provision and Performance: Contributions from Efficiency and Productivity Measurement*, Elsevier Science B.V., Amsterdam.

Caves, D.W, L. R. Christensen, and W. E. Diewert (1982) "The Economic Theory of Index Numbers", *Econometrica*, Vol. 50, No. 6, November.

Coelli, T. (1996) "A Guide to FRONTIER Version 4.1: A Computer Program for Frontier Production Function Estimation," *CEPA Working Paper 96/07*, Department of Econometrics, University of New England, Armidale.

Coelli, T., D. S. Prasada Rao and G. E. Battese (1998) *An Introduction to Efficiency and Productivity Analysis*, Kluwer Academic Publishers, Boston.

Fried, H. O., C. A. Knox Lovell and S. Schmidt editors (1993) *The Measurement of Productive Efficiency: Techniques and Applications*, Oxford University Press, New York.

Steering Committee for the Review of Commonwealth/State Service Provision (1997) *Data Envelopment Analysis: A technique for measuring the efficiency of government service delivery*, AGPS, Canberra. www.pc.gov.au/gsp/gsppubs.html

The Economist (1991) *Numbers Guide: The Essentials of Business Numeracy*, The Economist Books Ltd.