Performance Measurement in the Public Sector

*Alice O. Nakamura**
Faculty of Business, University of Alberta
William P. Warburton
British Columbia Ministry of Social Services

Performance measures are being actively developed in Canada at both the federal and provincial government levels to improve the performance of the public sector and to achieve greater accountability.[1] One important source of friction in discussions of performance measurement in the public sector is that closely related terms are used in different ways. Reviewing the types of measures and issues related to their production and use can help avoid misunderstandings. Another reason for calling attention to the spectrum of different types of performance measures is that choices made about how information is to be collected and stored for one particular set of measures can inadvertently facilitate or impede the later development of other measures. When information is to be produced at public expense, it is important for careful consideration to be given to cost-effective auxiliary uses that might be made of the information.

We begin by providing a nontechnical overview of the different types of performance measures and some of their uses.[2] Next we discuss how these relate to, and can build on, a number of long established private and public sector functions. Formal performance measurement practices have evolved and been used more in a private than in a public sector context.[3] Many of these practices require adaptive adjustment if they are to be successfully used in public sector applications. We call attention to some of these needed adaptations. Finally, the article summarizes our main conclusions.

A Full Spectrum of Performance Measures

Phase of Production

One way of classifying performance measures is by the phase of production that they pertain to (i.e. input, throughput, output, and outcomes performance measures).

Inputs are all the elements brought in from outside a production situation and used in a particular production process or in the multi-purpose ongoing activities of a productive unit. For example, for a welfare program, those applying for welfare are an input. The hours worked of the welfare workers who process welfare applications are an input. Purchased education and job-finding services made available by the welfare program to selected participants are a program input. Hence, possible input performance measures include the number of those applying for welfare, the number of hours of labour expended by program personnel on taking and processing applications for welfare assistance, and the number of spaces in education and job finding programs that are arranged for and made available by the welfare program to those eligible to receive this help.

Throughputs are flow through a production process or program or organizational unit per unit of time and possibly also per unit of some input factor or standardized by some aspect of the external circumstances. From a phase of production perspective, throughputs can be thought of as intermediate products. The welfare applications that are processed each month or year are a welfare program throughput. Likewise, the cases reviewed in a month or a year of

full-time equivalent caseworker time are another welfare program throughput, and the corresponding number of these cases per designated unit amount of caseworker time is a potential throughput performance measure.

The *outputs* are the final services or products directly resulting from the production activities. The size of the caseload handled per month or year is a possible welfare program output measure.

Public programs have been established because of widely held beliefs that the outputs of the programs would help bring about certain desired *outcomes*. For example, one desired outcome of a welfare program is that the children in the families who receive welfare assistance will be protected by this support from long-term poverty-related damage to their health, school performance, and other aspects of development. Another desired outcome is that there will be less crime. For reasons discussed later, the development of public sector outcome measures is typically much more difficult than the development of public sector input, throughput, and output performance measures.

Financial versus Nonfinancial Measures

The examples of possible performance measures that we have given so far are mostly *quantity* measures. There are also *financial* performance measures of inputs, throughputs, outputs, and outcomes. For example, the wage bill for the hours of work of the welfare office workers is a possible financial input performance measure. Likewise, the estimated dollar savings in incarceration costs due to reductions in crime that are attributed to the welfare program are an example of a financial outcome performance measure.

Financial performance measures intrinsically involve both quantity and unit price (or unit cost) dimensions. Obviously, it is the second of these two dimensions that distinguishes financial from nonfinancial measures. In the private sector, many businesses prefer nonfinancial performance measures, most of which are quantity measures, for operations management and for monitoring the work effort of production workers. On the other hand, financial measures tend to be relied on more for financial reporting and financial control within a company and for strategic planning purposes. Financial measures are also used for the evaluation of the performance of the managers, for the evaluation of the various divisions and other operational and administrative units of a company, and for the evaluation of overall company performance.[4]

Absolute versus Contextual Measures

We mentioned the number of welfare applications taken in a given time period as an example of an input performance measure. This is an *absolute* measure. For decision-making purposes, it is generally important to consider absolute measures in some context. Comparisons might be made over multiple time periods for a given geographic area. This would show how the level of applications has changed over time. Or, the figures might be compared among the different geographic regions. However, the populations of the geographic regions might differ greatly, for instance. To control for these population differences, the number of applications in a specified time period might be reported on some sort of a per capita basis. Notice that, in the short run at least, population will not be affected by the operation of a welfare program; rather population size is an *exogenously determined contextual variable*.

Two types of *contextual* measures deserve special mention because they are widely used in the private sector and because the standardization for the context for these measures is with respect to a *choice variable* of the production process. These two types of measures are input-output ratios and productivity measures.

Broadly defined, an *input-output ratio* is any given measure of input for a production process divided by some measure of output of the process. Hours of welfare caseworker time in a month or year per welfare application processed is an input-output ratio. This particular example involves a nonfinancial input and a nonfinancial output measure. The cost of running the welfare program expressed on a per case basis (i.e. expressed as a *unit cost* where the unit is a welfare case) is an input-output ratio involving a financial input measure (the cost of running the program) and a nonfinancial output (the number of cases). Because these ratios give the amount of input used per unit for some measure of output, they are also commonly referred to in private

sector contexts as *measures of efficiency* or simply as *efficiencies*.

Productivity measures are the reciprocals of input-output measures. For example, the average caseload per welfare worker is a productivity measure.

Although relative measures such as input-output ratios and productivity measures are often more relevant for analysis and decision making than absolute ones, it is important for the information storage to be in absolute rather than relative terms. The main reason for this is that different contexts may be relevant at different times or to different parties.

An example may again help clarify the concepts. The proportion of the population that is elderly has been growing over time. Moreover, the elderly are not usually eligible for welfare because they are eligible for other forms of income assistance. Thus it is desirable to switch from a general per capita measure of the number of welfare cases to the number of cases per household head who is within the age limits for eligibility for welfare. It would be easier to make such a switch, particularly for past years, if the total numbers of cases rather than just the calculated per capita numbers of cases were available. The absolute figures can also be valuable for exploring the validity of causal hypotheses relating program outputs to observed outcomes -- a topic we turn to later in discussing a special sort of contextual standard for performance measures called counterfactuals.

Aggregation in Data Collection and Performance Measure Construction

Most performance measures involve some types and amounts of aggregation. Aggregation can be carried out with respect to the definition of the operational unit, with respect to the unit of time, and with respect to the operational task.

Consider an input performance measure consisting solely of the number of applications for welfare. A first point is that it is not possible to compute this simple measure without adding some aggregation specifics, beginning with the operational unit.

Suppose that the applications of those seeking welfare assistance are received by individual welfare office workers. In this production scenario, the individual office worker is the lowest possible *operational unit level of aggregation* for the process of receiving applications for welfare. For many policy questions, this lowest level of aggregation is not the relevant level. Rather, figures are wanted at the welfare district office level, or for broader geographic regions of the province, or for the province as a whole.

Suppose that the operational units of interest are broad regions of the province corresponding to what are viewed as the main labour markets. Suppose furthermore that the various welfare offices send their figures on numbers of applications received to a central office that has responsibility for computing the input performance measure. Notice that the district level offices would not necessarily need to keep any records at all on numbers of applications received by the individual workers in order to compile their office level figures. They could simply maintain a running total of applications received for the office as a whole. Moreover, these district offices would not need to store even the office level figures over time. For the purposes of computing the designated input performance measure, there is no need for information to be collected at the level of the individual worker, and there is no need for the district office level information to be retained after it has been used by the central office to compute the broader area totals.

But if this more disaggregated information is not collected or retained, it will not be possible to go back subsequently and compute, say, the corresponding figures for an individual worker level throughput measure for the number of applications taken. The point is that data collected or stored at one level of aggregation in terms of the operational unit can usually be further aggregated, provided that information is available on the characteristics with respect to which aggregation is desired, but subsequent *dis*aggregation is not usually possible. Because of this, it is desirable to institute the data collection and data storage for a performance measures system at the *lowest* level of aggregation for which performance measures are wanted, now or in the foreseeable future, subject of course to cost considerations. Cost considerations have become far less of a constraint in terms of data storage because of advances in computer-related technology.

It is now desirable to undertake data collection at the lowest possible operational unit level of aggregation. This already happens by default in government departments that have massive record keeping systems. By definition, the lowest operational unit level is usually the actual level of operation: the level where the ultimate answers to many of the problems and questions arising from the monitoring and analysis of more aggregated performance measures are likely to lie. (Data collection at all operational unit levels above the actual operational level usually involves aggregation as part of the data collection process.)

In addition to aggregation decisions that must be made along the continuum of the unit of operation -- from the individual worker level to the whole welfare program in the context of our simple example -- decisions must also be made about aggregation with respect to *the dimension of time*. The number of applications taken could be counted on an hourly, daily, weekly, monthly, quarterly, or yearly basis, or for even longer periods such as a decade. The time dimension could also be specified in terms of events or circumstances, such as for business cycles, rather than in term of calendar measures of time. Again, information can usually be aggregated to higher levels after collection. However, it will not usually be possible to subsequently produce measures for lower levels of time aggregation than the level used for information collection and also information storage.

A choice must also be made about the *task dimension of aggregation*. Consider again the operation of taking an application for welfare assistance. If this is simply a matter of an applicant handing a completed application form to a welfare office worker, as depicted in our simplified example, then the receipt of the applications is the lowest possible task breakdown for information for this operation, and hence it is the lowest task level of aggregation at which information for this process can be collected and for which a performance measure could be constructed. But suppose instead that taking an application means that a welfare office worker first talks to a person interested in applying, and then helps that person understand the information required to apply, and enters the information verbally provided by the applicant into a computer file. This is a *sequence* of tasks having to do with the receipt of a welfare application, and information could potentially be collected and performance measures could be computed for each specific job task making up this sequence. In fact, some of these job tasks could potentially be carried out by different workers in a welfare office. Again, it will usually be possible to produce performance measures for higher task-specific levels of aggregation than the level at which the data were collected and stored, but not for a lower task level than this.

Integrated Systems of Performance Measures

Often, not all of an organization's intended uses for performance measures can be satisfied with one measure. For example, in a manufacturing plant there is often a set of quantitative performance measures that are used for operational monitoring and evaluation, and a second set of financial measures that are used for business planning purposes and for reporting to shareholders and lenders. With differing performance measures, and the differing objectives underlying the creation of the differing measures, an organization can be pulled in differing -- and sometimes conflicting -- directions.

To the extent possible, the goal is to develop performance measures that fit together within an integrative, managerially relevant framework. This has proved difficult to achieve even in for-profit firms. For instance, it is difficult to determine the appropriate implicit prices that are needed to explicitly relate quantity performance measures for the intermediate products of a firm to the financial outcomes for the business. Developing internally compatible and integrated systems of performance measures for public sector use is a more difficult challenge because there are no market prices for many of the final outputs being produced.

Uses of Performance Measures

We have classified performance measures according to the phases of production: input, throughput, output and outcome measures. The potential uses of measures follow these production phases too. Measures of inputs, throughputs and outputs can provide information on the efficiency of these phases of operation and can provide guidance to decision-makers on day-to-day resource allocation, scheduling, equipment

maintenance, and other operational decisions. Typically, these measures must be compared with similar measures from other jurisdictions or time periods, controlling for differences in quality. Performance measure information of these types can be useful as well for strategic planning.

In addition, performance measures can be used in reporting to stakeholders, including voters. Reporting of this sort can include figures for and analyses of quantity measures of output, quantity measures of throughput, and comparative measures of the quantities of specific outputs to demonstrate efficiency of operation. It can also include financial measures of business success, or measures that demonstrate that a business is practicing good financial control and is financially accountable. Reporting of this sort can include survey or testimonial information from users or recipients of the outputs concerning their levels of satisfaction. Finally, outcome measures can be used for strategic planning and policy formation, and can contribute to debate in a democratic society about the advisability of continuing to fund particular program and administrative units by providing information on their intended and unintended consequences.

Building on Established Public Sector Activities and Expertise

Actually, a number of the processes involved in performance measurement are long established public-sector practices.[5] In designing and implementing a new performance measurement system, cost effectiveness, avoiding mistakes, and acceptance and confidence in the new system are more likely to be achieved if this system makes appropriate use of information already being collected and builds on the expertise gained from the related and already established data collection, monitoring, evaluation, and research functions within the public sector. Our main purpose in this section is to draw attention to what these other established functions are.

Administrative Data Information Systems

The federal and the provincial governments maintain a number of large electronic data information systems. These support tax programs, public programs such as health care and welfare, government financial expenditures, vehicle and other licenses, and so on. In designing new public sector performance measures, costly duplication of effort can be avoided by carefully examining and making use of relevant administrative data available from these existing information systems. This way, the new system can benefit from the accumulated expertise acquired in running the existing administrative data systems. If improvements are needed for the new performance measures (such as correcting recognized data accuracy problems), these will further support the original functions of these data. Another advantage of making use of existing information is that the personnel who have been responsible for the previously existing data systems will not be as likely to view the new performance measurement system as a direct threat to their job security.

Financial Auditing within the Public Sector

The financial records of governments in Canada are subject to auditing. The auditing functions are conducted in accord with strict legal requirements and established accounting procedures. They are carried out by persons with formal professional qualifications in financial accounting who are institutionally insulated from the organizational units they are responsible for auditing. In addition, the formal rules and procedures for official audits cannot be readily altered by those being audited. These audit practices not only help to insure that public funds are *used* appropriately, but they also help insure the reliability of the resulting financial data. From this perspective, government financial records are a particularly desirable source of information for incorporation into performance measures.

There are two important principals of formal financial auditing that should be applied, as broadly as is practical, in the production of performance measures. Doing this will help safeguard the integrity of the performance measures and will encourage trust in them:

1) There is value in agreeing on and formally stating *which performance measures* will be computed (including the formulas and procedures to be used in

computing these). The stated choices of these agreed on measures should not be readily subject to change from period to period by those being evaluated, or whose personal interests are otherwise directly affected by what the measures may reveal.

2) Attention should be given to safeguarding *the integrity of the raw information* used for computing the performance measures.

Government performance measures and reports that are produced for the stated purpose of improving accountability to the public, but which consist primarily of measures that were proposed by and are controlled by the same administrative units that are being reported on, do not satisfy the above two principles.

Personnel Evaluation

Like large private-sector organizations, the federal and provincial governments have extensive and long established personnel monitoring and evaluation practices. The resulting information is another potential source of information for the construction of performance measures.

In addition, there are some aspects of the usual conduct of personnel evaluations that could contribute to better performance measures. The first of these is careful attention to contractual job descriptions. In constructing performance measures for workers or for organizational units, and in interpreting and using performance measure results, it is important to know the formal job descriptions of those involved. Performance measures focusing on aspects of a process over which those whose performance is being measured have only partial or no control may engender anxiety or feelings of unfair treatment. A second and related aspect of personnel evaluation practices that could be beneficially carried over into the performance measurement area is explicit attention to potential morale and behavioural incentive effects.

Public Opinion Monitoring

Many government departments conduct public opinion surveys that focus on public satisfaction with their services or other aspects of their functioning. These surveys are another potential source of information that can be used in constructing performance measures. However, it is important to consider the reliability and credibility of the information. The wordings for and also the manner in which public opinion survey questions are asked, as well as the procedures followed in choosing those who are asked to respond, can greatly affect the results of the surveys. Because of this, public opinion information collected by or for a government department without any form of external validation may be useful for internal purposes but may not be credible for external reporting uses.

Program Evaluation

Program evaluation, monitoring and review, are established (and sometimes even legislatively required) components of the ongoing operations of many federal and provincial government departments. For example, as part of operating the Unemployment Insurance program (now Employment Insurance), Human Resources Development Canada collects and analyzes administrative data on caseload totals, on breakdowns of these by attributes such as age and geographic location, on caseload growth, and on other factors for which information is needed for the on-going operation of the program. The department uses these administrative data and other survey data as the basis for both in-house and externally contracted program evaluation studies.

Program evaluation involves five basic components. The first two are often dealt with in the introduction and the institutional background sections of program evaluation reports. These two components are:

1) A *description of the program*, ideally covering its legal basis, its current administrative operating practices, descriptive statistics computed from nonfinancial and financial administrative data, and relevant descriptive statistics computed from other external data sources.

2) A *description of major interactions of the program with other public programs*, ideally including the current legal bases for these interactions, and statistics on these interactions computed from administrative data and from other sources.

The next two components are usually treated as the analytical and empirical core of an evaluation report, with the emphasis in any one evaluation study usually mostly on one or the

other of these components depending on the intended uses of the evaluation, the available data, and the expertise of the evaluator(s). These next two components are:

3) *Empirical estimates of the micro level behavioural effects* of specified program features.

4) *Empirical estimates of the macro level behavioural effects* of the program.

The final component draws on the other four:

5) *Implications of the descriptive review and estimated behavioural effects* with respect to the strategic planning or other purposes for which the evaluation of the program was undertaken.

As already noted, program evaluations usually cover specific time periods for an operating unit or some specific program. In contrast, performance measurement programs typically are continuing exercises that cover whole administrative units, or even an entire provincial or the federal government in a public sector context. Also, program evaluations are usually treated as stand-alone exercises, even when, in fact, they are conducted on a periodic basis for some major programs.

The information resulting from high-quality program evaluations can be valuable for making choices about performance measurement objectives and instruments. For instance, program evaluations have tackled the difficult challenge of determining the actual *outcomes* of public programs. It is also the case that good input, throughput and particularly output performance measure information can greatly facilitate program evaluation research into the causal effects of public programs. Thus, instituting or expanding a system of high-quality performance measures can contribute to improved program evaluations as well.

The Challenge of Adapting Private Sector Performance Measurement Methods for Public Sector Use

The performance measurement methods being promoted for public-sector use have been primarily developed in large-firm, private-sector contexts. These methods require adaptation for use in assessing performance in government.

Overall Goals, Operational Objectives and Operational Efficiency

In the private-sector context, it is generally accepted, on a conceptual level at least, that the main criterion for judging divisional performance is its contribution to the profitability of the firm as a whole (for which there are accepted measures). That is, the *overall performance goal* is fairly clear for private-sector firms. Usually, too, there is a fairly clear understanding of how the activities in each part of the firm are expected to contribute to the overall profitability. One implication of this is that the *operational objectives* of the various parts of a firm are generally accepted within the firm. Moreover, it is usually possible to assess, *ex post*, the extent to which these operational objectives were well chosen given the overall profitability goal. In many firms, it is also possible to measure the *operational efficiency* with which the job tasks associated with the operational objectives were carried out.

In a public-sector context, the overall goals and the operational objectives are often poorly articulated or subject to disagreement. They are typically difficult to measure even when they are clearly stated and agreed on. The hope of many proponents of public-sector performance measurement programs is that these programs will help the public, legislators, and government managers judge how well programs are performing and whether they are achieving what was intended. The process of collecting information and producing measures that are put forward as relevant to desired program outcomes will inevi-

tably make program goals more explicit. Beliefs (and differences in beliefs) about how program outputs relate to the program goals will also become more explicit. This, in itself, can be of considerable value in terms of both improving the actual quality of government and improving citizen understanding of the functions of the public sector.

The Quality Dimension

There is ample private-sector documentation of management disasters that can be traced back to performance measures that failed to account for certain crucial *quality* dimensions of performance. This problem is most likely to arise when the neglected quality dimensions are *operationally competitive* with aspects of performance that are being measured. In general, quantity and quality dimensions of performance tend to be operationally competitive, and quantity is usually easier to measure. To guard against quality being unduly traded off, quality must be monitored as an output. Also, changes in quality, however measured, must be related to the resulting changes in the relevant outcomes of the productive activity.

The average unit sale price may be a good overall indicator of quality in some private-sector contexts, and this does directly affect the profit figures private businesses typically use as their main outcome indicator. Consumers may be unwilling to pay as much if they have been disappointed with the quality of previous purchases. However, many public-sector outputs are not sold in markets where buyers have other alternatives. Many are not sold at all. In a public-sector setting, it will usually be necessary to try to find nonprice checks on quality.

In the case of publicly-subsidized services for which quality can vary greatly, it is vital for quantity-based performance information to be combined with measures that *do* explicitly take account of quality. Note that the relevant dimensions of quality include not only the satisfaction levels of end users and recipients of program outputs and of other stakeholders such as the voting public, but also things such as the protection of organization assets (things like equipment and worker morale), and compliance with laws and government edicts (for example, compliance with employment equity laws).

Often organizations introduce performance measure systems with the idea that they will yield substantial resource savings at the supervisory level, as well as through improvements in operational efficiency. However, a shift to more automated methods of performance monitoring may well require *new* expenditures for monitoring quality. Businesses have found this to be the case even with production processes for physical outputs.[6] Traditional delivery-level supervision methods inevitably incorporate some quality-control functions through the presence, and presumed watchfulness, of the supervisors. Reductions in traditional supervision will tend to result in reductions in the "production" of quality control as a byproduct. Also, private-sector experience suggests that, for performance measures to enhance performance, they must be utilized by thinking managers who ask questions first, rather than reflexively reacting, when the performance measure results turn out to be poorer than expected or hoped for.

Missing Prices and Aggregation Related Limitations

As already noted, most performance measures involve some types and amounts of aggregation. Aggregation is straightforward so long as only one sort of attribute is being measured. However, when there is a need to combine *different sorts* of attributes, there is the problem of determining how these should be weighted.

Suppose that a plant produces the quantity "yA" of good A and the quantity "yB" of good B in a unit time period. Then the data on "yA" and "yB" for that period constitute output quantity performance measures for the production activities of this plant. But how should the information from these two output quantity measures be combined for overall evaluative or planning purposes? One sensible possibility in a private sector, for-profit context is to sum the output quantities for the two goods with each of the output quantities weighted by the unit profit margin for that good.

In the public sector, unit cost information is generally available for inputs but not outputs. When unit price information is missing or not relevant, it is inherently more difficult to devise systems of performance measures that fit together in a meaningful integrative framework,

and that take meaningful account of how micro-level operational performance and decisions contribute to overall operational objectives and goals. Certainly there are many public-sector cases in which it will not be possible to form the sorts of overall financial performance measures that are used in the private sector. Rather, in the public sector it will typically be necessary to make do with information from incomplete and unintegrated, or very loosely integrated, sets of performance measures. In cases like this, knowledgeable decision makers will need to decide how to weigh and use the various sorts of information provided by the different performance measures.

Dealing with Uncertainty about the Underlying Causal Relationships

Another aspect of the challenge of formulating useful overall performance measures in the public sector is the need this raises for understanding the causal connections between program outputs and the desired program outcomes. Consider a desired outcome such as protecting Canadians who have lost their jobs, and their dependents, from permanent damage due to material privation. Operational objectives that have been legislatively enacted for the purpose of achieving this outcome are that the federal government should run an unemployment/employment insurance program, and that the provincial governments should run welfare programs. However, some critics question whether, in the longer run, these public income support programs *really* serve to protect Canadians from the ravages of material privation. Some argue that in the longer run these programs lead to reduced levels of employment

Governments have an obligation to the public to fulfill legislatively enacted (or publicly promised) operational objectives efficiently. Performance measures for on-going operational efficiency can be of value to the government in achieving operational efficiency, in fulfilling operational objectives, and in being accountable to the public about these matters. Operational efficiency measures can help managers recognize areas where more resources are needed or where there is a need for procedural reforms or retraining of personnel or more careful monitoring or vigorous enforcement of existing rules. Performance measures may also contribute to a perception as well as the reality of more equitable treatment of public sector workers if these measures are incorporated into the existing salary and promotions processes. In these respects, the potential value and nature of a public-sector performance measurement system is similar to the private-sector situation.

However, the public sector has greater need than the private sector to seek evidence on whether, and by what causal means, the current operational objectives are helping to achieve the overall goals that were the motivation for adopting these operational objectives. Investigations of causal relationships have traditionally been carried out both within and outside of governments under the heading of research, or the empirical analysis phase of program evaluations. Data having to do with program outcomes is a needed input for research efforts of this sort. The development of more and improved measures of what are believed to be program outcomes can facilitate the needed causal research, which in turn can provide an informed basis for program reform if this research reveals that the underlying causal relationships differ in important ways from what had been explicitly or implicitly presumed. (However, labeling these measures as measures of *performance* may be problematical. A government and government workers can reasonably be commended or blamed on the basis of evidence about the efficiency with which operational objectives are being met, but praise or blame are not necessarily appropriate responses to news of research findings showing that the causal relationships between legislatively or otherwise agreed on operational government objectives and outcomes following from the implementation of these differ from the beliefs on which the programs were premised and enacted.)

The Importance of Counterfactuals

One contextual standard of comparison of special interest in exploring causal impacts are *counterfactuals*. These are estimates of expected outcomes that are made under a hypothetical assumption that a specified action or circumstance did *not* occur, *when in fact it did occur*. An example of how counterfactuals are used may help to clarify the concept. Consider the circumstance

of British Columbia's new three-month residency requirement for eligibility for welfare. Caseload figures for the period since the residency requirement was enacted could be compared with hypothetical figures for what it is estimated that these figures would have been *if* there were still no residence requirement; that is, the actual caseload figures could be compared with the relevant counterfactual of estimates for the situation of no residency requirement. If properly done, this comparison could be interpreted as showing the *causal impact* on the BC welfare caseload of the new residency requirement.

The concept of counterfactual standards of comparison is conceptually akin to the control group or control situation in a scientific laboratory experiment. In fact, social experiments are sometimes used to generate counterfactuals for analyses of public programs.

A Go Slow Recommendation on Linking Compensation to New Performance Measures

Many businesses use performance measures as inputs into personnel evaluation and compensation decisions as well as for operations management. If the performance measures that are used are readily understood and are deemed to be reliable and appropriate by workers as well as management, the use of these measures in evaluation and compensation decisions can contribute to a shared perception of fairness and harmony within an organization. Opposite effects on worker morale can be expected to result from the use of impersonal performance measures which fail to credit, or may even penalize, actions that truly were taken in the best interests of the organization.

Any performance measures that are to be used for personnel evaluation and for the determination of compensation should relate in a clear-cut way to the job descriptions of the personnel involved. Personnel who are judged unfavourably on the basis of organizational goals that they personally cannot affect, or can only partially affect, are likely to feel badly treated.

Special care must be taken with aspects of personnel evaluation and compensation that could widen the inevitable gaps between organizational interests and the private interests of those being evaluated. One of the hopes often expressed when organizations adopt performance measurement systems and build these into their personnel evaluation and compensation mechanisms is that the performance measures will provide a means for communicating and gaining cooperation with organizational goals throughout all levels of an organization. Unfortunately, however, unfavourable changes in these respects are also possible. For instance, a performance measurement system viewed by workers as unfair can help incite and disseminate discontent. This should be borne in mind in implementing new performance measurement programs.

A new system of performance measures usually does not function quite as intended at first. For a phase-in period of a year or more, it may be desirable to use the system *just* for operations monitoring and research, and perhaps for some planning purposes, leaving compensation-related uses until the performance measurement system is functioning properly and its strengths and limitations are understood. This will help avoid unintended damage to the human capital of an organization.

Most mistakes made in handling non-labour productive inputs can be quickly reversed once they are recognized. Non-labour inputs have no feelings. In contrast, once morale has been badly damaged within an organization, recovery can be slow.

Also, leaving the compensation system untouched in the initial phases of implementation for a performance measures system will make it easier to detect behavioural changes that are being caused by the measurement system itself. This will provide a base case against which to judge the impacts of subsequently linking compensation practices to the performance measures.

Reliability and Public Credibility Issues

Most private-sector performance measures are used only within the firm that produces them, often mostly by those within the firm who manage or are responsible for authorizing the production of these measures. Managers can usually be presumed to want the information that is collected at their own request, and solely for their own use, to be accurate. When there are other users as well, of course, managers may

have reasons for wanting to shape or even falsify or suppress performance measurement information that could affect what they see as their interests or the interests of the firm. These sorts of incentives to deliberately distort the information produced by a performance measurement system are likely to be present in public-sector contexts as well.

The managerial usefulness of performance measures is rooted in the reliability of the factual information being summarized. At their best, performance measures can help to focus the attention of management and workers at all levels and in all parts of an organization on the organization's central goals and problems, and can improve the objectivity of resource allocation and personnel evaluation. But poor performance measures can cause systemic misdirection of effort and resources, and systemic morale problems that can be far more damaging to the functioning of an organization than the sorts of problems that typically go along with the spot failures and lack of full coordination that come with managers operating with more informal and less comprehensive information. In this regard, performance measures can be thought of like an automatic guidance or collision-avoidance system for a vehicle.

A driver of a car can only be looking in one direction at a time, and can be distracted by unexpected happenings in one direction and may fail to notice happenings is other directions that could cause a collision. In contrast, an automated collision avoidance system can potentially keep watch all around a vehicle all the time. Also, a computer can much more rapidly and accurately process factual sensory information having to do with distance and speed of approach to obstacles. A driver could potentially be far safer with than without an automated collision avoidance system. This will not be the case, however, if the scanners for the automated system have been mounted inappropriately -- say, so that they can detect obstacles to the right and left but not in front or behind. Nor will it be the case if the scanners are poorly adjusted, or have become spattered with mud, or for any other reason are malfunctioning so that the signals from them to the computer processor are providing misleading information about the locations of other vehicles and obstacles. Even an inexperienced or inattentive driver might do much better at avoiding accidents *without* an automated collision avoidance system that is systematically producing faulty information: information that can lead to *systematically wrong* decisions.

Deliberate falsification of the information in a performance measurement system is one potential reason why these systems can produce wrong information. Even the perceived possibility of this sort of distortion will inevitably impair the usefulness of the system. This brings us back to an earlier point: information that is collected at arms length or taking other effective and observable measures to help insure its integrity is of special value for public sector performance measurement systems, particularly when one central objective of instituting these systems is to provide credible information to the public.

Conclusion

We began by providing an overview of the different types of performance measures. They can be classified by phase of production: there are input, throughput, output and outcome performance measures. Data collection and storage in support of performance measurement systems, and the production of the performance measures themselves, all involve aggregation choices. These choices include the degree of disaggregation or aggregation with respect to operational unit; with respect to the dimension of time, including the nature of time measurement (clock or calendar time versus, say, event-based ways of noting the passage of time such as from peak to peak of business cycles); and with respect to the job-task dimension of productive processes. Financial versus nonfinancial, and absolute versus contextual performance measures were defined and discussed. In addition, we briefly considered various uses for different types of performance measures, and noted that, ideally, the various performance measures adopted by an organization should fit together within an integrative, managerially relevant framework.

Next we noted that a number of the processes and practices that are, or could be, involved in performance measurement are long established public-sector practices. These include the maintenance of large electronic administrative data systems; financial auditing practices within government, and audit-related procedures for pro-

tecting the integrity of information; arms-length data-collection practices; personnel evaluation activities and practices; public opinion monitoring; and, particularly, program evaluation. (In fact, high quality and comprehensive program evaluation potentially spans all of the processes and practices that are part of performance measurement programs, except that performance measurement is normally an on-going activity whereas program evaluations tend to be carried out on an occasional basis.) We argue that building on these established practices in instituting performance measure programs can have important cost-saving, efficiency, and effectiveness benefits.

Finally, we considered some of the challenges involved in adopting private-sector performance measurement methods for public-sector use. Attention was focused on the issues of goal clarification, the monitoring and protection of quality, and missing price aggregation complications. Special attention was paid to the complications resulting from uncertainties about goals and about the causal linkages between program outputs and outcomes, and to the importance of counterfactuals for learning more about these output-outcome linkages. We then concluded with a discussion of reasons why it is important to go slow on linking public sector employee compensation to the new systems of public sector performance measures being developed.

Notes

* This research was funded in part by the Western Research Network on Education and Training (WRNET).

1. See, for example, Government of Canada (1996), Auditor General of British Columbia (1996), Government of Alberta (1996), and the Treasury Board Secretariat Internet site, http://www.tbs-sct.gc.ca/tb/key.html.

2. This portion of our paper draws on Armitage and Atkinson (1990), Diewert (1992a, 1992b, 1996), Diewert and Nakamura (1998), Lawrence, Houghton and George (1997), and Nakamura, Diewert, Lawrence and Russell (1998).

3. However, there are many early examples of the production and use of performance measures in public sector situations going back to the early 1900s, as documented in Diewert and Nakamura (1998).

4. See Armitage and Atkinson (1990:12) for field survey evidence on this issue.

5. For specifics on this see, for example, the references cited in the sources listed in the first footnote. For a useful short discussion, see also the Canadian Evaluation Society (1992).

6. See, for example, Armitage and Atkinson (1990).

References

Armitage, H. M. and A. A. Atkinson (1990) *The Choice of Productivity Measures in Organizations,* The Society of Management Accountants of Canada.

Auditor General of British Columbia (1996) *Enhancing Accountability for Performance: A Framework and an Implementation Plan: Second Joint Report.*

Canadian Evaluation Society (1992) *The Value of Evaluation.*

Church, A. Hamilton (1909) "Organisation by Production Factors," *The Engineering Magazine,* Vol. 38, pp. 184-194.

Diewert, W. E. (1992a) "The Measurement of Productivity," *Bulletin of Economic Research,* Vol. 44, pp. 163-198.

Diewert, W. E. (1992b)"Fisher Ideal Output, Input and Productivity Indexes Revisited," *Journal of Productivity Analysis,* Vol. 3, pp. 211-248.

Diewert, W. E. (1996) "The Measurement of Business Capital, Income and Performance," Working Paper (preliminary draft), Department of Economics, University of British Columbia.

Diewert, W. E. and A. O. Nakamura (1998) "Benchmarking and the Measurement of Best Practic Efficiency: An Electricity Generation Application," forthcoming in *Canadian Journal of Economics.*

Government of Alberta (1996) *Measuring Up: Second Annual Report of the Performance of the Government of Alberta.*

Government of Canada (1996) *Getting Government Right: Improving Results Measurement and Accountability: Annual Report to Parliament by the President of the Treasury Board.*

Griliches, Z. (1992) "Introduction," in Griliches (ed.) *Output Measurement in the Service Sectors* (Chicago: University of Chicago Press for the National Bureau of Economic Research Studies in Income and Wealth), pp. 1-22.

Lawrence, D., J. Houghton and A. George (1997) "International Comparisons of Australia's Infrastructure Performance," *Journal of Productivity Analysis.*

Nakamura, A., W.E. Diewert, P. Lawrence and A. Russell (1998) "Performance Measurement, Evaluation and Accountability," A PEER Group working paper.