22

PERFORMANCE MEASUREMENT PRINCIPLES AND TECHNIQUES

An Overview for Local Government

Harry P. Hatry

A simple definition of performance measurement for governments is the systematic assessment of how well services are being delivered to a community—both how efficiently and how effectively. The term "efficiency" refers to the relation of the amount of input required to the amount of output produced. "Effectiveness" refers to the impacts and quality of the service delivery, whether the service achieves its purpose, and how responsive it is to community needs. The increased knowledge about a government's service delivery system can improve the decision making of its elected officials and managers, and can improve their accountability to the public. This overview of performance measurement in local government covers four topics: the criteria for selecting measures, the various types of measures that should be considered, the data collection procedures for collecting data on the individual performance measures, and how targets for individual measures might be established.

Criteria for Selecting a Set of Performance Measures

Eight criteria are useful for selecting an appropriate set of performance measures. Both individual measures and the entire set proposed for collecting performance information should be assessed against the following criteria.

"Performance Measurement Principles and Techniques: An Overview for Local Government," *Public Productivity Review* 4 (December, 1980): 312–339.

Validity/Accuracy. Is the measure valid, does it measure what it should? Both the measure itself and the specific data collection procedures should be considered. Do they measure what they are supposed to accurately enough? It is not necessary for performance measurements to be extremely precise, but they should be reasonably accurate. A measurement might be quite accurate but the data might be invalid; it might not measure anything meaningful or might measure the wrong things.

Understandability. Will the measure be reasonably understandable by government officials? Generally, esoteric measures that are overly technical or that involve complex combinations of elements have limited use, at least for higher level officials. There are, however, circumstances where the more exotic measures can become understood and so become useful to public officials. For example, air pollution indices initially developed in technical terms are often then translated into the degree of hazard to the public. Thus they are useful both to public officials and citizens.

Timeliness. Can the information be gathered in time for it to be useful to public officials? Certain impacts of some programs take years before they can be detected, such as the long-term effects of education and other human services. Thus, for performance measurement purposes, by the time such information becomes available many aspects of the original program would have changed, and the data might be of little use.

Potential for Encouraging Perverse Behavior. Will the measure result in behavior that is contrary to the objectives of the organization? A measure such as "the number of tickets per police officer year" has a large potential for encouraging harassment of citizens.

Uniqueness. Does the measure reveal some important aspect of performance that no other measure does? This criterion is needed to keep down the number of overlapping, duplicative measurements. Multiple performance measures will inevitably be needed for most government activities, but too many measures can quickly cause both data collection and information overload.

Data Collection Costs. What does it cost to collect and analyze the data for the measure? This is probably the major constraint on performance measurement. Ultimately, the cost of performance measurement has to be justified by its value in improving decision making, reducing or avoiding service costs, improving service effectiveness, or improving service management. Unfortunately, at the onset of a performance measurement system, the value of such benefits can be, at best, only very roughly estimated.

Controllability. To what extent is the measure controllable by the agency whose performance is being measured? This has been one of the most controversial topics of debate over the utility of individual performance measures, especially effectiveness measures. Other things being equal, the more control government managers have over a measure, the more the agency can be held accountable for it. However, if an agency has only partial control over a measure, this should not exclude it from being used. Partial control is the usual situation in government. Operating agencies generally prefer to focus measurement on the more immediate workload outputs of their programs because of the direct link to their resources. They are less enthusiastic about effectiveness measures. These are often affected by external factors over which an agency's influence is questionable. Street cleanliness is not just a result of the performance of the public works department, but is also affected by citizen littering. Note, however, that even in situations of partial control, local governments can often exercise enough influence to have substantial effects, especially over the long run. In this example, government officials could pass and enforce anti-littering ordinances or conduct community campaigns aimed at litter prevention. Many effectiveness measures that are not controllable in the short run can be measures of accountability of top policy officials—including elected officials and chief executives. They provide an indication of how well these officials keep government policies and priorities responsive to the changing needs of the community.

Comprehensiveness. Does the set of measures cover all or most performance aspects of the organization's functions? If not, these omissions should be made explicit, or the measures should be designed to fill the gaps identified.

Types of Performance Measures

Measures can take a variety of shapes and forms, each measuring some different aspect about a program or service. Careful thought should be given to the purpose for a performance measure, and the appropriate type selected to suit that purpose. A variety of different types of measures are described here, with a brief discussion of their uses.

Cost Measures. These measures are simply indicators of dollars spent. It is, of course, legitimate to measure actual costs against budget costs. However, cost measures are normally not included under the label "performance measurement" and, consequently, are not covered in this overview. Cost does not by itself measure efficiency or effectiveness.

Workload-Accomplished Measures. Measures of the amount of workload that has been accomplished are the most commonly found measurements collected by operating agencies. They are frequently displayed in budget documents. By themselves, these output indicators are not performance measures as defined earlier. They say little if anything about the quality or effectiveness of an activity and, until they are related to the amount of resources expended to achieve that output, they say little about the efficiency with which the activities are provided. Of course, information on the amount of workload accomplished can be used for internal management purposes. But uses of these measures should be avoided when they merely encourage increasing the workload regardless of whether it is needed and affordable. Workload measures used as standards that are to be met or exceeded without an appropriate linkage to cost and without consideration of the work accomplished has considerable potential for encouraging perverse, make-work behavior.

Effectiveness/Quality Measures. Often the most difficult type of measure from which to obtain data, effectiveness measures are used to measure the impact and quality of a service, and whether it achieves its purpose. Each service has explicit or implicit objectives directed at its service clients. These objectives can serve as the basis for identifying effectiveness measures. Service clients may be the general public, some sub-set of the public (such as target neighborhoods, the business community, or the handicapped), or operating agencies (such as a public works department) served by internal support services (such as purchasing and data processing).

Often missed in setting objectives and developing measures is the potential for unintended, negative effects, for example, the pollution effects of some transportation systems. Effectiveness measurement, then, should include likely negative impacts of services or at least those impacts that can be identified ahead of time.

There is no complete agreement as to what effectiveness is, but some examples may suggest what governments should consider. Available goal and objective statements are a major source for deriving measures of effectiveness—if they are stated in terms of ends and not means. Examples of such measures include the following:* for recreation and library services—indicators of client satisfaction and use of facilities and services; for employment, health, and social services—measures of the extent of improvement in clients' employment and earnings, health and functioning; for

*For a comprehensive discussion of effectiveness measurement covering nine municipal services, including lists of measures for each, see *How Effective Are Your Community Service?* published by the Urban Institute and the International City Management Association.

street cleaning—an indicator of the cleanliness of streets; for fire—amount of spread after arrival of first fire fighting vehicle; and for support services such as purchasing, data processing, and vehicle maintenance—indicators of the timeliness and quality of the service provided.

In addition to receiving a government service, there are a number of "quality" characteristics that concern service clients. These include the timeliness, accessibility, courtesy, and equity with which each service is performed. Each of these qualities is also a candidate for effectiveness measurement. Each quality measure sheds a different light on the way the service is delivered.

Response time measures in particular have become popular; Sunnyvale, Milwaukee, and others use them for a variety of services. Response times can be measured for most programs, including the times to repair traffic lights, to respond to citizen complaints, to fill purchase orders, as well as the response time of police and fire calls. Note that these measure the timeliness of the service but not the result of the response.

Efficiency/Productivity Measures. Efficiency measures are generally defined as the relation of the amount of resources applied to a service or input to the amount of output. Ratios of output to input have typically been labeled measures of "productivity." The converse, ratios of input to output, are called "efficiency" or "unit-cost" measures. Both forms are equivalent. If five employee hours result in ten units of output, the unit cost is one-half employee hour per unit of output, and productivity is two units per hour. Units of input can be the amount of any resource. Typically, inputs are expressed in terms of the amount of employee time and in terms of dollars. Hours are not affected by inflation as are dollars. Dollars should be adjusted to "constant dollars," using a price index to give a better measure of changes in productivity. Units of energy may become used as an additional unit of input.

Units of output are measures of the amount of workload accomplished. Unfortunately, the readily available workload counts often say little about the real product of the activity. For example, the number of park acres maintained says little about how they look, the number of gallons of water treated does not indicate the quality of that water. Whenever possible, defective outputs should be identified and should not be counted as output. For example, in most cities a defective street repair, reported as an open crack or pothole and repaired again, would probably result in two output units being counted where only one was completed properly. It is perverse to encourage employees to work quickly by performing work badly. Because it is often difficult to track defective outputs, effectiveness measurement becomes an important "quality control" complement to efficiency measurement. We will return to this problem with some suggestions for alleviating it.

Actual Unit-Cost to Workload Standard Ratios. These are a special form of efficiency measurement. A "work standard," or standard amount of time that the particular activity should take, is determined. Work standards have been used for many years in the private sector, usually developed by industrial engineers. Local governments have used them in some instances, and recently they have become more widespread in all levels of government.

For a particular performance period, an agency would determine the amount of output and the total actual time applied to produce that output. The average actual time per unit would then be calculated and compared to the work standard time per unit. The resulting ratio is a comparison of the actual work accomplishment to a target based on the work standard. For example, a work standard might be that a crew of three should fill one pothole in one hour. Actual performance for the period might average one per forty-five minutes, for a production of 133 percent of the standard (that is, four potholes for every three based on the work standard).

Preferably, the standard should be an engineered work standard, calculated through some version of time and motion studies that provide a standard time systematically and reliably derived. Unfortunately, in some instances state and local governments have derived "standards" merely by having personnel keep track of their own time (for a few days), and using whatever job procedures the employees happen to choose. The existing average times are then used as the work standard "should take" times. A much better approach is first to examine the procedures to identify good practices, and then to determine the times to accomplish these good practice procedures by a systematic, objective method, such as time and motion studies. Also, work standards need to be periodically reviewed to ensure they keep up to date with current methods, technology, and requirements.

Work standards are useful only for some government activities—those for which a specific procedure can be established and a standard product identified. For example, standards for welfare eligibility determinations should be possible but may not be appropriate for client counseling; standards for the fingerprinting activity can be readily developed, but standard times for crime investigations would be much more difficult to determine. Work standards have been used for such relatively repetitive tasks as data processing, key punching, clerical work, street repairs, vehicle maintenance, park and building maintenance, and various types of inspections.

Efficiency Measures and Effectiveness Quality. As discussed earlier, most ratios of the amount of workload accomplished to the amount of resources expended say little about the quality or effectiveness of the service. Performance measurement should attempt to close this gap. If the information from performance measurement is to be used for purposes of substantial concern to employees (such as changing their budgets, appraising their performance, or as a basis for monetary incentives), the absence of quality considerations can encourage employees to perverse behavior such as focusing on the immediate workload outputs at the expense of quality. There are three approaches that can be used to alleviate this problem.

First, in some cases the output units can actually be designed as an effectiveness measure, rather than just a measure of the work completed. For employment, mental health, and social services, rather than using the number of clients served, the number of clients actually helped could be used to produce the efficiency measure: "the number of clients helped per dollar." This, of course, requires the use of reliable procedures for assessing whether clients had been helped. For parks and recreation, a government could use the "cost per household that used a recreational facility during the year and also expressed overall satisfaction with their experiences." Data for this measure could be obtained through the use of client surveys, as discussed later.

The workload measure could still be used as the output units for the ratio, but formal quality controls could be built in to ensure that only outputs passing pre-specified quality control tests would be included. Defective units would not be counted. For example, street repairs that did not pass an inspector's check, or that deteriorated before some pre-specified time, would be excluded. For its services for children, New York City has begun using the measure "percent of institutionalized placements found inappropriate after documented review." This percent could be multiplied by the total number of placements to yield a better output count for efficiency measurement. Sunnyvale, California, includes in output counts only the number of acres of park vegetation "evaluated to healthy." This approach requires a more formal quality control process in local government than is often currently the case. Private industry often has formal quality control procedures; governments much less so.

Finally, if neither of these two approaches seem feasible, then data on effectiveness-quality measures should be presented along with the input-output ratios to display how effectiveness has been changing at the same time that efficiency has been changing. If, for example, it is found that unit-costs have improved but effectiveness measurements examined during the period have worsened, then perhaps efficiency has been improving at the expense of effectiveness. Officials could then question whether there was a true efficiency improvement and take appropriate action to get effectiveness and efficiency back in line.

Resource-Utilization Measures. The term efficiency measure sometimes has been used to cover another type of measure not expressed as input-output ratios: measures of "resource utilization." These measures are often

expressed in terms of the proportion of the available time that equipment or personnel were actually providing, or available for, service. They include measures of the amount of downtime for equipment and personnel. For personnel, the measure sought is usually the percent of time that was spent on "productivity" activities and not, for example, spent waiting for material or equipment such as excessive court waiting time for police. Other examples include the average number of crews on duty, "unused capacity" measures such as load factors for public transit vehicles, and the amount of water pumped into a local system that is not billed because of leakage, etc. These measures can be used to highlight operating problems or poor use of resources.

Resource utilization measures primarily indicate the potential for added resource value. They do not reflect the added amount of output or output-per-unit-of-input that is lost or gained. For example, the time personnel spend waiting for material and equipment might be reduced, but that does not by itself guarantee that the freed time will be used to produce added output. Thus, these measures are not direct measures of efficiency but can be used to point out that potential for greater efficiency. They are probably best labeled as "quasi-efficiency" measures.

Productivity Indices. These can be readily calculated from any of the previous performance measures. An index shows the relative change of a measure from a base period to another time period. The actual value of whatever performance measure is used (e.g. the number of repairs made per unit of input) for the base period would be given the value of 100. In succeeding years the index would be the ratio of the performance measured for that year as compared to the base year. Thus, if a future year's performance is seven percent higher than that of the base period, then the index for that year would be 107.

Indices for individual activities can be combined to give an overall index even with quite different activities. This is done by some form of weighting by the amount of input. For example, if one activity required twenty-five percent of the total number of employee-years whereas a second activity required fifty percent of the employee years, then the index for the second activity would be weighted twice that of the first in calculating the overall productivity index.

Productivity indices are commonly used for national assessments of private sector productivity. These have rarely been used in the public sector. The federal government has begun to attempt to measure annually the productivity of federal employees. The Bureau of Labor Statistics calculates productivity indices for each federal agency and combines these to produce an overall productivity index for federal workers covered by the productivity measurements.

Pseudo-Measures. A number of measures have been used on occasion as performance measures, but are so ambiguous and so potentially misleading that they should not be considered as performance measures. Nevertheless, they are likely to be of interest to public officials—but they do not measure efficiency or effectiveness since the amount of product, the quality, or the value obtained is not indicated.

Cost per capita or cost per dollar of assessment value, that is, the total cost for a service divided by the total population in the jurisdiction or the total assessed value, is not a measure of efficiency. Cost per capita is really an indicator of the amount of resources expended. For most services the population of the jurisdiction tells little about the quantity or quality of the output. Even where a service is directed at all citizens, cost per capita says little about the product. For example, cost per capita for police protection does not indicate anything about what is obtained for those dollars. Multi-year trends of total service costs per capita (or per dollar of assessed value) may be useful indicators of the fiscal behavior of a community. Since cost per capita measures say nothing about the output of the service, however, they should not be used to represent the efficiency or effectiveness of services.

Other examples of pseudo-measures include the number of books per capita and staff-client ratios. These measures provide no information about the product being delivered. They are primarily measures of the amount of resources applied and say nothing about the efficiency of the service. If evidence was developed that the number of books per capita (or staff-client ratios) is highly related to client satisfaction (or client progress), then the measure could be used as a proxy for service quality (though not efficiency).

Cost-Benefit Ratios. These are sometimes proposed, especially by economists, as being the ultimate in performance measures. Traditionally, cost-benefit ratios are ratios in which the output side of the ratio has somehow been converted into a dollar value. A ratio of 1 to 2 indicates that for every dollar of cost, two dollars of benefits were achieved. As long as the value of the benefits exceed the value of the cost, presumably a program is worth its cost. Programs with better cost-benefit ratios would be more desirable than ones with worse ratios.

In actual practice it is usually extremely difficult to determine a meaningful dollar value for many if not most government service outputs. How does one determine the value of additional cleanliness in the neighborhood, additional recreation satisfaction, or better quality water? Methods have been developed for some valuations such as estimates of the value of travel time and of recreation, but such procedures tend to be partly arbitrary and often are based on questionable assumptions. The use of cost-benefit ratios is not likely to be a very viable procedure in the foreseeable future. Such ratios can be useful for specific selected in-depth studies, but not for routine performance measurement.

Comprehensive Performance Measurement. Thus far, we have been talking about individual types of measures. Any single government, and any single agency, is likely to find it necessary to use a variety of measures that will form a comprehensive set of measures.

One approach is called Total Performance Management (TPM). This includes a combined set of information including workload oriented unit-cost measures, quality measures (particularly client satisfaction), and employee attitude measurement. In TPM employee attitude surveys are used to identify potential obstacles or problems that need to be alleviated to accomplish improved productivity, rather than to measure employee satisfaction. A number of local governments have tried TPM, including the cities of Sunnyvale, San Diego, and Long Beach, California, and Cincinnati, Ohio. Some governments may want to include employee attitudes as part of an overall measurement package, particularly where employee problems are believed to be an important productivity problem in the jurisdiction. Employee attitude surveys, however, are not considered part of performance measurement for the purposes of this paper.

Data Collection Procedures

For convenience, we will group data collection procedures into five categories: 1) the use of data in government records; 2) trained observer ratings; 3) a special measurement approach—work standards; 4) citizen/client surveys; and 5) miscellaneous—a catchall category to cover any procedure that does not seem to fit into one of the first three categories and to provide leeway for individual jurisdictions' ingenuity. Each data source can yield meaningful information. Each, however, is subject to large errors if not done properly.

Government Records. Use of existing record data, all other things being equal, is the most attractive source of data collection for performance measures. It is already collected and thus by definition requires little added cost. Some examples of existing data routinely collected in many communities, and that are directly applicable to performance measurement include: counts of workload completed (for use in efficiency measures) such as the number of repairs, number of records processed, number of gallons of water treated, number of tons of garbage collected; cost data for efficiency measures; service quality measures such as the number of complaints received; number of traffic accidents/casualties; number of reported crimes and number of arrests; incidence of communicable diseases; and response time to fire and police calls.

In many cases, however, existing data, or the way it is collected or calculated, will have to be at least partly modified for use in performance measurement. For some information, particularly on the effectiveness and quality of services, existing records are likely to be inadequate and will need revision. Cost data in many if not most governments are not maintained in such a way as to permit cost or employee-hours breakouts by specific activities associated with specific workload measures (e.g. to distinguish commercial collection and residential collection efficiency). Some local governments do not collect the necessary workload, effectiveness, or other data required for efficiency measures (e.g. the weight of solid waste collected is needed but not always measured to determine the cost-per-ton collected.) Though incoming complaints are sometimes tabulated, data on counts of complaints that are actually resolved (one way or another) are much less frequently available.

Thus, even where government records seem an appropriate data source, often modifications to the existing data collection procedures are likely to be needed. Following are some examples where government records can be made more useful. For determining the "number of arrests per police-employee year" rather than using the number of arrests, the "number of arrests *that survive the first judicial screening*" can be substituted to provide an estimate of arrests that are "productive." This would also reduce the temptation of police employees to make excessive arrests. However, to collect this data requires obtaining information from the judicial system as to the disposition of the arrest at the first judicial screening. Tests of such procedures have been conducted in the District of Columbia, St. Petersburg, and Nashville. Secondly, client complaints can be tallied by agency and by subject. Finally, response times can be calculated for numerous activities such as responding to client complaints, making repairs of streets and signal lights, repairing automotive vehicles, and filling requests for services. Data collection procedures would have to be added for accurately recording the time of the initial request and when it was satisfied.

The accuracy of government record information varies considerably. The accuracy of counts of workload accomplished depends on the particular procedures used to make those estimates. If, for example, scales are used to measure the amount of waste collected, the data should be fairly accurate, assuming that the scales are checked periodically; if, however, estimates of weight are made without scales, then the estimates may be unreliable. Do not assume that because data is obtained from records that the data is accurate. Following are some examples of common data accuracy problems. First, the number of reported crimes, widely used as a measure of crime prevention success, is subject to errors from nonreporting by citizens, by as much as thirty to fifty percent for some crimes. There are also problems in defining crimes consistently; police officers have considerable discretion in reporting and categorizing incidents. Secondly, in various human services such as community mental health and social services programs, counts of the number of cases closed can be misleading as an output measure—whether used for measuring efficiency or effectiveness. Problems abound in defining what a "successful" case closure is and even determining when a case is closed. Thirdly, data on the amount of time and cost for employees by specific activities will be accurate to the extent accurate procedures are used, and will depend on the motivation of employees keeping track of their time. This is not a problem when a worker has only one activity during a performance period, but it is a problem when it is necessary to allocate a person's time among more than one activity. Finally, changes in record-keeping procedures can affect the accuracy of the data (e.g. complaint counts could be affected by a change in the complaint telephone number).

Trained Observer Techniques. This is a procedure in which an observer is trained to rate characteristics of some service, primarily to assess a physical characteristic of the results of the service. Some examples of trained observer uses are: to make ratings of street cleanliness (e.g. as used in New York City's "Operation Scorecard" and in Charlotte, North Carolina); to evaluate the rideability/roughness of streets in the assessment of street maintenance activities (e.g. as in Phoenix, Dallas, and Nashville); to assess park and playground maintenance results (Honolulu, New York City); to assess the level of hazard at solid waste disposal sites (Nashville); to assess exterior housing conditions (Dallas); and to develop work standards (numerous jurisdictions).

A form of trained observer rating may be the most feasible approach for providing quality tests for some efficiency measures, such as inspecting repairs to ensure that they meet basic quality standards, or reviewing eligibility decisions to determine if they were reasonable and should be included in the output counts.

The accuracy of trained observer measurements depends on a number of elements. If the procedures are too loose or too haphazard, they may be quite inaccurate. Local governments should not treat such procedures casually. Steps for making ratings reliable include the following. 1) Provide some type of anchored scale specific as to what each grade of the scale represents, so that different observers, seeing a range of conditions at different points in time, would generally provide the same ratings. For example, the rating categories might be defined by pre-selected photographs representing each grade of street cleanliness or each grade of park maintenance quality. Written rating descriptors can also be devised to describe the individual grades. The descriptions for each grade might include quantitative descriptions such as the height of grass or amount of litter. 2) Test the procedure before full use. Have different observers rate a number of different conditions to see whether their ratings are sufficiently close to be reliable. 3) Choose raters who are independent of the unit whose work is being rated to avoid the potential for, or appearance of, bias. 4) Periodically check each observer's ratings for possible deterioration in the ratings' quality. Experience shows that ratings tend to "telescope" after a time, so that they are squeezed together toward the center of the scale. 5) Train new raters in the procedures and retrain current observers when periodic checks indicate rating problems.

Trained observer procedures require time to make the ratings. Thus, they are likely to involve extra costs unless there are persons available with time to do the ratings. In some instances, they may be personnel for whom undertaking such ratings would be a reasonable activity under their current job descriptions. For example, cities that have used solid waste collection inspectors, such as the District of Columbia, expanded the inspectors' jobs to cover cleanliness ratings. Foremen and other supervisors probably are already responsible for work quality, and they could be made responsible for such ratings. However, where the ratings are to be used to assess and compare the performance of each supervisor's work crew, rather than individual workers within the crew, supervisors might not be sufficiently unbiased to make reliable ratings.

There are two ways to reduce the costs of trained observers. First, include the ratings of a sample of items rather than all (e.g. only a sample of items being repaired, or of eligibility determinations, might be reviewed to estimate the number of outputs produced that meet quality standards). Second, have inspectors do combined ratings of different activities (e.g. only a sample of city streets might be rated, and rated "simultaneously" on both cleanliness and rideability).

The size of the sample has a substantial effect on cost. The sample size needed depends on the precision desired for the estimate and on the number of different sub-groups for which comparisons are wanted. For example, if a city wants to compare eight neighborhood areas, adequate sample sizes will be needed for each of the eight. If sampling is used, the sampling procedures should be sound; some form of random sampling should be used to produce a reasonable representative sample.

Trained observer time requirements were estimated by the Urban Institute on the basis of 1975 solid waste collection ratings in Nashville and St. Petersburg. One-half employee year was estimated to be needed for twice per year samples of streets with 250,000–750,000 people—about half this for cities under 250,000. Actual out-of-pocket costs would depend on the extent to which existing personnel are used. In addition, some resources

are needed before full implementation of the rating procedures to ensure that an adequately reliable procedure is designed and that the personnel involved with making the ratings are adequately trained in the procedures.

Most rating procedures are reasonably easy to learn. The trained observer role can often be undertaken by lower-skilled, inexpensive personnel, such as clerical workers or college students. For more complex processes such as eligibility determinations and quality checks of road repairs, this would not necessarily be so.

Special Measurement Approach—Work Standards. The development and use of work standards requires both agency record information and a special form of the trained observer approach. Trained observers are used to develop the work standards. They need to use such techniques as time-and-motion studies, process flow charting, and work sampling to analyze the methods used by public employees to perform an activity. The observers or analysts then determine a work standard in the form of a "should take" time (in employee hours) to produce a unit of service. The work standard should be based on the best methods feasible considering the equipment and staff available. Generally, when a community goes to the trouble of such a study, it will try to implement the improvements determined by its methods analysts.

Industrial engineers, methods analysts, or other specially trained observers are needed to develop the work standards. But once the standards are developed, the analysts generally design a routine reporting system or modify the existing one so that employees' output and work time can be regularly reported. The new or modified reporting system essentially creates new/modified government records, from which data is periodically collected to determine the level of performance or percent of standard achieved.

When work standards are used, the agency should use the specially skilled trained observers periodically to check and update the standards. This is required if, for example, new methods or technology have been introduced. For repetitive activities, those that involve standard procedures and identifiable products, the use of these approaches to set time standards can be quite accurate. A principal source of error comes from developing standards for word procedures that are not the procedures actually used by the employees doing the job. After time standards are established, the accuracy of the actual time reported per unit depends on the particular reporting procedures, especially the extent to which there is potential for misreporting, including the lack of skill or motivation of employees to report their times accurately. Problems can also occur to the extent employees have the option of allocating their time to activities covered by work standards or to activities not covered.

Developing engineered work standards can be costly, especially because of the time required of skilled personnel or consultants. But the payoff in efficiency savings after an intensive study can be high, especially for an activity that had not been examined analytically in many years. The City of San Diego, for example, reported a 9 to 1 ratio of savings to program costs in the first five years of its engineered work standards programs.

Building periodic reporting into employees' jobs is generally not very costly, as usually no new personnel are required, but it can be a source of annoyance to the employees.

Citizen/Client Surveys. Surveys of the general citizenry, or only those citizens who have been clients of a particular service or facility, can be used to obtain ratings of various aspects of service quality, and to obtain certain factual data on service effectiveness. Annual citizen surveys have been used regularly as part of performance measures by such cities as Dallas, Dayton, Kansas City, and St. Petersburg.

To obtain factual data, citizen surveys can be used for the following. They can obtain data on victimization of citizens to get better counts of crime rates than is obtainable from reported crimes. They can obtain participation rates (i.e. the percent of different persons or households using a service) for such services as parks and recreation, libraries, and public transit. Data from fare boxes and site counts tells how many trips or visits were made, but they do not indicate how many different persons made trips or visits. They can provide estimates of the number of rat sightings by respondents of various neighborhoods to obtain an estimate of the incidence of rat populations. For clients of human services programs such as employment, training, mental health, and social services, they can obtain information from clients as to their post-service employment duration, earning levels, and the extent of improvement in mental distress or social functioning (to estimate the percentage of clients helped).

Surveys can also be used to obtain citizen or client ratings of specific characteristics of the following services: satisfaction with recreational opportunities and facilities; availability and quality of library services; perception of the cleanliness of their neighborhood; feeling of security from crime (often an important performance measure for crime control activities); adequacy and quality of transportation services; and the odor, taste, appearance, and pressure of the water they receive. Among the quality attributes these ratings can obtain are the timeliness of the services received, their accessibility (location and hours of operation), and the courtesy and dignity of treatment by government employees. These qualities are also relevant to internal services such as purchasing, data processing, payroll, and personnel. The "clients" of internal services are employees in other government departments.

Citizen survey ratings are influenced by the expectations of citizens as well as by actual conditions. Changes in citizen ratings could be due to changes in either or both. In general citizen surveys, respondents not using specific services can also be asked to identify reasons for their non-use. However, caution should always be taken when interpreting these reasons, which are not always penetrating. Responses to questions for reasons of non-use can be tallied to provide data for measures as "the proportion of citizens that do not use basic municipal services for reasons that are controllable by the government (such as accessibility, hours of operations, and safety)."

In client surveys, as well as in general citizen surveys, when a respondent expresses dissatisfaction with a particular service characteristic, the respondent can and should be asked why. This will provide diagnostic information and information on the frequency of reasons controllable by the government.

Surveys for performance measurement should be distinguished from surveys for citizen opinions on various policy issues—those that ask what the government "should do" in the future. Many, if not the majority of, current citizen surveys used in the past by local governments have tended to emphasize the latter type of question. Survey procedures then are used as a quasi-referendum. There are numerous problems with these surveys, such as leading questions and the possibility of tying the hands of elected officials. The inclusion of a few questions for these other purposes can increase the willingness of local officials to undertake the survey. However, these questions are particularly subject to bias. Special care should be used in developing such questions to minimize this.

The accuracy of information obtained from citizen/user surveys depends on a number of factors. Following are some suggestions for maximizing accuracy. First, avoid sampling errors. Give considerable attention to the precision needs; don't overdo precision, but do provide adequate sample size for each subgroup (e.g. about 100 per subgroup). Second, use a random sample. Third, make sure the list from which the sample is drawn covers all groups of interest. Accuracy depends on the completeness of the population list. Fourth, proper interviewing techniques should be used and provisions made for an adequate number of call-backs. Fifth, provide for adequate training and monitoring of interviewers. Inadequately trained and monitored interviewers can lead to errors. For example, interviewers can influence the respondent to answer in some particular way. Sixth, pretest the questionnaire on at least ten to twenty persons from different educational levels and ethnic groups. Accuracy depends on the quality of the questionnaire. It is important that the wording be reasonably clear, unambiguous, and unbiased. Finally, in performance measurement surveys, clients should be asked to respond about their own personal recent experience.

Both costs and efficiency are affected by the mode of interviewing. In-person, telephone, mail surveys, and combinations thereof, are options. Mail surveys, though the least expensive, may be inadequate because of low response rates (about ten percent return rates are common). Although they can be inexpensively sent to a massive number of households, more representative (and so more accurate) results can be obtained from a small number of households, such as 500 or less, if they have been randomly chosen and a high response rate (about seventy to eighty percent) is achieved. Mail surveys also have the problem that literacy is required. And, the questionnaire has to be quite short. Mail surveys become more of a candidate if follow-up mailings are used for non-respondents and if they are supplemented by telephone or in-person interviews for those that do not respond to the mailings. These steps are needed to get response rates up to reasonable levels, about fifty to sixty percent.

In-person interviews have long been the favorite for many professional survey organizations because of the belief that these are likely to be the most accurate, provide high response rates, and permit the longest interviews. Because of the travel time and costs involved, however, these are by far the most expensive. Recently, telephone interviews have become popular, especially those that use random digit dialing to ease the problem of obtaining a representative sample. Telephone interviewing has the advantages of lower cost, permitting many more call-backs by re-dialing at different times, easier access (in some locations it is difficult to get interviewers into respondents' homes), and controllability (where interviewing is done at one central location, a supervisor can monitor the interview). In cases where the sample is not representative due to the lack of phones in enough households, telephone interviews can still be used for most of the population. These calls can be supplemented by in-person or mail interviews.

The cost for a survey depends on the total sample size and mode of interviewing. St. Petersburg, Nashville, and Dayton have used citizen surveys that have cost ten dollars or less per respondent for sample sizes from 600 to 1,000. However, costs up to twenty dollars per respondent can occur, especially for a first-time survey. This assumes an interview of about thirty minutes in duration. Heavy reliance on in-person surveying can easily double or triple the cost. Mail surveys would cost less, but would have to be considerably shorter. Telephone surveys should average no longer than about forty-five minutes and preferably less. These cost estimates also assume that the product of the survey is primarily a series of numerical tabulations without any extensive analysis of the data. The estimates do not include the start-up costs such as the time required in developing a questionnaire.

For surveys of the total population, a considerable savings in the cost per household per question can be achieved by covering several services in the same survey. Though this lengthens the interview, some questions are applicable to all surveys (e.g. demographic questions) and thus have to be asked only once. More importantly, the major expenses of locating clients and certain other costs are affected to only a small extent by the number of questions.

In some instances it may be convenient to use a much shorter interview to obtain data for performance measurements of only one service. For example, short, five to ten minute telephone surveys have been conducted about solid waste collection. Such interviews should be possible for a cost of five dollars per interview or less.

Client surveys normally cover just one service, such as surveys of users of buses, libraries, and recreation programs. They also cover complainants about any service—to determine client satisfaction with the handling of their complaints. Client surveys often can be done less expensively than general citizen surveys. The questionnaires can be shorter and often can be handed out to users as they enter the government facility and collected as they leave. They can even be given in-person interviews at the facility. This ready access means that the expensive tasks of locating persons in the sample is greatly eased. For some other services, the number of clients is likely to be smaller and their telephone numbers or addresses conveniently available, thus reducing location problems.

The use of citizen surveys has become more common in recent years. Cities such as St. Petersburg, Kansas City, Dallas, and Dayton, as well as small locations such as Randolph Township, New Jersey, have used them periodically for performance information. Many cities have done at least one survey for special studies.

At least a small amount of expert assistance will be needed by most jurisdictions to avoid major problems. Preferably, the survey work should be contracted out to an experienced survey firm to avoid administrative and quality control difficulties. This requires dollar outlays. Dayton believes that it has been able to keep costs below ten dollars per respondent by using competitive bidding procedures.

Special Data Collection Procedures. This is a catchall category to pick up miscellaneous approaches and leave room for a government's own ingenuity. There are a number of performance measures, especially measures of effectiveness, that require special measuring equipment to obtain data. Air, water, and noise pollution testing procedures fit into this category. For many jurisdictions, these tests are becoming sufficiently common that they could fall into the category of government record data.

Another example of the use of special equipment is the use of the "roughometer." It is dragged behind a vehicle and gives physical measurements of road roughness (a number of state transportation agencies are using such procedures). Since the physical measurements obtained from technical equipment are somewhat esoteric, they preferably should be converted into more understandable categories. For the roughometer, the technical readings can be translated into overall street rideability. Rideability categories can be derived, for example, by using a group of blindfolded citizens who judge the riding comfort of each street as they ride on a sample of streets. Roughometer readings for those same streets are then related to the judges' ratings to identify ranges of roughometer measurements that are associated with particular citizen personal comfort ratings. Then for subsequent roughometer measurements, these can be converted to rider-comfort categories. A similar procedure can be used for other physical measurements.

There are likely to be many ingenious ways to get at what otherwise may appear to be difficult to measure aspects of government performance.

Consideration of Workload Difficulty and Equity

Two additional topics of importance to government performance measurement are often neglected: the need to consider somehow the difficulty of the incoming workload and the measurement of the equity of services. Each of these are briefly discussed below.

Measuring Workload Difficulty. Agency performance, whether efficiency or effectiveness, is clearly affected by the difficulty of the incoming workload. This would not be a problem if the workload was homogeneous, but in most services this is not so. For example, crimes vary considerably as to the evidence presented to investigators, street or automotive repair problems will differ considerably as to the difficulty of making repairs, and purchase requests from agencies will differ considerably in complexity. These differences greatly affect the time and ability of government agencies to perform their role in providing service.

This would still not be a significant measurement problem if the mix of workload is stable overtime and, in a comparison among units, if each unit had about the same workload mix. Unfortunately, however, significant differences in workload mix are likely to be the rule rather than the exception.

Performance measurement should explicitly consider workload mix. Preferably, performance should be measured for each category of workload difficulty. For example, criminal apprehension rates could be measured for various categories of case difficulty. The categories might be based on the amount and type of evidence provided to the investigator. The number of hours required to make repairs could be tallied for each category of street or automotive repair job (based on its apparent difficulty). This requires that specific definitions be developed for the various categories of difficulty to permit reliable categorization of incoming jobs by difficulty—with subsequent tracking of the outcomes and resources expended for jobs that fell into each category.

If such procedures are too difficult to develop, at least the mix of the incoming workload with respect to degree of difficulty should be identified. This will provide public officials with information to help them interpret whether measured changes in performance were due to changes in the mix of the incoming workload. This is a measurement subject that has only recently begun to receive attention. As yet, there is not much experience in handling this issue.

Measurement of Equity. The degree of equity of service delivery and community conditions among neighborhoods, or among other types of citizen groups, should be an important part of performance measurement. Data collection procedures such as citizen surveys and trained observer ratings lend themselves readily to tabulation by client groups. Ratings for street cleanliness and rideability, for public facility cleanliness and physical condition, and for security from crime and victimization, should all be measured for each major neighborhood. The resulting comparisons can provide valuable information as to which neighborhoods or client groups have greater needs than others.

Indicators of equity can be gathered for each major type of data collection procedure. Citizen survey responses can be tabulated for each major neighborhood area (e.g. those that represent major socioeconomic groupings) or at least areas with substantial commonality of needs and problems. Survey findings should also be broken down by major respondent characteristics such as age, sex, and income class. Trained observer procedures often lend themselves to grouping by geographical area within the jurisdiction. Finally, government records in some cases permit breakouts by client group characteristics. Where the procedures lend themselves to this, such breakouts can be made and may indicate disparity in equity of services to various client groups. Savannah, Georgia, for example, has used data from all of the above sources as part of its "Responsive Public Services Program." The purpose is to make service improvements and new public facility selection more responsive to each neighborhood's differing needs.

Determining Whether Performance Is Good or Bad

A major issue in performance measurement is how to assess whether the measured level of performance is good or bad. Information is desired to permit government managers and other local officials to compare actual performance to benchmarks. Several ways to make comparisons of performance are described below.

1. Compare actual performance against performance standards. Unfortunately, few valid standards exist today against which actual performance can be compared. Ideally, and

perhaps eventually, such standards may be constructed for many performance measures. The most common standard for performance measurement are "work standards." The comparison of the actual time it takes to produce a certain amount of output can be compared against a systematically obtained "should take" time. Long Beach, California, included questions in its citizen survey to help develop standards for certain performance measures such as police response times for non-emergency calls (by asking respondents what they felt would be a reasonable wait for a police officer to arrive).
2. Compare current performance measurements to performance in previous time periods. Previous performance can generally be used, and in practice is so used, to help assess current performance.
3. Compare the performance for different units where a service is being delivered by more than one unit doing essentially the same activities. Furthermore, the performance of any one group can be compared to that of the average of all units.
4. Compare outcomes for various client groups within the jurisdiction. For some measures it is appropriate to compare the performance levels for different types of clients. For example, wherever citizen or client ratings of services are provided, these ratings could be grouped by such characteristics as neighborhood of residence, age, sex, race, and income group. Grouping performance measurements by major geographical neighborhoods in a jurisdiction seems to be particularly informative to local officials. Average performance for the whole jurisdiction, or the area or group for whom performance was highest, could be used as a benchmark for each individual group.
5. Compare performance to that of other jurisdictions. Officials often want to be able to compare their own performance with those of other, similar, jurisdictions. A major problem here is the lack of similar performance measurement by local governments. Another problem lies in the hidden assumptions that are made when saying that two jurisdictions are similar. Population and density similarities do not necessarily mean that the type and difficulty of incoming workload as well as other factors that affect measurement are the same. Also, data collection procedures are likely to differ, and special factors unique to individual jurisdictions abound. However, when local governments are using approximately the same procedures such comparisons can be made. Even now some comparisons

are possible on crime and clearance rates, and citizen ratings of certain services can be made among a few governments that have been using similar questionnaires.
6. Compare performance to that of the private sector. For some government activities, similar activities are carried out by the private sector. For example, commercial automotive repair times are likely to be comparable to government repair times when the vehicles involved are similar. Other examples include private solid waste collection, data processing, and purchasing activities. In some instances, a government may itself use both public and private delivery such as solid waste collection and thus provide an opportunity for comparison.
7. Compare performance against pre-set and planned targets. If a jurisdiction sets targets for the forthcoming year on its performance measures, then at the end of that year actual performance can be compared to those targets. Local governments already set targets that use management by objectives. The key question is how the targets should be set. Some ways are by using the comparisons already mentioned, that is, by basing the targets in part on prior years' performance, on work standards, on other areas of the community or other organizational units, or on the private sector. A more difficult approach is to undertake an in-depth analysis of the resources, program characteristics, past performance, and conditions likely to exist over the next year, and thus analytically derive a target. Even targets not analytically derived, however, when set a little ahead of previous performance, can serve to challenge service managers to improve past performance.

A caution here about benchmarks. Benchmarks should explicitly consider the incoming workload mix. As discussed earlier, the difficulty of the incoming workload may vary—between private and public sector workload, from one government to another, from one year to another, and so on. Preferably, performance benchmarks would be set for each workload-difficulty category. Comparisons would then be made of the performance in each category.

Conclusion

There are a variety of measures and data collection procedures available. Procedures that provide substantial information on effectiveness and quality of service, whether used for effectiveness measurement alone or as part of efficiency measurement, will often require additional effort and cost

BOX 22.1 Difficulties in Measuring Human Services Performance

Paul Epstein

Measuring human services provides difficult challenges. Human services, while measurable, often present "moving targets," making performance difficult to define and data difficult to interpret.

While measuring activities of many human services is not difficult, measuring results or "outcomes" of the activities is often problematic, which makes defining and measuring the performance of many human services less certain than that of "hard" services. While many activities (*e.g.*, caseworker visits to clients or client visit to a clinic) can be counted and their costs calculated, the results of the activities are often difficult to ascertain (*e.g.*, whether a client is physically or emotionally functioning better).

The difficulties of measuring human services can be used as an excuse *not* to make the effort to measure service effectiveness and efficiency. But these same difficulties point to the importance of measuring human service performance and why human service managers should commit the resources required to do it well. Precisely *because* the performance of human services is difficult to define and measure, these services run the risk of not having clear goals and objectives for managers and staff to work toward. Even if an agency has an eloquent statement of its broad human service mission, if it is not backed up by more narrowly defined measurable objectives and measurement of changing trends in client needs, there is a danger of confusion and organizational drift as different managers and staff define and apply the mission differently in their work. Measurement can help keep the staff of a human services agency focused on common objectives which are relevant to changing client needs.

Some suggestions are:

- Clearly identify all the clients (or client groups) of each service provided by the human service agency.

(continues)

over that which governments are currently taking. Furthermore, in-house undertaking of trained observer ratings, citizen or client surveys, and data analysis probably requires skills sometimes not found in local governments, thus requiring that these new skills be obtained.

Many of the procedures such as citizen surveys and trained observer ratings fortunately can be cut back in size and cost for small cities. Smaller sample sizes generally can be used in small communities; some small communities have undertaken surveys of perhaps 300 citizens. Often they have used local volunteers such as members of the League of Women Voters or local students to help with interviewing and data processing.

The data collection procedures discussed will not by themselves identify why the performance levels are as they are. Measured reductions in performance do not necessarily mean that the agency was to blame for the reduc-

BOX 22.1 (continued)

- Define service effectiveness from clients' points of view.
- Look for key aspects of human services that are easy to measure.

  Some examples are:

  Service response times and other process items;

  Error rates and administrative burdens, especially for services involving eligibility determination; and

  Revenue, such as revenue generated from special efforts (*e.g.*, efforts to qualify more clients for Medicaid or Medicare eligibility) or from "outside sources" (*e.g.*, special purpose grants) to supplement institutional or governmental funds.
- Examine efficiency and other measures in the context of effectiveness.
- Set goals and objectives which balance various aspects of performance.
- Periodically reevaluate and review performance measures, goals, and objectives.
- Always remember the primacy of the clients who need human services.

(Source: Paul Epstein, "Difficulties in Measuring Human Services Performance," in: "Difficulties in Measuring Human Services Performance: Keeping Your Eye On the Moving Targets," *Journal of Health and Human Service Administration* 14 (Summer 1991): 27–43.)

tion. The classic example is that an increase in crime may not mean a city's crime prevention effort has worsened but could be due to worsened economic conditions. Similarly, an increase in measured performance is not necessarily due to better agency performance. To determine the extent to which an agency is responsible for observed changes in performance measures, more in-depth program evaluation and analysis are needed. A major purpose of agency performance measurement is to identify which activities require closer scrutiny. This is extremely valuable information since resources required for in-depth evaluations are always limited.

Finally, a word of caution. There is a great temptation to local governments to cut corners on performance measurement procedures and not to worry about the accuracy and validity of the information. Thus far, because of the limited use made of performance information, this may have been justified. But as more important use of performance measurement is made (such as to guide major program and policy choices, provide performance incentives to employees, as well as for performance contracting), then much more care will be needed. Good information does not come for nothing. You get what you pay for. Sound data collection practices and quality control of the data will be required and should be provided.