July 2014

Transforming Performance Measurement for the 21st Century

Harry P. Hatry

The Urban Institute
2100 M Street NW
Washington, DC 20037

Acknowledgements

The author received a number of very helpful comments and suggestions to various drafts of the report for which we are very grateful. These reviewers were Dr. Kathryn Newcomer, professor of Public Policy and Public Administration at the Trachtenberg School of Public Policy & Public Administration, George Washington University; John Kamensky, a senior research fellow at the IBM Center for the Business of Government; and Katherine Barrett and Richard Greene, columnists and correspondents for *Governing Magazine,* who focus on state and local government. Finally, the author is very grateful to Price Philanthropies Foundation and the Urban Institute, for their financial support for this work.

Preface

During the latter part of the 20th century considerable progress was made in gaining widespread acceptance for performance measurement as an ongoing part of performance management—at all three levels of government and increasingly within private nonprofit organizations. This is a good thing. However, for the most part, the information provided by performance measurement systems has been both shallow and not always as timely as is needed to help managers operate throughout the year.

Major advancements have occurred in the first decade or so of the 21st century that show great potential for enhancing the value of the performance information provided by these management systems. The opportunities for public and private service organizations to provide more timely and substantive information for managers are exploding. Major advances have occurred, and continue to occur, in areas currently being labeled with terms such as "Data Analytics, "Data Visualization," and "Big Data." The availability of such tools presents government and private for profit organizations with tremendous opportunities to improve the information provided by their performance measurement systems.

This report provides a number of recommendations for making use of such tools to help speed up the development and use of modern technology. Technology-related problems exist, especially the need to provide user-friendly devices that can enable the manager of the 21st century to download at any time and in any location, from some form of electronic device, information that enables them to drill down into the latest available data. This is data that in the past would have required an excessive amount of time and resources to obtain. And, all of this achieved without requiring more than a basic knowledge of analytical methods.

We hope this report will encourage implementation and use of these great opportunities for performance measurement and performance management in the 21st century.

Contents

Section One
Background and Purpose

How Did We Get Here?

Everybody is doing it. Performance measurement is being used to some degree by so many government agencies in the United States that it can be considered almost universal for all but the smallest agencies. Throughout the world, performance measurement, often called "monitoring," is becoming common in developed countries and increasingly in developing countries.

> "Performance measurement" is a process in which a governmental or non-governmental public service organization undertakes regular collection of *outcome* and/or *output* data (preferably both) throughout the year (not only at the end of the year) for at least many of its programs and services.
>
> "Performance management" means the practice of public service managers using performance data to help them make decisions so as to continually improve services to their customers.

Beginning in the 1970s, the key element of the performance measurement movement has been the focus on measuring outcomes, that is, the results of services, which enables the organization to track progress in achieving public service objectives to improve the lives of citizens. The previous focus of performance measurement, before the outcome movement began in the 1970s, was on measuring finances (inputs) and sometimes outputs, (the amount of work completed by the organization).[1]

Performance measurement systems should be designed to provide information to managers and their employees that will help and encourage them to work continually to improve the effectiveness of their services to customers.

> Surprising as it may seem today, prior to the performance measurement movement, regular collection and reporting of the outcomes of services was seldom done by public or private service agencies. *The focus of reporting was financial information.*

[1] Ridley, Clarence and Herbert Simon, "Measuring Municipal Activities: A Survey of Suggested Criteria and Reporting Forms for Appraising Administration," Chicago: International City Managers' Association, 1938. This groundbreaking report moved governments from reporting costs without any examination of what the funds produced, to also tracking the amount of work done, that is, the outputs. However, the focus of that work was on outputs, not measuring the outcomes of the services.

Some of the major developments were the following:

- Local governments effectively started the performance measurement movement in the 1970s. Cities as diverse as New York, Charlotte, North Carolina, Dayton Ohio, and Sunnyvale, California, implemented performance measurement systems. New York City might have been the first government to implement a reporting process that has regularly included outcome information (in its annual Mayor's Management Report) since the early 1970s, a report mandated by the city charter since 1977.
- State governments followed suit starting in the 1980s, with states such as Oregon and Texas implementing their versions of performance measurement.
- The Governmental Accounting Standards Board, an organization that recommends financial reporting standards for both state and local governments, expanded its focus in the late 1980s by encouraging these governments to include performance measurement information in public reports.[2]
- A major breakthrough occurred in 1993 when congress enacted the Government Performance and Results Act (GPRA). The GPRA requires federal agencies to undertake some form of performance measurement. And, because the federal government provides major funding directly or indirectly to state and local governments and to many private nonprofit organizations (NPOs), it has become central to progress in the field. Many federal agency programs require regular performance data from lower levels of government and NPOs.
- The GPRA Modernization Act of 2010 encouraged the use of performance measurement for performance management. A major addendum to the 1993 GPRA requires quarterly data-driven performance reviews of progress made towards each agency's "priority goals."
- The movement is beginning to spread substantially into non-governmental, nonprofit organizations that provide services to the public. It originated around 1996, when United Way of America—with input from a number of local United Ways—produced its report "*Measuring Program Outcomes: A Practical Approach*." Many philanthropic foundations that fund NPOs have also encouraged performance or outcome measurement, often under the label "evaluation."
- In 2010, a National Performance Management Advisory Commission, comprised of representatives from 11 prominent national state and local governmental interest groups, issued a "framework" for state and local performance measurement and reporting. The

[2] For example, see its primary early report on this topic: "Service Efforts and Accomplishments Reporting Its Time Has Come: Overview," Governmental Accounting Standards Board, Norwalk, CT, 1990. It provided a starter set of performance indicators for each of 12 basic state and local government services.

framework can be said to provide "universal" agreement that performance measurement and performance management are here to stay for state and local governments.[3]

[3] "A Performance Management Framework for State and Local Government: From Measurement and Reporting to Management and Improving," National Performance Management Advisory Commission, Chicago, IL, 2010.

THE PURPOSE OF THIS REPORT

Performance measurement and performance management are entering a major new phase. This report aims to provide recommendations to governments at all levels as well as private nonprofit service agencies, to help them transform and upgrade their performance *measurement* systems so as to vastly strengthen their ability to help managers improve performance *management*, and thus provide more effective and more efficient services to citizens.

A number of the recommendations take advantage of the significant advances in relevant technology and also incorporate basic elements of *program evaluation* into performance measurement systems. (Program evaluations are ad hoc studies that delve deeper into understanding particular programs, including, when possible, distinguishing the extent to which the program has caused the measured outcomes.) Adding more analysis into the performance measurement process can provide considerably more rigor and explanatory power. "Evidence-based" decisions and use of program evaluation have in recent years become a focus of attention for the federal government and private foundations, and are trickling down to state and local governments and to private nonprofit organizations.

Advances in information technology (IT) and lowered costs of basic hardware and software, make many of the recommendations here feasible for even small organizations. Not long ago, some of the recommendations would have been prohibitively expensive and time consuming.

CURRENT PERFORMANCE MEASUREMENT LIMITATIONS

Current performance measurement systems have many virtues but also important limitations. These include:

- Collecting only limited outcome information, especially limited information coming from customers, including information on the sustainability of improvements achieved after customers exit services.
- Examining only aggregated data without digging deeper so as to link outcomes to important demographic and service characteristics of those receiving the service. This also limits the ability of the performance data to address equity concerns.
- Undertaking highly limited analysis of the performance data and not tapping into important available opportunities presented by the data to identify more specifically where the strengths and weaknesses of the services are located.
- Not providing for the incorporation of qualitative information to help interpret the data findings.
- Too often providing out-of-date data, thus limiting the data's usefulness to managers.
- Reporting the data in an unclear and/or uninteresting way; discouraging use and encouraging misuse of the data.
- Not providing the training and technical assistance to managers on how to access and use the performance information.

> All the problems listed lead to the information provided by many, if not most, performance measurement systems being too shallow.

Some agencies have done a fine job on one or more of the items listed above but sometimes only when specific issues arise; they are not imbedded into their on-going performance measurement practices. Addressing these limitations and increasing the usefulness of performance measurement is the focus of the recommendations presented in later sections of this report.

The three over-arching performance measurement system limitations are:

- The useable information provided by performance measurement systems to public administrators has been highly limited.
- The data that are available have not been analyzed on a regular basis in order to enable sufficient interpretation of the performance information.

- The performance information has not been fully utilized by public administrators to help them make decisions, most likely in part because of the limitations mentioned above.

Fortunately, a number of developments have occurred in recent years that allow for major improvements in performance measurement and performance management:

- Advances in technology.
- The lower cost of hardware and software.
- Widespread growth in the acceptance of performance measurement and performance management.
- Widespread growth in demand for reliable evidence.
- Increased familiarity with technology among many young (and older?) professionals.

The remainder of this report provides recommendations for addressing the limitations discussed above and for making performance measurement a more effective tool for helping managers manage performance. The focus here is on using performance information to improve services, and not merely using performance measurement as an accountability tool to help assess whether the service agencies are doing well enough.

> Managers, whether in government or non-governmental organizations, make a countless number of decisions throughout the year. Seldom do they have the advantage of drawing on a randomized, controlled trial or on any type of program evaluation. But they still have to make those decisions with whatever information is available to them.
>
> How do we strengthen the information available to public managers? That is the purpose of this report.

REPORT ORGANIZATION

The remainder of this report is organized around the three over-arching limitations mentioned above:

- Section 2 provides recommendations on what performance information should be collected and the data collection procedures needed to obtain that information.
- Section 3 focuses on how the information collected can be analyzed and reported to make the collected information as useful as possible.
- Section 4 focuses on how that information can be used to improve services. The section concludes with comments on the implementability of the recommendations in these three sections.
- An Appendix contains a complete list of the recommendations.

We believe the recommendations to be appropriate whether: (a) your organization is just introducing a performance measurement process; (b) you want to improve your current performance measurement process; or (c) you are adding new services or programs for which a performance measurement process needs to be developed.

> Throughout the report, the word "customer" is used to encompass all forms of service recipients, including those who might otherwise be called "citizens," "residents," "clients," "patients," "beneficiaries," or "students." Some programs do not directly serve human beings. Managers may be seeking improvement in non-human units, such as improving the quality of roads or of water and sewer systems. The recommendations generally apply to such units as well as to people.
>
> Throughout the report, the word "data" is used to mean information expressed in quantitative units. The term "performance information" is used to encompass both quantitative data and qualitative information.

We have attempted to restrict the recommendations to suggestions that do not require on-going use of highly specialized personnel or are likely to require substantial added resources. However, for a few recommendations advanced technology may be required for some organizations.

Section Two
Are You Measuring the Right Things and in the Right Way?

To move from performance measurement to performance management, managers need to get the right data on the right things. Performance management, like performance measurement, begins with selection of the performance indicators for which data will be collected. The recommendations in this report are dependent on the organization being able to generate the relevant performance data.

The following section discusses ways to strengthen this key task. Recommendations are grouped into:

1. Identification of what performance indicators should be tracked; and

2. Selection of data sources and data collection procedures for obtaining the data.

WHAT PERFORMANCE INFORMATION SHOULD BE COLLECTED?

For this report, we assume that an earlier step has been taken to identify the mission of the organization and its programs, perhaps as part of a strategic planning effort. The recommendations below focus on identification of the performance information, both quantitative and qualitative, needed for performance management. The performance indicators selected should, of course, support the mission statement and strategic plan goals.

D1: Seek input from representatives of key stakeholder groups to ensure identification of appropriate outcomes.

Do not rely solely on agency staff to identify outcomes important to customers. Seek input from representatives of key stakeholder groups such as:

- Your customers and other members of the public expected to be affected by the program, including persons such as inmates and substance abusers who in a sense are "customers";
- interest groups with significant subject matter interests in the particular service or program, such as environmental groups and particular demographic groups (e.g., those representing seniors, children, Hispanics, etc.);
- business representatives; and
- staff, particularly program field staff, who are likely to be aware of the concerns of the customers served by the program.

Input from stakeholders can be obtained inexpensively through such means as meeting with small groups of customer representatives.

D2: First, consider indicators for which you already have data. Then, consider indicators needed to measure outcomes for which you do not currently have data.

Do not only include performance indicators for which the data are readily available. Consider new measurement approaches. It is a great temptation to only seek performance indicators for which data are already collected or can easily be obtained. *Unfortunately, some of the most important outcome indicators may require data not previously collected.*

> A classic example comes from the Center for Medicare & Medicaid Services (CMS) and picked up by some state governments in their state health reform efforts. CMS has recommended two indicators for tracking tobacco cessation: (1) the percent of patients who were queried by physicians about tobacco use; and (2) the percent of patients who smoked that received cessation intervention. No indicator has yet been included that addresses whether the patient *stopped smoking subsequent to the intervention*. [See, for example, "Priority Measures for CMMI Monitoring and Evaluation," Center for Medicare and Medicaid Innovation, DHHS-CMS, September 2013.]

As the above example illustrates, programs commonly track the number of customers served but not the number who were actually helped after receiving services.

Organizations are reluctant to introduce new measurement procedures. A major reason has been their perception of the (high) cost of collecting the new information, such as introducing the regular use of customer surveys. Certainly, data collection costs need to be considered in selecting indicators. However, also to be considered is the added value of obtaining more key information such as on the quality of the program's services. In some instances, the cost problem is potentially generated by overly strict or overly rigorous collection requirements, which is discussed further in the following subsection on data collection procedures.

D3: Use "logic models" ("outcome sequence charts") as a tool to help identify performance indicators.

Logic models are diagrams that trace the sequence of expected/desired outcomes from the planned program activities through the ultimate end outcomes desired. Logic models have become common in federal agency studies, such as program evaluations. They can be just as useful for helping program managers identify performance indicators for performance measurement systems. They can be used for programs that have not yet identified outcome indicators; when a new program is being implemented; or when an agency is not happy with its indicators.

Logic models are also helpful in highlighting the relative importance of the various "intermediate" and "end" outcomes (intermediate outcomes are those expected to result from the program activities and are expected to lead to the desired end outcomes that measure benefits to the public). For most, if not all programs, both outcome categories should be included in the performance measurement system. The intermediate outcomes come earlier and are more controllable by agencies.

A classic example of the application of logic models is the attempt by many human service programs to reduce drug and alcohol abuse. Programs might first seek to improve customers' *knowledge* of the problems caused by particular poor behaviors. Improved knowledge is presumed to lead to improved *attitude* about the behaviors, which in turn is presumed to leading to reductions in those poor *behaviors*, finally leading to *healthier and more productive individuals*. For each of these outcomes, specific measureable outcome indicators are needed to enable program managers to track progress and improve services when progress is not found to be satisfactory.

A related term sometimes used in program evaluations is "theory of change," which refers to a more detailed description of the theoretical basis for a program and the assumed causal mechanisms that are needed to produce desired ends. The logic model is a condensed version of theories of change. Some federal programs have begun requiring organizations to provide a logic model as part of their applications for funds. For example, CMS has required states participating in its state health reform, "SIM" program, to prepare a somewhat more complex version of logic models called "drivers diagrams."

Figure 1 on page 17 is a basic, simplified example of a logic model. Each result shown in a block indicates the need for one or more performance indicators that measure the extent to which that result has been achieved. The question is asked after each block: "What do you next want to result?" Logic models can become much more complex, especially as they address more complex programs. Figure 2 on page 18 is a somewhat more complex version that illustrates the logical flow from intermediate outcomes to end outcomes. Most logic models begin by adding blocks on

the left-hand side that identify the program activities being used or proposed, which are believed to lead to the outcomes.

Developing logic models is also an excellent teaching/training device to encourage employees to think more in terms of outcomes. Logic models have been described in a number of publications in recent years.[4]

[4] For example: "The Logic Model Guidebook," Lisa Wyatt Knowlton and Phillips, Cynthia C., Sage, second edition, 2013; "Developing a Logic Model: Teaching and Training Guide," Ellen Taylor-Powell, and Henert, Ellen, University of Wisconsin Extension , Madison WI, 2008; and "Logic Model Development Guide," W.K. Kellogg Foundation, Battle Creek,. updated 2004

D4: Include outputs, intermediate outcome, and end outcome indicators in performance measurement systems—and distinguish which is which.

Outcome indicators have been the missing metrics in performance measurement systems in past decades. Now that outcome measurement has become far more common, the temptation is to skip *output* metrics. Avoid this temptation. Among the important advantages of measuring outputs are that they can be readily measured; programs have considerably more control over them; and, most important, they provide highly useful information for budgeting as well as for managing program activities. The assumption is that an agency needs to complete work that in turn leads to outcomes. If the number of outputs is low, the expectation is that this will lead to a reduction in outcomes.

Typically, it is important to include both intermediate and end outcome indicators. In this report the terms "intermediate" (or "interim") and "end" outcomes are intended to relate to the *importance* of the indicator, not its *timing*. For example, some very important outcomes can occur very soon after program actions, such as hot lines that prevent suicides or certain health actions that can have significant and immediate impact on a patient's major health problem. An intermediate outcome will, by definition, precede an end outcome with which the intermediate outcome is associated.[5]

Intermediate outcome indicators (such as the "number who completed the stop-smoking program" in Figure 1) enable managers to obtain earlier feedback on progress toward end outcomes. However, if end indicators (such as the "number who actually stopped smoking for a certain length of time") are not tracked, users can lose sight of what their program is really about. An exception occurs when the agency believes it cannot measure the end outcomes and has to rely on intermediate outcomes as a surrogate (proxy) for measuring end outcomes.

It is likely to be helpful in reports to categorize each indicator as to whether it is an output, intermediate outcome, or end outcome indicator. This will help users of the information put the relative importance of the indicators into perspective.

[5] Many professional analysts prefer to use the terms "short-term" and "long term" to distinguish outcome indicators. Thee important element is that the sequence of the outcomes is clear. (Logic models are intended to help make the sequence clear. The actual timing of outcome indicator values (the number of years before the numerical values for individual outcome indicators are expected to occur) becomes very important when *projecting* outcome values as part of strategic planning and budgeting.

D5: Collect outcome data that identify the outcome at a time *after* the customer has completed the program's services.

For many programs this has been, and continues to be, a major gap in performance measurement. Failing to obtain information on how customers are doing after they have exited services has been a badly neglected category of performance indicators at all levels of government and among NPOs.

> A major limitation in most performance measurement systems is the lack of outcome indicators that measure the condition of the customer especially after they have exited the service. Such information provides considerably stronger information on program success.

For many services, a major outcome desired is that customers remain better off for a sustained period *after* they exit from the service. The needed outcome indicator would be something like: "Number, or percent, of customers who X months after completing service had the desired 'condition.'"

The following are two examples of this all too prevalent gap in performance measurement systems:

> The Corporation for National Community Service, through its grantees and service members, is providing services to veterans and their families. The corporation has established performance indicators on which grantees need to report. Thus far, the required indicators track the number of veterans assisted. No indicator is currently required to address whether the veterans received any benefit from that assistance or even took steps toward obtaining such benefits.
>
> The CMS example of smoking cessation performance indicators provided in D2 is another example. The current indicators track those who received intervention but not the outcomes of those interventions.

In both of these examples, surveys of program customers after the services were provided would likely be needed to obtain the outcome information.

The follow-up process can provide key information on the helpfulness of the service in providing sustainable relief from the problems for which customers entered the service—and also obtain suggestions for improving the service. This information should help managers identify needed program improvements. In some program areas, such follow-up has been required by federal agencies, including past employment and substance abuse treatment programs.

The need for post-service follow-ups applies to most "human services" other than long-term programs where customers remain in care and are receiving periodic performance measurement. The need also applies to programs that focus on non-human physical conditions. For example,

programs to improve or maintain air quality or bodies of water track conditions on a regular basis so that change data should be available after a program's corrective actions have been taken.

When should the follow-ups be done? How often and when should outcome information be collected after service completion? To limit costs, one follow-up is likely sufficient for performance measurement systems for most programs but would ultimately depend on the nature of the service. For performance measurement purposes, and when surveys of customers are needed to obtain the information, the follow-up should be done no longer than 12 months after the customer's exit. It would likely be difficult to find customers after that period. Longer-term follow-ups are more appropriately done as part of in-depth program evaluations.

It should be noted that if post-service data collection is to be collected by person-to-person contacts with the customer—to avoid potential bias and lack of external credibility of the data—the contact should not normally be undertaken by the customer's own service provider, such as a caseworker or doctor. The interview might instead be conducted by supervisory or administrative staffs; likely less costly in time and money is to use mail or electronic devices (particularly for surveys of businesses).

A major obstacle to following-up with customers is the perceived cost of obtaining the information on a regular basis. The position taken in this report is that *often the costs of post-service customer follow-ups need not be as expensive as many managers expect*. The added time and cost expended in obtaining such follow-ups should not detract significantly from service delivery, and the value of the information appears considerable.

To add considerably to the value of such customer follow-ups, the information sought can also ask the former customer for information on the circumstances relating to the success or failure to achieve the desired outcomes. Such information can provide important leads to ways to improve the service.

Another obstacle is that program managers and staff sometimes believe that following-up on clients is beyond the scope of their work; they should not be responsible for customers' condition after exit. However, the intent of many programs is to alleviate/correct problems not just while the person is under care but afterwards as well. Having information on the extent to which the service has helped its customers is important for program managers in identifying whether the program is working well or not well and whether program corrections are needed.

Follow-up information can be obtained by: (a) surveys of customers; (b) from administrative records (possibly from other programs or other agencies); (c) trained observer ratings; and (d) special sources such as tests (e.g., for educational progress) and special equipment, (e.g., for measuring air and water pollution levels and road ride-ability). Such data collection procedures and recommendations for alleviating the costs of follow-up data collection are discussed below under "Data Sources and Data Collection Procedures."

Figure 1

Example of a Logic Model

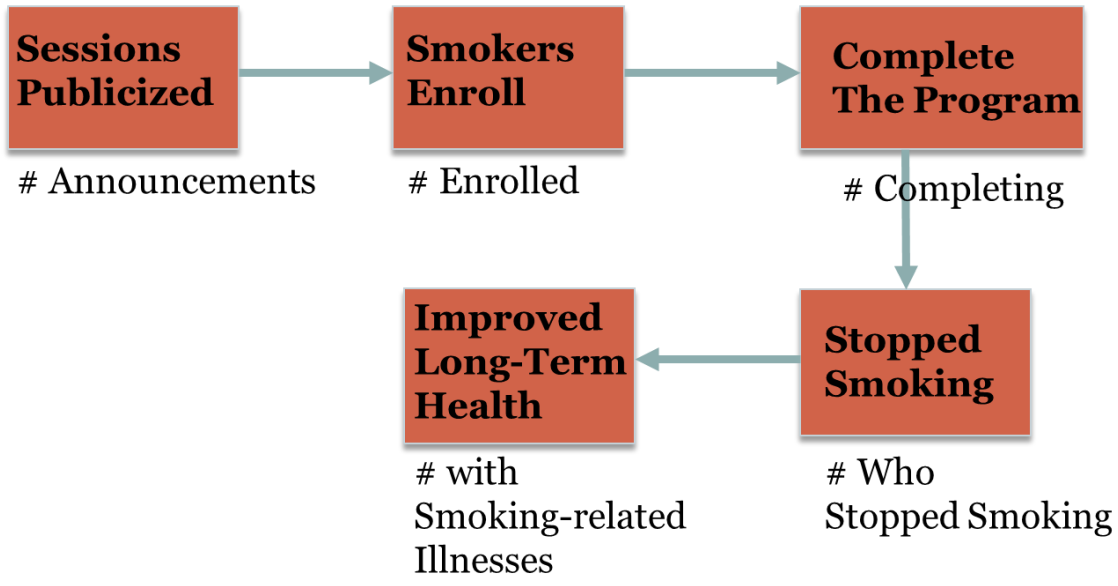
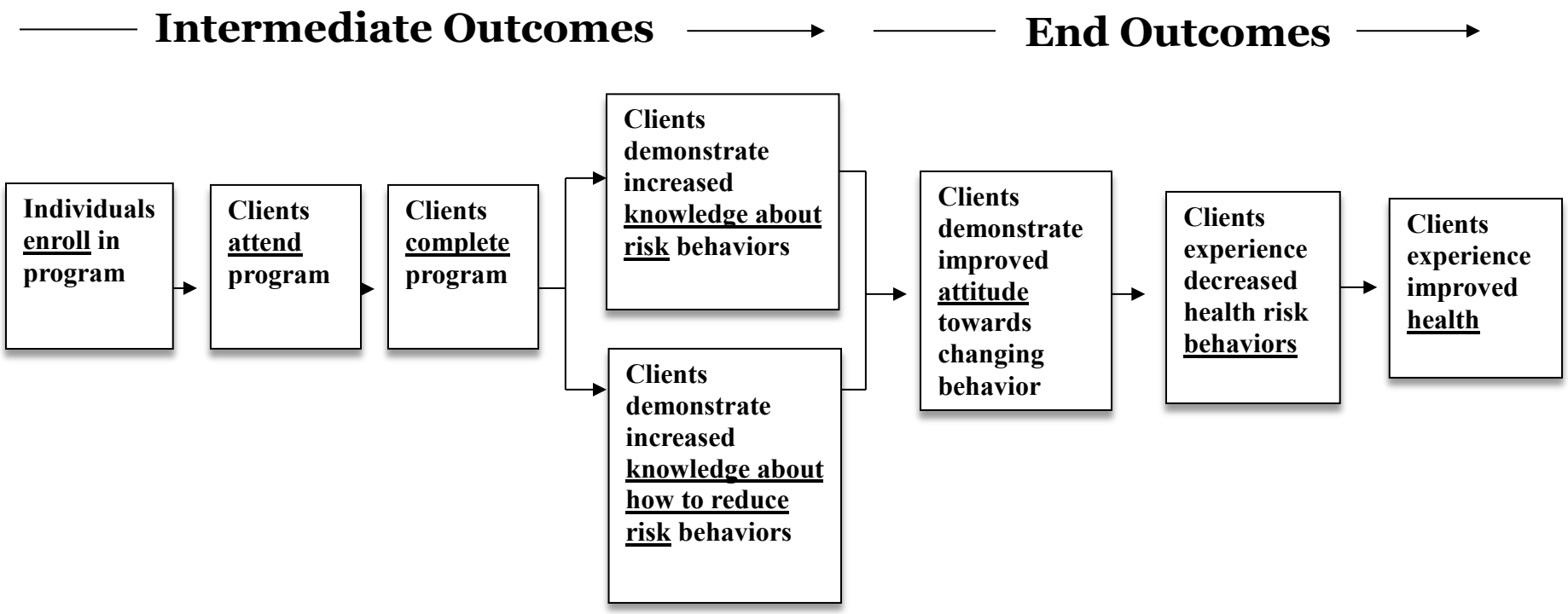


Figure 2

Example of a Health Risk Reduction Program



D6: Include indicators for outcomes over which you have some responsibility even though only limited control. Perhaps identify in external performance reports those indicators over which your organization has only highly limited influence.

A major challenge in measuring and reporting performance information is that too many people, including the media, citizens, and even high-level officials assume explicitly or implicitly that, if an agency reports on an outcome indicator then that agency has complete or at least a substantial ability to affect those outcome values.

However, other factors over which the agency has little control often significantly affect the outcomes, for example, unexpected changes in economic conditions; the weather; new legislation; and even internal events, such as health problems affecting key employees. Agency managers are inevitably, and understandably, concerned that they will be blamed unfairly for outcomes that are not as good as expected. (The issue applies primarily at the federal level and somewhat at the state level. For example, these governments pass funds through to lower government levels, limiting their control.)

> A major problem for performance measurement reporting is the interpretation by the media, legislators, and even high-level officials that reported poor outcomes are always the fault of the program and public employees.

> The US Department of Education reports on national test score results even though the department plays only a limited role in education delivery. Conversely, the Equal Employment Opportunity Commission plays a major role in identifying and enforcing national laws on employment discrimination but has a limited ability, by itself, to affect overall national levels. Nevertheless, should it not also report on estimated national levels of employment discrimination? (It currently does not, in part because of the perceived cost of data collection.)

Outcome data identify the amount of benefit received but not WHY. Organizations need to do a much better job of making this clear. One way to do this is to categorize each performance indicator by the extent of control the program and organization actually have in producing results, such as by categorizing each outcome indicator as to whether the agency has "extensive," "moderate," or "limited" control. A rare example is from the state of Virginia:

> The State of Virginia's "Virginia Performs" website identifies for each of its reported indicators whether the state has "significant" or "limited" control. Also, for each indicator it includes a section that briefly describes the state's role, including the services it provides that affect the indicator values.

A second way to address this concern is to attach explanations for not meeting expected outcome levels to performance reports. As is discussed further below, explanatory information

such as that obtained from administrative records, feedback from citizens or more expensive program evaluation could be the source of such information.

D7: Track efficiency but do not settle for output efficiency. Focus, when possible, more on outcome efficiency: measuring efficiency in producing outcomes.

Saving costs has always been an important concern for public service agencies. Today, and for the foreseeable future, it is of major concern across the nation.

Performance measurement of efficiency has traditionally used indicators of the form: "average cost (or number of FTEs) per unit of output." However, "true" efficiency is reducing costs without sacrificing service quality. A more important indicator is to relate cost to the amount of outcome achieved. The performance measurement is then "average cost per unit of outcome." If only the traditional performance indicator is used, reported values can be achieved by reducing service quality. For example, speeding up response times can be achieved at the expense of service quality.

Changing the focus from performance indicators of the form "cost per client served" to "cost per client served who achieved a specified amount of improvement," might, however, require special data collection procedures to identify clients' condition after clients completed the program's services.

Not all outcome indicators will be amenable to efficiency measurement. Indicators that measure problem levels do not lend themselves to efficiency indicators. For example, "cost per crime that occurred" does not make sense as an efficiency indicator. The "cost per crime prevented" would be a great efficiency indicator if the number of crimes prevented could be measured on a regular and reasonably sound basis.

D8: Include "milestone" indicators for programs with substantial start-up steps.

New or substantially modified programs can require significant implementation time. Special metrics are likely to be needed that reflect the achievement of major implementation milestones. (Such indicators are a form of output indicators.)

> For example, a number of state governments have begun to implement major statewide health reforms. These reforms will require a number of years for full implementation and some important outcomes cannot be expected for several years. The US Department of Health and Human Services-CMS State Innovations Model Program (SIM) provides support to states to help them implement major health care reform innovations. These innovations include providing "patient-centered medical homes" for patients (that provide integrated services for their patients) and encouraging pay-for-results programs. These are complex efforts requiring participation and information sharing across many organizations within a state, and they require legislation changes as well. The states are likely to require a number of years before the effort can be fully implemented and when improved health outcomes can be expected to appear.

Other examples where milestone indicators are likely to be useful are research and development (R&D) programs. The outcomes of most R&D program projects can take years until the research is competed, tested, promulgated, implemented, and for the desired end outcomes to occur.

For such programs, indicators can be included that assess progress on key milestones. Milestone-tracking is seldom included on current performance measurement systems. Progress information is needed for both program managers as well as high-level officials to enable them to track progress and identify the need for mid-course corrections. A milestone indicator might simply be whether the milestone was completed or not. Another option is to provide estimates of the percentage completed at the end of the reporting period for each milestone.

D9: Provide key disaggregation of performance data routinely by major customer characteristics, such as by demographic and risk characteristics.

Regularly providing performance data broken down by customer characteristics (such as age group, gender, race/ethnicity, and where they live) is still rare but this information can provide managers and their staff with much richer information than merely providing aggregate values. Disaggregation of *outputs* by demographic characteristics (e.g., breaking out the number of customers served) has been considerably more common than disaggregation of *outcomes* by demographic characteristics. (Disaggregation of outcomes is commonplace in the field of program evaluation.) Such information can be used for many purposes as will be discussed in Section Three below.

> Reporting outcomes by major customer characteristics presents a major opportunity for improving the information available to managers, public officials, and the public.

Many programs already collect demographic information on their customers. This should then be linked to the outcome information on the customer and tabulated for each customer category. Major technology advances in data processing enables organizations to process such data and calculate the values for each outcome indicator broken out by each category. Thus, obtaining these calculations need not be the significant burden it once was. Advancements in technology have made such disaggregated performance information readily available and accessible to managers if breakout categories of interest are inputted into the database. For example, customer zip codes can be coded so that outcome data can be disaggregated and tabulated for each individual zip code.

Disaggregations are equally relevant to public programs for which people are not the direct focus. For example, road maintenance outcomes can be disaggregated by category of road and by average daily travel levels; water pollution control programs can be disaggregated by individual segments; water main breaks can be broken out by material and type of failure; business development program success can be disaggregated by individual categories of businesses; and crime control programs are often disaggregated by crime category, geographical location, and time of the day and week.

A major variation is to characterize each unit of incoming workload (whether customers, bodies of water, road segments or sewer-line segments) by level of difficulty (level of risk) in achieving desired outcomes. To rate risk/difficulty, each customer might be characterized at or near intake by a supervisor or other "judge" into one of a few levels of difficulty. Outcomes would subsequently be calculated for customers at each level. For some programs, one or a small number of demographic factors might be sufficient for categorizing difficulty levels. For example,

the percent of students enrolled in the school lunch program, a reasonable proxy for family income, might be used to indicate the extent of disadvantaged students in different schools, indicating the need for special help in getting those students up to grade level. For health services, patient outcomes can be disaggregated by patients' incoming health condition, as an approximation for case difficulty.

More sophisticated and more accurate statistical methods can also be used to develop difficulty levels. This is being done in the health field to provide fairer comparisons across hospitals or other health venues. Mapping software is also making the production of attractive maps displaying outcome data by geographical area considerably easier (using colors, numbers, shading, etc.).

A potentially BIG challenge: Providing so many outcome indicator breakouts can lead to managers being drowned in a sea of data. Fortunately, current technology allows performance measurement systems to enable managers and their staff to extract the more detailed disaggregated information *only when needed*. For example, such disaggregations can be made available through electronic means by providing links ("hyperlinks"). These links enable data users to pull up outcome tabulation breakouts only when needed, such as for customers with specific characteristics (e.g., those customers residing in particular zip codes with specific employment problems).

Recommendations for the analysis of disaggregated data are presented in Section Three below.

D10: Provide key disaggregation of performance data routinely by major service characteristics, such as by service type and amount and by provider.

Information that can be obtained by disaggregating outcome data by different service characteristics is largely untapped in performance measurement systems. Yet it can be of considerable help to managers in identifying what is working well and what is not...

For example, it can be done by examining and comparing the outcomes of *different service units*. Examples of such service units are: regions; districts; schools; correctional facilities; health care facilities; parks; offices; teams; individual caseworkers/teachers/ doctors; and individual contractors/grantees. The manager of each such service unit should, of course, be provided with the outcome data for their own unit as well as the aggregate data for all units.

It can also be highly useful to disaggregate outcome information by *amounts of service and/or types of services provided*. For example, an agency might record the amount and type of casework customers received (such as the extent to which individual, group or electronic means was provided to individual customers). Some public programs have been doing this for many years. For example, it is not uncommon for water and sewer maintenance programs to track the maintenance record for individual pipe segments.

> Human services case management software has reached a reasonably mature stage. A number of software programs have become available that enable agencies to track the use of particular services for *individual* clients. The software enables the agency to link demographic and service characteristics to client outcomes (obtained from outside the case management system).

D11: Require agency programs to provide explanations for unexpectedly poor and unexpectedly good outcomes as a standard part of the performance measurement system.

Many outside factors in addition to a program's own efforts can affect the outcomes sought by the program. (These factors are sometimes called "contextual factors.")

Such explanatory information can be very useful to program managers in providing clues as to why the measured outcomes have occurred and what might be done. *Providing explanations for unexpectedly poor and unexpectedly good outcomes, even if resources permit only a small effort, should be considered part of a modern performance measurement system*.

> The Texas state legislature, for example, has for many years required explanations "when actual performance of key measures varies five percent or more from targeted performance. These explanations should describe the circumstances that caused the agency's actual performance to deviate from its performance targets." Reviewers of the information ask "Did external factors affect performance to the extent that targets could not be met?"[6]

> Even simply asking the program manager to provide explanations and insisting that the explanations be substantive can stimulate constructive investigation by the program.

Such explanatory information might be obtained from administrative records (including complaint and 311 systems), customer surveys, and trained observer ratings. Another emerging source is the tracking of social media content such as Twitter and Facebook.

> An example of explanations for poor ratings obtained from *administrative data* is the work of transportation and public works agencies to track traffic or highway accidents. These agencies typically also identify the reasons for the accidents (such as driver error, vehicle malfunction, weather conditions, poor signage, etc.).

Performance indicators that obtain quantitative data from *customer surveys* can also ask respondents for important qualitative information. Respondents can be asked to:

a. Provide explanations for giving poor ratings to particular service characteristics about which they are asked; and

b. Provide suggestions at the end of the questionnaire for improving the service.

Similarly, agencies using "*trained observer*" rating procedures as part of the performance measurement systems can also obtain very useful explanatory information. The trained observers,

[6] The State of Texas, "Guide to Performance Measure Management: 2012 Edition," March 2012, Report No. 12-333.

for example, might be rating the cleanliness of streets, road ride-ability, the condition of buildings, or the readiness-for-school of pre-K children. The raters can be asked to provide an explanation as to what was wrong in situations where the observer gave low ratings.

Whatever the source, explanations can be categorized and then tabulated. This yields an indicator of the form: "Number /percent of persons indicating that the reason they gave a poor cleanliness rating was that too much litter was present." A program can also seek explanatory information in a number of other ways. Explanations might be sought by forming a special working group to examine the causes of unusually good or poor outcome values. Field personnel familiar with the way the program operates can be a productive source of knowledge on why the outcomes were unexpectedly poor (or good).

Exhibit 1 provides suggestions as to sources of explanatory information.

Exhibit 1

Obtaining Explanations for Unexpected Outcome Values

1. Examine findings from the previous steps to identify which client and service characteristics appear particularly related to the unexpected outcome values. While this will not explain why the outcomes occurred, it will enable the program to narrow in on where to search for explanations. For example, the group therapy program whose data are presented in Exhibit2 might focus its search for explanations on problems with females and those that attended less than two sessions, particularly the clients of caseworker C. In this example, the problem may be that clients of caseworker C only attended one or two sessions.

2. Discuss the reasons for shortfalls with supervisors and their service providers. Such discussions might be with individual supervisors and providers, or might be in groups. One very useful approach is to hold "How are we doing?" sessions shortly after each outcome report becomes available and is disseminated. At this session ask staff to provide their reasons for unusually good or poor outcomes identified by the report.

3. Examine client responses to "open-ended" questions included in surveys, particularly responses to questions asking respondents to provide their reasons for low ratings to particular service characteristics or their suggestions for improving services. Programs that survey their clients for outcome information should include such questions and examine the responses. The responses should be categorized and grouped by common themes to identify specific problems and client suggestions. For example, a number of clients might have stated (though probably in somewhat different words) that they had difficulties getting through to program staff by telephone, or getting to the program's facilities. Such information can suggest corrective actions that the program can take to ease such problems.

4. Hold focus groups with clients (former and/or current) to elicit their views about the problems. (These are 90- to 120-minute meetings with perhaps 8–12 clients in which a facilitator solicits comments and thoughts about the program's service.)

5. Form a working group of staff, and perhaps volunteers, to examine the problem and why it occurred (and ways to correct it).

6. For major problems, seek an independent in-depth program evaluation. Perhaps recruit an outside organization, such as a local university or community college, to examine the reasons for the problem.

Source: Hatry, Harry P., Jake Cowan and Michael Hendricks. *Analyzing Outcome Information Getting the Most from Data*. Washington, DC: The Urban Institute, 2004.

DATA COLLECTION ISSUES

You do not have a performance indicator until you also have a feasible data collection procedure.

Most performance data have traditionally come from agency records, often called administrative data. These are data that an agency routinely collects on its activities. It is the basic source of performance data for output indicators and efficiency indicators that focus on output efficiency. Other important data collection procedures include the use of customer surveys, trained observer rating procedures, physical measurements (e.g., measuring air and water pollution and road conditions), and tests (e.g., educational achievement).

The following recommendations focus on a few of the special issues that can arise in data collection.

D12: Identify opportunities to obtain data from other programs or agencies.

Outcomes for some programs may be available from other programs or agencies. However, obtaining the desired data can be a problem if provisions, such as memorandums of understanding (MOUs), are not made with the other agencies to access the data. Two examples of data available from other programs or agencies are as follows:

> State unemployment data have been used to assess the extent to which employment programs have helped people obtain jobs. The federal–state unemployment insurance program requires (most) businesses to provide quarterly earnings information on covered employees. However, this source has several drawbacks, including confidentiality issues; delays (lag time) before the data can be obtained; employment information availability only on an aggregated three-month basis; and the accuracy of the information on individuals, especially those who have changed jobs or moved.

> Prisoner reentry programs use data from law enforcement and other agencies to monitor success of those programs.

Increasingly, governments are recognizing the need for more holistic examinations of major public issues, such as health care improvement, environmental programs, reducing poverty, reducing unemployment, etc. These involve participation and data sharing across agencies and organizations, perhaps both public and private.

> A number of states in the CMS State Innovation Model (SIM) health care reform program are attempting to pull data together from many health care organizations to provide more comprehensive data—such as information from Medicaid and Medicare claims data—on levels of quality-of-care on their citizens. The effort requires substantial participation from many agencies, both public and private sector sources.

Advances in technology are beginning to generate new opportunities for such data-information sharing. The ability to process large amounts of data from many different organizations ("Big Data") may help agencies incorporate and analyze data from multiple sources.

D13: Regularly survey your customers and former customers to obtain key outcome information.

For many basic outcome indicators, surveys of customers may be the only practical way to obtain vital outcome data. For example, useful administrative data for assessing service quality or the condition of customers after receiving services is often not available for many services. Customer surveys can be used, and generally should be used, to obtain measurements of both customer satisfaction and the condition of customers. Surveys are a growing but still greatly underutilized data collection procedure in performance measurement systems.

Many current surveys have been done of the whole population of a city, a state or a nation. Likely to be of more value to program managers are surveys of only the program's own customers. *For performance management systems, surveys of an agency's own customers can be a major source of outcome data*.

When dealing with large numbers, surveys of a program's customers can involve a representative sample rather than the entire customer base As long as the respondents are reasonably representative of the agency's customer population, this information provides a rare occasion to hear from a *representative set* of customers, an important advantage of customer surveys. Systematic surveys are more likely to provide valid performance measurement information than such procedures as relying on complaint and 311 tabulations; such data are likely to come primarily from people who know how and are willing to use government processes. The squeaking wheel, often the complainers, gets attention while the rest of the wheels are not heard from.

For some outcomes, such as ratings of the quality of the service received, surveys can be undertaken at the time of service completion or shortly thereafter. However, to obtain important information on the sustainability of the service's assistance more time needs to elapse, usually at least a few months. Post-service surveys require locating the former customers and then getting them to complete the survey.

Whether or not post-exit feedback surveys are conducted, it is a good idea for agencies to survey their customers at or near the time they exit programs to obtain feedback on the quality of their experience. Early feedback has begun to be used at all levels of government, but its regular use has likely only been implemented by a small proportion of agencies. The following example provides a good list of service experience characteristics that can be obtained by customer surveys:

> The US Department of Veterans Affairs (VA) has begun to sponsor regular surveys of both its inpatients and outpatients (using contractors to administer the surveys). Its goal is "to improve patient satisfaction by having a survey service provide real-time feedback,

identify root causes of dissatisfaction and other solutions for progress."[7] The VA's Central Iowa Health Care System is obtaining the following set of quality-of-care outcome information:[8]

a. Front staff courtesy and helpfulness;

b. Scheduling ease and accuracy;

c. Speed of registration process and length of wait (findings might have helped early identification of the problem the VA had in early 2014 over its data on waiting times);

d. Nursing staff courtesy;

e. Provider courtesy;

f. Quality of explanations;

g. Amount of face-to-face time with patients;

h. Office operations (hours of operations, cleanliness, etc.); and

i. Sensitivity to the patient's needs (scheduling, privacy concerns, convenience of office hours).

The VA surveys are used to obtain timely data on the quality of patient care; the returned survey data are to be submitted to a centralized system within 48 hours of receipt. Such information should be of considerable help to the agency and staff in identifying both problems and successes in their quality-of-care.

The set of quality-of-care elements listed above can be adapted for use for many other public programs. (This survey, however, does not ask respondents for information relating to the health effects of the VA assistance.)

The cost of surveys is a major concern and major obstacle to their use. especially if the surveys need to be undertaken on a regular basis and after customers have exited services. Surveying customers near or at the time of exit is likely to help program managers with quality-of-care information. However, it is not likely to be sufficient for assessing the sustainability of the

[7] VA RFQ818805 issued August 30, 2013.

[8] VA RFQ818961, 8-30-2013.

improvement in the customers' condition.[9] Information about a customer's condition after departure from services can be vital to understanding what effects the services have had on them. Follow-up surveys should be given strong consideration.

Customer surveys, however, need not be as costly as managers perceive them to be. The following are some suggestions for making follow-up surveys less expensive, so that they can become a routine part of performance measurement systems:[10]

1. Keep the questionnaire short. Performance measurement systems are not research. They are intended to help agencies manage and improve their programs on a continual basis. The basic outcome information on quality-of-care and extent of condition improvement, along with a few "context" questions (e.g., asking respondents to provide a few demographic characteristics) does not typically require more than the equivalent of a two-sided questionnaire.

2. At the time of entry and exit, indicate to customers that they will be asked to provide information on their experience with the service and ask for their help.

3. Obtain contact information from customers.

4. Avoid requiring a high degree of precision and rigor in the data sought by your performance measurement system, such as in precision and response rates. Many if not most public programs do not require such high levels, given the uncertainties that other sources of information are likely to include. Excessive requirements can lead to considerably higher measurement costs with only marginal value added. Data collection costs can escalate considerably when sample sizes are increased; high survey response rates are required; and the survey questionnaires contain large numbers of questions. Questionnaires should be kept brief but focused.

5. When appropriate, use relatively inexpensive distribution methods such as mail or e-mail to administer the survey, at least for segments of the customer population who can be reached by these means. The use of a combination of mail and e-mail surveys, with two or three subsequent mailings is often feasible. In the future, electronic responses might be obtained through mobile devices or laptop computers.

---

[9] For federal agencies, another obstacle to customer surveys has been the requirement that OMB review customer surveys if more than nine respondents are to be surveyed. This has inhibited the use of customer surveys, probably leading to less useful information becoming available to public service managers.

[10] For further discussions of ways to make follow-ups of former customers practical, see Nayyar-Stone and Hatry, "Finding Out What Happens to Former Clients," The Urban Institute, 2003.

6. While they are receiving services, let customers know that they will be asked to provide feedback to help the agency improve its services to future customers.

An approach that can make seeking feedback from former customers of more interest to programs is to incorporate it as part of "after-care"—as a regular task of human services agencies. This is likely to make follow-ups more agreeable to program managers and caseworkers. At the time of follow-up, former customers should be asked about their own need for further assistance as well as be administered the follow-up questionnaire. The survey might be conducted by mail and e-mail (or telephone if resources are available). Note that if the customer is interviewed by his or her own caseworker, the responses may not be frank. Thus, any telephone interview should be handled by a person other than the client's caseworker, perhaps a supervisor or the agency's administrative personnel.

Surveys undertaken by human services agencies have some advantages over those undertaken independently by survey organizations. The human service agencies are likely to have had good relations with their former clients, which is likely to lead to higher response rates especially if they emphasize that the information being sought is to help improve the service for future customers. The agencies have the additional advantage that they are likely to have contact information on former customers, making them easier to locate for the follow-up.

D14: Avoid excessive precision and rigor when designing data collection.

Precision and rigor cost money. Data collection costs appear to discourage public and nonprofit agencies from undertaking ongoing performance measurement; in particular, when it requires new data collection procedures such as customer surveys, which can be the only feasible way to obtain data for some important outcomes.

Requiring too much precision usually means drawing larger samples than needed, thus escalating survey costs as noted briefly under D13. The same issue applies to other data collection procedures such as determining response times from administrative data.

Similarly, requiring high response rates means that considerable added effort is needed both to find former customers and then obtain responses, again escalating costs. Requiring surveys to be administered by outside contractors can also be costly, though the rigor of the survey process may be greater. However, this is not to say that survey administration should be haphazard. Obtaining guidance from outside experts to help write questions and administer the survey process, for example, is a good idea and this onetime cost need not be large.

When sample surveys are used, requiring high-precision levels in determining sample sizes (such as 95 percent or even 90 percent confidence limits) can be overkill, demanding larger than necessary sample sizes and thereby escalating survey costs.[11] It is likely to be overkill for most public programs—other than for high-stakes public policy issues.

However, it may be important to oversample certain customer subgroups to ensure that enough responses are collected for certain at-risk groups, such as a particular race/ethnicity. For example, from a survey of all city residents reliable information may be needed on residents who are of a particular Asian nationality but only a small number live in the city. To obtain an adequate representative sample, the city would need to seek more responses from persons of that nationality.

Response rates of 60 percent or higher have become very difficult and costly to achieve, in part due to the prevalence of mobile phones at the expense of land lines—and the ease of screening out callers. Lower rates can, however, be useful as long as the survey is otherwise conducted in a sound manner with good valid questions and a reasonable attempt has been made to obtain a representative sample.[12]

---

[11] Statistical confidence limit information indicate the probability, with a desired confidence level such as 90 percent, that the sample findings are within plus or minus a specified number of percentage points of the true value for the whole population represented by the sample.

[12] A step, often currently practiced after the survey has been completed, is to examine the demographic characteristics of the set of respondents and compare them to the characteristics of the whole population the sample is to represent. If these are reasonably close, this provides more confidence that the findings are representative of the population. A second approach, used to adjust the overall findings is to choose a key

Determining just how much precision and accuracy is required is a complex question. The choices will likely involve political as well as programmatic considerations such as public health and safety and how much public attention is likely to be garnered. While analysts or other staff can help lay out the issue, ultimately, the managers will need to make the final call.

An additional problem here is that the recent push for "rigor" by the federal government can easily be interpreted narrowly to require considerable precision. *For most public services it is better to be roughly right than precisely ignorant*. *Excessive rigor can lead to rigor mortis* with nothing being done.

It is better to be roughly right than precisely ignorant.

Too much rigor can lead to rigor mortis.

demographic characteristic, such as race/ethnicity, and weight the survey findings for each such demographic group (such as race/ethnicity) by the percentages of the total known population that has that characteristic. This hopefully provides more accurate survey estimates.

Section Three

Analyzing and Reporting the Data: Making the Data Considerably More Useful

Now that you have performance data what do you do with it? *Lack of at least some regular basic analysis of the performance information is probably the major missing element today in many if not most performance measurement systems.*

Basic analysis is a key step required to make sense of the data collected in order to provide useful information—such as identifying what is working well and what is not—so that managers can take improvement actions. While performance indicator data are important, by themselves they do not tell much of a story. Too often performance measurement data are reported to potential users without any thoughtful examination.

A major element in improving analysis is to incorporate into performance measurement systems some of the basic concepts of program evaluation so as to increase the depth and rigor of the information collected. In this report, we are not talking about sophisticated statistical analysis. While ideal, as a practical matter such analysis is not likely to be available to most managers and certainly not on a frequent, regular basis. This report, therefore, concentrates on basic, straightforward, approaches can be almost universally applied.[13]

The following recommendations are grouped into two categories: (1) analyzing the information; and (2) reporting the information.

[13] Only very recently are we beginning to see reports identifying this need. For example, the IBM Center for The Business of Government has in this decade published studies on this topic, including: "From Data to Decisions I: The Power of Analytics" (2011); "From Data to Decisions II: Building an Analytics Culture" (2012); and "Predictive Policing: Preventing Crime with Data and Analytics" (2013); and "From Data to Decisions III: Lessons from Early Analytic s Programs Lessons from Early Analytic s Programs" (2013).

Analyzing the Information

Analyzing the information has likely been the most neglected component of performance measurement systems in the United States and probably throughout the world. As noted in Section One, "data analytics," have recently gained much needed attention.

The previous section addressed the collection of information that performance measurement systems should provide. To give managers a much richer understanding of performance and to provide the wealth of information that can be obtained, the following activities are suggested:

A1: Assign staff to analyze the performance data.

To help managers get the most from the performance information they should assign one or more staff members to undertake basic analytic procedures such as those described below. The analysis should preferably be conducted soon after the data become available and before the performance report is publically disseminated. The data should be analyzed before performance review sessions (such as the data-driven reviews recommended below in Section Four) and whenever a need arises during the year, such as for queries from the legislative body or the media.

Even most small agencies should be able to identify someone who has an analytical bent to examine the performance data. Managers would be wise to do some of their own basic analysis to help them better understand performance and obtain clues to improving it. For small organizations, volunteers—perhaps those associated with nearby colleges or universities—might be available to help with some of the analysis.

A2: Compare findings to a variety of benchmarks.

Comparisons are a key step in making performance data meaningful. They are the name of the game in program evaluation and no less so for making performance measurement really useful for managers. Comparisons can be useful in: (a) interpreting the extent to which performance has been good or not so good; (b) identifying where corrective actions are needed; (c) helping determine which programs are working well and which not so well; and (d) determining whether outcomes have improved after corrective actions have been taken.

Outcome comparisons that can provide key information to managers include:

1. reporting periods;
2. targets;
3. customer characteristics;
4. service providers;
5. type and amount of services provided;
6. other agencies delivering similar services; and
7. level of customer difficulty.

The first two outcome comparisons listed above as well as comparisons to other similar agencies are commonly used. However, the remaining comparisons listed have been badly neglected in performance measurement systems. Each of the outcome comparisons is discussed briefly below.

- Compare the latest performance data to the results for one or more **previous reporting periods**. Identify unusual jumps or dives in the values of individual outcome indicators. This comparison is commonly used by agencies that have performance measurement systems. For quarterly or monthly reports, the comparison might be to the latest quarter or month, or might be to the same quarter or month in previous years.

  Comparisons made with multiple previous reporting periods will show time trends, which will help identify situations where changes in outcomes over time indicate particular patterns. Regular examination of time trends is less commonly done in performance measurement systems.

  When significant program changes occur, later outcome measurements can be examined to indicate whether these later outcomes are those expected from the program changes. (This is a form of "pre vs. post" analysis incorporated into many program evaluations.)

- Compare the latest performance data to **targets** that the agency or program may have set. Agencies that set targets for their performance indicators usually do so at the beginning of the year. If targets are set at frequent intervals, such as quarterly, comparisons can provide more timely information than if only annual targets are used.

  Target setting is something of an art and can be subject to political pressure. However, for internal management purposes, target setting is likely to be helpful in encouraging performance improvement—if the targets are realistic.

  > Tip: In some instances, setting targets for the next year will involve considerable uncertainties. In such cases, consider expressing targets as ranges rather than single values.

- Compare the disaggregated outcome data across the various **customer characteristics**, such as those characteristics identified in D9. For example, compare the outcomes for males to females; among various race/ethnicity groups; and among different geographical areas (zip codes, neighborhoods, cities, counties, regions, states or urban versus rural areas).

  Disaggregated outcome data helps "pinpoint" which demographic groups had desirable outcomes and which less than desirable outcomes. For example, a decrease might have occurred in the overall percentage of teenagers who dropped out of school from the last year to the current year, while the disaggregated data might show that dropouts had *increased* for one or more race/ethnicity groups. Such information provides a very different story than if only the overall data had been considered; attention was drawn to an issue that would not have occurred if only the aggregate data had been considered.

  > Outcome data disaggregated by demographic characteristics can be an important tool for examining who gets served and outcome *equity*.

  An agency can drill down deeper by comparing differences using more than one demographic or geographic category. For example, one might compare the percent of white males to the percent of white females achieving certain levels of outcomes for particular time periods and for specific geographical locations. Such steps are variations of what is sometimes called "data mining," "root cause analysis" or "drilling down." Emerging software is enabling most agencies, large and small, to easily and rapidly tabulate disaggregated data.

  For services that do not have human beings as the immediate customer, the appropriate units can be compared. For example, for water quality protection programs compare different bodies of water; for road maintenance programs compare road ride-ability for different road segments, perhaps based on location or on average daily traffic.

Disaggregated data can be tremendously powerful information, but we emphasize that such analysis does not identify WHY conditions worsened or improved.

- Compare the outcomes of different **service providers**. Examples of service units that might be compared include different regional offices; district offices; schools or school districts; fire stations; police precincts; correctional facilities; health care facilities; parks; road maintenance districts; libraries; sanitation crews; and social security and unemployment offices. Comparisons can also be made of external service providers, such as comparison of the outcomes of private contractors or NPO grantees.

  Comparisons can be made among individuals, such as individual caseworkers, teachers or doctors. However, it is not likely to be appropriate to include names of individuals in most reports.

- Comparing outcomes for customers who have received different **amounts and/or types of services** can be very informative for managers through identifying what procedures are working well or not so well. The comparison applies to programs for which the variation in service approach naturally occurs for their customers. For example, a program could compare the outcomes of customers who receive different types of casework, such as whether the service was delivered in individual or group sessions—or by in-person, telephone, or electronic means. In another example, managers will have stronger evidence that workers are safer in facilities that are inspected more frequently if safety data are tracked along with data on the frequency of inspections for each facility.

  > Tip: A variation on the amount-of-use has begun to be used, with customers termed "super users" and "frequent flyers." Customers are characterized by the amount of use they make of a service. Special analytical attention can be focused on frequent users who are generating major portions of the costs. These users can be examined to assess why they consume high levels of services and what actions might be taken to reduce this use, and thus significantly reduce service cost.
  >
  > For example, health care agencies such as hospitals are beginning to focus their attention on patients who are very heavy users of services. A small proportion of patients might, for example, have high hospital readmission rates. In one instance when an agency looked at those cases, it found that many of the super utilizers had developed pneumonia after hospital discharge, requiring readmission. An investigation found that this was likely due to poor heating/cooling in their homes, which then provided an opportunity to take measures to reduce readmissions.

Exhibit 2 is an example of a performance report showing comparisons by both demographic and service characteristics. (The latter include the amount of service provided and specific providers.) An analyst has highlighted the poor outcomes believed likely to warrant the program manager's attention.

> The South Carolina Department of Public Safety used its Incidence–Based Reporting System database to examine disproportionate minority contacts among juveniles in the state. It compared the contact rate for juveniles of various other race/ethnicities to white juveniles, an easy to calculate ratio. Ratios were calculated separately for each characteristic, including gender; county; category of offense (e.g., property, drugs, and violence level); and location of the crime (e.g., in a residence, school or a store).[14]

- Compare the latest results to the performance of **other agencies delivering similar services to similar customers**, although such data may be difficult to obtain. State governments have used such comparisons to compare themselves to other selected states, especially when the federal government has been able to obtain similarly defined indicators across states.

  For example, the State of Virginia has been using such comparisons on its public "Virginia Performs" website. For a number of its outcome indicators it has also provided outcome values over time for three other nearby states; the state with the best outcome numbers; and the national aggregated values.

- Compare outcomes by the **level of difficulty (risk)** in achieving desired outcomes. If customers have been characterized at or near intake by level of difficulty (as described in D9), outcomes can be compared by examining differences in outcomes between customers of different difficulty levels. Such analysis, for example, will help managers avoid missing an important reason for differences in the outcomes accomplished by different service providers. It will also help avoid incorrect judgments when interpreting current years' outcomes as compared to previous years' outcomes (if the mix of incoming workload contains substantially different proportions of difficult-to-help customers).

[14] "An Overview of Racial Disproportionality in Justice Arrests and Offenses in South Carolina," South Carolina Department of Public Safety's Office of Justice Programs, May 2012.

Exhibit 2

Sample Comparison of All Breakout Characteristics

Characteristic	Number of Clients	Clients That Reported Improved Functioning after Completing Group Therapy: Considerable Improvement (%)	Some Improvement (%)	Little Improvement (%)	No Improvement (%)
Gender					
Female	31	10	19	55	16
Male	43	30	40	21	7
Age Group					
21–30	13	23	31	31	15
31–39	28	21	32	36	11
40–49	24	21	29	38	13
50–59	9	22	33	33	11
Race/Ethnicity					
African-American	25	32	20	32	16
Asian	5	0	60	20	20
Hispanic	20	15	40	40	5
White/Caucasian	24	21	29	38	13
Sessions Attended					
1–2	13	15	8	54	23
3–4	21	24	33	33	10
5+	40	23	38	30	10
Facility					
Facility A	49	24	27	35	14
Facility B	25	16	40	36	8
Caseworker					
Therapist A	19	26	26	42	5
Therapist B	18	11	39	33	17
Therapist C	18	6	17	56	22
Therapist D	19	42	42	11	5
All Clients	**74**	**22**	**31**	**35**	**12**

Source: Hatry, Harry P., Jake Cowan and Michael Hendricks. *Analyzing Outcome Information Getting the Most from Data*. Washington, DC: The Urban Institute, 2004.

Sometimes different service units whose outcomes are worse than other units say in their defense: "My customers are more difficult than those served by the other units." Are such claims valid? Basic analysis can provide evidence that helps address this issue.

The key step is to define categories of levels of difficulty. As described in D9, these are based on customer demographic characteristics and the extent of the problems the incoming customer has. These definitions are then applied to each new customer, assigning the customer to a difficulty level. A supervisor or other staff member would select the difficulty level for each customer using the level definitions. Such categories would ideally be based in part on statistical analysis of factors relating to success. However, a program may believe it is sufficient for its own staff to establish difficulty-level definitions.[15]

> Pretrial Service Agencies (PSA), such as the PSA of the District of Columbia, assess the level of risk (a variant of level of difficulty) of each charged defendant in order for judges to determine whether or not the client should be released while awaiting trial and to help determine the conditions of release; for example, level of oversight the pretrial agency needs to take with defendants released while waiting for their trials. For this service, a special risk-assessment rating instrument is needed, one that considers a substantial number of defendant characteristics. For this application, because public safety is a concern, statistical expertise is needed to develop the risk assessment definitions.

The idea of level-of-difficulty or "risk adjusted" workload has considerable potential for the future. However, as in the case of pretrial risk assessments, it may require special efforts to develop an appropriate customer-difficulty rating scale.

Managers do not need to be deluged with all this information but should be able to link easily to disaggregated data when they believe they need it. Also, the analysts should provide alerts when they identify disaggregations they believe warrant the manager's attention.

[15] For more discussion of this procedure, see "Performance Measurement: Getting Results," Urban Institute, 2006, pages 131-132. An early application of customer difficulty was in vocational rehabilitation. See "Field Test of a Service Outcome Measurement Form: Client Change," Oklahoma Rehabilitation Agency, March 1975.

A3: Provide software that enables managers and their staff themselves to "drill down" to obtain performance information easily.

The software used by an agency should enable managers and their staff to review outcome data and obtain real-time reports at their desks (or wherever they might be). If performance-related data is entered throughout the year, managers will be able to obtain reasonably current information as issues arise. For example, a manager might need to check on the timeliness with which key eligibility information is being processed, perhaps because they are asked for such information by top management, the legislative body, or the media.

Many types of software are available at various price levels. Agencies should shoot for optimal capability within available resources.

A4: Identify explanations for unexpectedly low or high outcome levels. Categorize and tabulate these explanations.

As part of data examination, the analysts should seek to identify those factors that have contributed to unexpectedly low or high outcome levels. As noted in D11, as part of their performance measurement data collection procedures agencies sometimes have opportunities to obtain qualitative information that would help provide explanations for some outcomes. Such explanatory information can often provide useful information that helps the agency make improvements.

Explanations can be obtained in a variety of ways, as noted in Exhibit 1. As discussed earlier, they can be obtained from survey respondents, by trained observers, by administrative records (if the data collection procedures provide for such information), and by special interviews with staff and customers. The explanations can be categorized, tabulated, and provided to managers and their staff. For example, administrative record information (such as counts of traffic accidents) can be grouped and tallied (e.g., whether caused by a vehicle mechanical failure, driver mishandling, unclear signage, or bad weather). The performance measurement system would provide not only the aggregated data, such as the total number of traffic accidents, but also the number and percentage of accidents due to each type of cause.

Similarly, survey respondents' reasons for providing poor ratings to particular service characteristics might be examined and categorized. For example, the survey might ask respondents why they had not used a particular service during the previous 12 months. Responses might then be grouped into categories, including that the service was too far away; the location did not have enough parking spaces; the hours of operation were not convenient; and they had tried the service before and found the staff discourteous or unhelpful.

When substantial proportions of respondents give poor ratings and provide the same reason for doing so, this provides more guidance as to the remedial actions. Even a single response might trigger recognition of the need for a service modification among program staff.

The agency should also *consider distinguishing between explanations for outcomes where responsibility lies inside the agency or program and those where the responsibility lies outside.* Some factors affecting outcomes might be beyond the program's control, such as unusual weather conditions; unexpected economic problems; an incoming workload that contained a substantially higher proportion of difficult-to-help customers than expected; or unexpected budget cut backs. Internal operational factors may also be responsible, such as illness among key personnel or major computer system problems.

Survey questionnaires can also ask respondents to offer suggestions as to how the service might be improved. Such a request is usually placed at the end of the questionnaire. If the same improvement is suggested by a substantial number of respondents (e.g., "have different service

hours"), this indicates a potential route for improvement. Such suggestions represent implied explanations for service problems, and can be grouped by subject, collated, and tabulated for examination by the manager.

Categorizing statements from survey respondents who provide explanations for their ratings or who provide improvement suggestions does require extra work, especially if large numbers of respondents provide explanations or suggestions. Software programs such as NVivo can help somewhat in the categorization process but still require staff time.

A5: Take advantage of mapping software to examine relationships among different outcomes or among outcome data and other program-related information that might have important location relationships.

A technology development that is rapidly gaining widespread use is geographical mapping of outcome data. Maps can visually display how different segments of a geographical area vary on outcomes, including crime rates, fire calls, household income, rate of code violation, and graduation rates.

Mapping the location of particular events can considerably enhance the ability to identify patterns associated with geographical distribution of incidents. These can and have been used to examine the distribution of such incidents as location of crimes; origin of calls for emergency rescues; location of fires; and location of persons with particular diseases. The examination of such distributions will enable those examining the maps to identify specific geographical areas with a high density of such undesired outcomes. Such information can then be used to help identify where resources should be located; to begin an examination of why those particular areas have particular problems; and to enable managers to focus on correcting those problems.

Sometimes it is also useful to overlay maps to identify potentially important relationships. For example, maps that show the location of air or water pollution might also show the location of potential pollutant sources, such as industrial or agricultural premises—thereby helping to identify the source of the pollution. In another example, the base map might be overlaid with maps showing the distribution of incidents by time of day and day of the week or month (such as mapping of criminal activity) to help managers better allocate their resources.

> What is new here is that with considerable improvements in mapping software technology such analyses can now be done much more routinely, enabling managers to track problem areas on a regular basis. GIS and mapping software provide many such opportunities.

A6: When using an index also provide ready access to the values for the individual outcome indicators that comprise the index. Also provide information on the weights used to combine the indicators into the index.

Some officials are likely to prefer having a single composite number that identifies how a program is doing, rather than looking at the data on a set of outcome indicators. To do this, an analyst will need to assign weights to each individual outcome indicator that comprise the index to combine the values for the individual indicators into one index number.

However, indices used alone to make decisions can conceal very important information on what has occurred. To help managers make decisions, *the data on the individual components of the index should be presented along with the final index value*. If this becomes too unwieldy, the analysts should at least provide easy-to-use hyperlinks to the data on the index components. For example, an air pollution index is normally sufficient information for the public. However, for managers, knowing which pollutants are outside acceptable limits is vital information when taking action. In order to understand and enhance the ability to correct the air pollution, it is of course necessary to examine the various components to help identify the potential causes of that pollution, a first step toward taking action to reduce it.

For full transparency, it is also necessary to provide information on what the weights were and how they were derived, which will enable users of the index to better understand the its validity and perhaps even enable users to recompute the index using their own weights. For indices that are highly technical, such as air and water pollution indices, most users are not likely to be able to do much with the information and it is hoped that whoever set up the weighting process did it well.

A7: Consider using "mini" experiments to test alternative service delivery procedures and provide more rigorous evidence.

The key limitation of performance measurement systems is that while they track outcomes, they do not tell WHY those outcomes occurred. Too often the media and general public assume that if the outcomes are not good then the program should be blamed. This report has suggested ways that performance measurement systems can provide some explanations for these outcomes (see item A4). However, they are limited in their explanatory ability.

To obtain stronger evidence, randomized, controlled trials (RCTs) have been considered the "gold standard" approach. The key characteristic of RCTs is that participants are randomly chosen for placement into alternative ways to provide a service. Generally, RCTs have required specialized professional help and can be quite expensive. (Other advanced statistical tools are available for obtaining stronger evidence on "causality," known as "quasi-experimental designs." However, this report only addresses procedures that public service agencies can use without requiring advanced statistical procedures or highly trained analytical staff).

Believe it or not, even a small government agency or non-profit organization can sometimes use this "gold standard" approach without significant added expense—assuming the program already has a performance measurement system in place. For example, a program might want to decide whether casework for a particular service would be more effectively delivered using group sessions or one-on-one sessions, or if assistance could be provided electronically rather than in-person. The agency could randomly assign incoming clients to each of the approaches. The outcomes for each group of clients would then be calculated and compared to identify which procedure produced better outcomes.

Random assignment can be done very simply, such as by flipping a coin for each incoming customer; drawing numbers out of a hat; using a random numbers table on a computer or handheld device; or by assigning every other person or every third person to a different approach.

> One rare example of this experimental approach undertaken solely within a government agency was done by the city of Wilmington, Delaware many years ago. The city randomly assigned city-owned vehicles to two different vehicle maintenance strategies: preventive maintenance or maintenance-as-needed.
>
> A recent example from a much larger unit of government illustrates the use of RCTs in a low-cost way. The judicial agency that administers the collection of fines in the United Kingdom sponsored an RCT that compared newly proposed procedures for collecting fines. The unit randomly selected people to be contacted using a new procedure (texting through mobile phones) or using the past (non-texting) procedure. The amount obtained using each procedure was then compared. (The agency also compared a variation of text messaging approaches.) Each approach was examined as to the amount that subjects subsequently paid in fines and it was found that texting substantially increased the

amount paid. The work was carried out by a combination of agency and university personnel.[16]

Earlier, in 1996, the Minnesota Department of Revenue tested a potential new procedure to collect owed taxes. It randomly selected 100 taxpayers with unpaid liability to be sent a letter giving them an opportunity to resolve their liability prior to being billed. It then randomly selected another 100 taxpayers with liability to which a letter was not sent. The department then compared the percentage of taxpayers in each group that paid their liabilities. Approximately 75 percent of those sent a letter paid their liability; only 25 percent of those not sent a letter paid.[17]

In sum, when conducting an RCT experiment customers are randomly assigned to one of two or more alternative service delivery approaches. The outcomes for each group of customers are separately tabulated and compared. Such a procedure provides particularly strong evidence as to the relative effectiveness of each approach. (The procedure is likely to be especially appealing to managers who like to try new service delivery approaches.)

Exhibit 3 provides further suggestions as to the steps necessary in undertaking such small-scale experiments.

Undertaking these mini-experiments are primarily applicable to alternative service delivery practices for which: (a) the data on the important outcomes are available; (b) the major outcome data needed can be obtained in a relatively short timeframe, such as within a year or less; and (c) the alternative service delivery approaches are not complex.

While the test period is under way, the agency needs to attempt to ensure that no other outside factor occurs that would likely have a strong effect on the outcomes of the strategies being examined.

[16] Collection of Delinquent Fines: An Adaptive Randomized Trial to Assess the Effectiveness of Alternative Text Messages," Laura C. Haynes; Donald P. Green, Peter John, and David J. Torgerson, Journal of Policy Analysis and Management, Vol.32, No. 4, 718-730 (2013).

[17] "Making Results-Based State Government Work," Liner, Hatry, Vinson, Dusenbury, Bryant, and Snell Urban Institute, 2001, page 54.

Exhibit 3

Testing Alternative Service Delivery Approaches

1. Identify the service approaches to be compared. Typically, one would be the existing approach and the other a new one.

2. Choose a method for deciding which incoming clients will be served by the new approach. The method should be one that selects a representative sample of clients for each approach. Some form of random assignment is necessary. Randomization helps assure that the comparisons will be valid and greatly increases the strength of the evidence. (For example, with random assignment, approximately the same proportion of difficult to-help clients will likely be included in each group.) Methods of random assignment include flipping a coin and using a table of random numbers. Another method is to assign incoming participants alternatively to each practice.

3. As each client enters the program, assign the client to one of the two groups by the procedure the program has identified.

4. Record which clients are assigned to which service approach.

5. Track the outcomes for each client in each approach over whatever period of time the program believes is necessary to identify the outcomes of these approaches.

6. Tabulate the values on each outcome indicator for each approach.

7. Compare the findings and make adjustments to program practices, as appropriate. The program may want to drop the approach that shows the poorer outcome. Alternatively, it might decide that it has not yet obtained a clear enough picture from the outcome data to make a decision, in which case the program might continue the comparison for a longer time.

Source: Hatry, Harry P., Jake Cowan and Michael Hendricks. *Analyzing Outcome Information Getting the Most from Data*. Washington, DC: The Urban Institute, 2004.

Reporting the Performance Information

A8: Shortly after each performance report is produced, provide for its summarization and highlighting.

Performance reports can include substantial amounts of data. Managers and other officials should ask an analyst to examine the indicator values, summarize the findings, and identify highlights. Such highlights would identify unexpectedly poor and unexpectedly good performance values as well as identify trends. However, this straightforward step is usually neglected. A summary and highlight piece can greatly increase the use of the report. Recent "scorecard" summaries have begun to include arrows or other visual cues that enable readers to quickly identify performance indicators whose values are improving or worsening. Such approaches can be programmed into computers.

A9: Examine and report the extent of uncertainty in the measured outcome values.

Some outcome data might be subject to considerable uncertainty. The extent of uncertainty should be identified and brought to the attention of decision makers to help them interpret the data.

Instances where data may be highly uncertain can include very low customer survey response rates and findings based on old data. For customer surveys, *statistical significance* levels are usually available and are sometimes reported. Such information should be available to those using the data. However, users should also be aware that *substantive significance* is probably a more important consideration. That is, for the particular indicator being measured, how much certainty is needed to affect decisions? For example, if the difference in satisfaction level between two years has slipped two percentage points, say from 91 percent to 89 percent, from the previous reporting period (which could be statistically significant if the sample size was large enough), how much attention does this small difference warrant?

A10: Make sure reports are clear, understandable, and meaningful—internal as well as external reports.

It may seem obvious but too often performance reports have incomplete and unclear, labels; are over-crowded; or are in overly small, difficult to read fonts. Current technology enables reports to be highly attractive and readable. It is relatively easy for data reports to contain a variety of graphics, such as bar charts, line charts, and maps in addition to tables. Color can also liven up the reports. "Data visualization" is becoming a hot topic, and a number of resources are becoming available that can provide good ideas for making reports look attractive.[18]

However, there is the danger that report developers will be tempted to go overboard when using graphics and provide overly complex communications, rather than concise and easy to understand ones. Exhibit 6 (discussed below under U9) is an example of presenting information attractively while conveying information clearly.

A special challenge and need here—though likely often neglected—is to not only make performance reports for external use attractive and informative but *also to make internal reports that managers use throughout the year clear, understandable, and informative*. Attractive, user-friendly reports will be more useful, and used, by both internal and external audiences.

A common problem is not providing information on the timing of the data reported. What time period is covered by the data being reported for each indicator? For some indicators it might have taken many months to assemble the data. A substantial lag time can occur between collection and reporting, particularly for data reported by federal and state agencies.

[18] For example, see "Presenting Data Effectively," Stephanie D.H. Evergreen, Sage, Thousand Oaks, CA, 2014. For a more analytical perspective, see "Now You See It: Simple Visualization Techniques for Quantitative Analysis," Stephen Few, Analytics Press, Oakland, CA, 2009.

A11: Incorporate relevant findings from completed program evaluations into performance measurement reports.

The above mentioned recommendations by no means remove the need for the use of program evaluations to provide a considerably deeper look into not only the outcomes of a program but also the explanations for those outcomes. Program evaluations can provide considerably more in-depth information on a program.

The manager's analyst should identify any relevant completed program evaluations and examine the extent to which the findings agree or disagree with the messages that the performance measurement system data collected. If the program evaluation findings paint a picture considerably different from the performance measurement system data, it should be identified and discussed in the performance report. At a minimum, the performance report should summarize the relevant findings, perhaps as an attachment to the performance report.

Section Four

Using the Information to Improve Services

Now that the right data and related performance information has been collected (Section Two) and analyzed (Section Three), the information needs to be used to help continually improve services.

Even the best data derived using top-notch analyses and provided to potential users will be wasted if the information is not used effectively. The following section discusses a number of ways to use performance information so as to increase its usefulness.

Performance information from performance measurement systems are also used to achieve external accountability. That is important but is not the focus of this report. Instead, we focus on performance management and using performance information to improve services to make them as effective and efficient as possible.

The first set of recommendations focuses on some of the basic uses for performance information. The second set contains recommendations for making those uses more effective.

BASIC USES FOR PERFORMANCE INFORMATION

U1: Hold regular data-driven performance reviews with staff, using data from the performance measurement system as a starting point for such meetings.

Data-driven reviews have sometimes been called "PerformanceStat (STAT)" or "How Are We Doing?" sessions. These in-person performance reviews should be undertaken on a regular basis, such as after each scheduled performance report has been prepared. At these sessions a manager meets with staff to identify and discuss where performance has been better and/or worse than expected. The participants discuss why this is the case and what actions are needed, by whom, and by when. A starting point for each review session could be the information provided by analytical staff to summarize and highlight the latest performance report (as suggested in Section Three).

> The PerformanceStat approach began in 1994 when the New York City Police Department introduced its CompStat program. It was then picked up by the City of Baltimore and called CitiStat, and covered all city departments. A number of local governments have implemented their own versions of PerformanceStat. The White House Office of Management and Budget introduced the concept in 2010, labeling the effort "data-driven performance reviews." Under the GPRA of 2010, each major federal agency is required to hold quarterly sessions on the agency's priority goals.[1] At the state government level, this approach has been started with initial variations in the states of Washington and Maryland. Thus far, the approach does not appear to have reached many non-governmental organizations (NGO's), although it appears to have large potential rewards for them.

Performance reviews typically focus on the work of individual departments or programs. Alternatively, they might focus on cross-cutting themes, such as juvenile delinquency or homelessness. The latter meetings would involve individuals from a number of agencies. For example, reducing homelessness would require substantial involvement from a government's housing program, public safety agencies, social services, and health services. A jurisdiction might use a combination of both approaches.

Such meetings can be made more productive if the STAT analysts have already examined the performance data to help identify the issues that the meeting leader will address, using such means as those discussed in Section Three. Moreover, each participating agency should be encouraged to ask its own analysts to do the same so that meeting participants are able to recognize what is happening and why.

Such data review approaches might involve *multiple sector* approaches, as is being explored by New York State as part of its COReSTAT pilot effort to help distressed neighborhoods in that state. COReSTAT meetings are expected to include representatives from a number of different state-level departments and representatives from a number of agencies in local communities. These theme and cross-sector efforts are likely to be more effective if the sponsoring government has first gained experience with the use of performance reviews within individual agencies.

The process is a potentially terrific approach with little downside. However, weak implementation can adversely affect its usefulness. Practices to avoid include:

- Not having the active participation of senior officials (engaged leadership is essential);
- Not holding regular sessions;
- Making the performance review sessions overly negative in tone;
- Not identifying corrective actions; and
- Not following up on the results of those actions in later sessions.[19]

[19] The literature on the PerformanceStat has expanded rapidly in recent years. Among these are many papers by Professor Robert Behn of Harvard University and his book: "The PerformanceStat Potential: A Leadership Strategy for Producing Results," Robert D. Behn, Brookings Institution Press, Washington, DC, 2014. More recommendations on the procedures for such reviews are provided in "A Guide to Data-Driven Performance Reviews," Harry Hatry and Elizabeth Davies, IBM Center for the Business of Government, Washington, DC, 2011.

U2: Encourage use of regular data-driven performance reviews at lower levels of management, not only at the top levels of an organization.

Thus far, at all three levels of government (federal, state, and local) most applications of data-driven reviews appear to have been undertaken at the top level of the agency, either by the agency head or by the head of the government, such as the mayor or governor.

> As with most recommendations in this report, the basic data-driven performance review approach can be readily downsized to any level of an organization, including use by first-line supervisors.

Managers at lower levels in an agency should be encouraged to hold their own sessions, and not merely as rehearsals for higher-level meetings. These managers should consider using such reviews to help them improve their services. The basic concepts of data-driven performance review ideas are straightforward. Each program manager should consider holding regularly scheduled meetings with their staff, using the latest performance data as a starting point, to: (a) discuss where performance data indicate weak or strong results; (b) identify required actions to improve performance; and (c) in subsequent meetings, review the progress of those actions.

Performance reviews do not need elaborate production or extensive technological support. Nor do these reviews need extensive agendas and special analysis staff to guide the manager in identifying where problems appear to exist. These elements are fine if resources permit but are not essential. A primary building block for regular performance reviews is that the program has an ongoing performance measurement process in place that provides performance data.

> Regular performance reviews have the substantial added benefit of encouraging staff to continually focus on improving service delivery to achieve better outcomes.

U3: Use performance measurement system outcome data as a major basis for developing and justifying policy choices, including budgets and strategic plans.

Performance measurement information should be one of the starting points when managers formulate budgets, and can be useful when subsequently justifying budget requests. Similarly, past performance data should provide a major starting point for developing strategic plans. The data also provide a starting point for estimating the *extent of the budget year and out-year needs* to be considered in making choices.

In addition, historical performance data, along with cost data, can be used to help estimate *target values* for outcomes achievable through the budget or strategic plan. For budgeting, future-year projections are only needed for one or two years. For strategic planning, the projections typically need to be made for three to five years.

> Major uses of performance information include providing managers with evidence as to program and policy needs and problems; and providing past data for use as starting points for estimating future budget needs and likely benefits achievable with budgeted amounts.

Many government agencies are already doing some form of "performance-informed budgeting," if only by including output and outcome data in their budget documents. However, it has been difficult to find examples of agencies in which outcome data has played a significant role in the formulation of budgets or strategic plans.

In part this is because many organizations have not had much relevant outcome data. Also, it is often very difficult to determine a direct relationship between outcome values and costs. For example, how much will it cost to increase from 76 percent to 80 percent, the percentage of students who graduate from high school? At present and for the foreseeable future, we do not have formulas for most such estimates.

Nevertheless, the process of thinking through how better outcomes might be achieved can itself be very helpful in formulating and subsequently justifying proposed budgets or strategic plans.

For some outcome indicators, however, particularly "intermediate outcomes," and for many outputs, reasonable budget estimates can be made. For example, the budget consequences of reducing response times to citizen service requests by a specific amount can be reasonably estimated (assuming that the cost data available is reasonably accurate). The costs to reduce police (or fire) response time by a certain amount can be reasonably estimated given specific information on how the reduction is to be made. The relationship of the reduced response time to

amount of crime reduction or citizens' feeling of security (or to the amount of fire property damage reduced) is much less clear.

Exhibits 4 and 5 illustrate the use of performance information for making policy and program choices. Exhibit 4 illustrates how performance data can help inform program choices. Exhibit 5 is an example of using data from the performance measurement system to compare the number of incidents of youth contacts with the criminal justice system by age of the youth over time. The information shown in the exhibit can be used to help make a strong case for formulating a policy of focusing crime-prevention interventions on high-risk youth before they reach a particular age.

> Budgeting and strategic planning are about the future.
> Performance measurement is about the past.

Exhibit 4

Sample Two-Characteristic Breakout

Percent of Clients Employed Three Months after Completing Service				
Education Level at Entry	N	Short Program	Long Program	Total
Completed high school	100	62% employed (of 55 clients)	64% employed (of 45 clients)	63% (of 100 clients)
Did not complete high school	180	26% employed (of 95 clients)	73% employed (of 85 clients)	48% (of 180 clients)
Total	280	39% (of 150 clients)	70% (of 130 clients)	54% (of 280 clients)

Is action needed? Encourage clients who had not—rather than had—completed high school to attend the long program. Use these figures to help convince clients of the longer program's success with helping clients secure employment.

Source: Hatry, Harry P., Jake Cowan and Michael Hendricks. *Analyzing Outcome Information Getting the Most from Data*. Washington, DC: The Urban Institute, 2004.

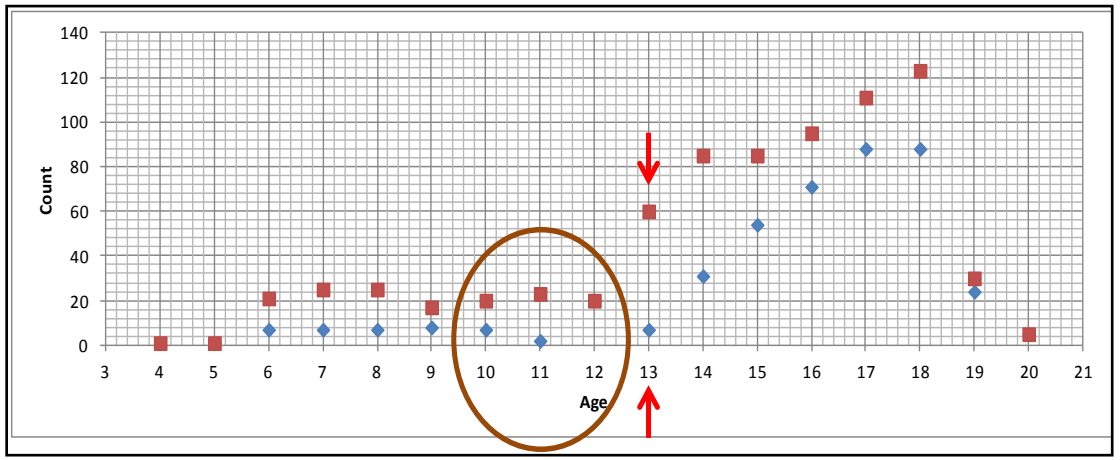
While we have tools for measuring past outcomes, only much weaker tools are available for estimating future outcome values, given the many unknowns. Statistical projections (useful for projecting future revenues given certain population and other assumptions) and simulation modeling can be used for some important estimates. Regardless, uncertainties in estimating the future, even if only for one or two budget years, can be considerable.

Even in the 21st century, unless a time machine is invented, it is hard to envision major improvements in the accuracy of estimates about the future. Simulation modeling encouraged by continuous technological developments will likely become more widely used–enabling managers to better estimate future effects given various assumptions, such as about economic and climate conditions. It is, therefore, likely to be more useful to provide cost and outcome estimates as ranges or alternative values based on what external factors actually occur, which would provide a set of alternative performance target values against which to compare actual indicator values.

For both strategic planning and budgeting it is also important to do an "environmental scan" to look for likely changes in key factors that can significantly affect outcomes as well as costs. For example, programs addressing substance abuse need to look for trends in the particular drugs being abused and in the composition of likely abusers, including demographic characteristics (such as age) that can affect the use of various drugs. These factors can have important implications for the costs and outcomes relating to the proposed budget. The scan for budgeting focuses on near future factors that would affect program costs and outcomes; for strategic planning, the scan should address longer term factors.

Exhibit 5

Use for Making Policy Choices



Red Icons represent the number of those with one and only one interaction with the police department. Blue icons represent the number with more than one interaction

Figure at left shows all interactions of youth, those with high absenteeism rates, with the police department including victims and witnesses. Overall interaction jumps at age 13 as does absence rates. This indicates the last opportunities for prevention activities before actual arrest escalation.

Source: AJW, Inc.

U4: Use performance information to motivate employees and contractors—but with caution.

Both nonmonetary and monetary awards can be used to help motivate employees (and contractors) to perform better. Monetary incentive programs are an attractive concept to some public officials. However, such programs can also lead to dissatisfaction if not implemented carefully, especially regarding the fairness of monetary awards. Pay-for-performance plans are used in a number of private business occupations, such as those that involve selling goods or services. In the public sector, however, many pitfalls exist, especially if the award is money. The jury is still out as to whether, and for which services, monetary incentives or other high-stakes consequences are worthwhile.

> The 2014 Department of Veterans Affairs scandal illustrates the potential negative motivational effects of overemphasizing performance results in a threatening environment—leading to employees falsifying waiting-list data.

In the past, performance awards in the public sector have typically been based on supervisor judgments. Perceptions of favoritism or of bad judgment often arise among those not happy with award decisions. *A major advantage of using data from the agency's performance measurement system to help determine awards is that it makes award decisions more objective*. However, monetary awards based on performance data are likely to be successful only if the data are used in a clear, systematic manner that is accepted by employees and avoids major overrides by supervisors. As important outcome indicators are generally produced by multiple employees, it may be advisable to give monetary awards to teams or even whole agencies rather than to individuals.

Although likely to have less motivational value, nonmonetary awards (e.g., recognition awards and small perquisites like a desirable parking space for a month) face fewer problems. One advantage over monetary awards is that they do not cost money; in this age of highly limited public and nonprofit budgets, this is a considerable advantage.

Private nonprofit organizations and even local governments agencies might be able to use nonmonetary awards donated by business organizations in their communities, including coupons for free nonalcoholic drinks, a free food item, a movie, etc.

Using performance data as a factor in monetary award programs will likely need to focus on outputs and intermediate outcome indicators (rather than on end outcome indicators). These data are more closely linked to the work that is the responsibility of the employee. Care needs to be taken so that focusing on selected indicators as part of the award criteria is not done at the expense of other indicators.

U5: Use performance information as the basis of reports to persons outside the organization, such as legislative bodies, the media, and citizens.

Getting the legislative body, the media, and citizens on your side is highly valuable. Use the performance information to help tell the story from your perspective. A challenge for managers is that the performance measurement system will sometimes provide evidence of unexpectedly poor outcomes and result in them being blamed, whether fairly or unfairly, for those negative results.

> Tip: In this time of increased transparency, it is highly likely that poor results will surface in one way or another. It is advisable for the agency to be the one to report the bad news. The agency can mitigate the negatives if at the same time it provides (credible) explanations as to why the poor results occurred. (This could come from the explanatory elements of the performance measurement system discussed above.) Adverse publicity will likely be mitigated further if the agency provides its plans for correcting the problem simultaneously with release of the negative data.

Visual display techniques have improved tremendously in recent years. If chosen wisely, very attractive performance report presentations can be provided for each of the various stakeholder groups. The presentations and the data should be timely. Presentations should be shown well, updated regularly, and be readily accessible by the various stakeholder groups.

U6: Link up with other agencies to tackle complex issues involving multiple agencies and programs. Enable performance data sharing across agencies while protecting confidentiality.

As noted in D12, the public management community is increasingly addressing issues where solutions for complex problems require coordinated actions among multiple agencies. Coordination of performance measurement across programs and agencies is likely to become an increasingly important element for all levels of government and for private nonprofit organizations. The "Collective Impact" movement is a version of this approach.[20]

Many public issues involve multiple programs, multiple agencies, and even multiple sectors. Combating juvenile delinquency and promoting effective economic development are just two of the many examples of challenges governments address on a regular basis. To combat complex problems it is often necessary to link data on the amount and type of services provided to *individual* customers in order to outcome information for each of those customers. Such linking can be important in determining which services work, which do not, and for which particular types of customers. More than one agency program is likely to have relevant data on the outcomes for individual customers.

> Programs to help disadvantaged youth such as the Harlem Children's Zone and Promise Neighborhoods involve many services and agencies. The program objectives can be numerous, including educational (e.g., improved learning achievement), economic (e.g., jobs creation), and social (e.g., strengthening families).
>
> Another example is provided in Exhibit 5, which illustrates the need for cross-agency data coordination. In this example both the school system and criminal justice system need to provide data and to work out confidentiality issues. Tracking the success of individual customers can require regularly measuring outcomes that come from two or more different agencies (e.g., the school system, social services, and criminal justice agencies).

Two major obstacles have limited the ability of the public sector to make cross-agency linkages: technology and confidentiality concerns. The first concern, lack of appropriate technology, is easing. However, confidentiality remains a major problem. Public agencies are sensitive to citizens' concerns that data on themselves, and especially data on their children, should be kept private. Some federal and state legislation has imposed strict requirements to protect certain data. For example, congress has enacted the Health Insurance Portability and Accountability Act (HIPPA) relating to health programs, and the Family Educational Rights and Privacy Act (FERPA) relating to education programs, which constrain the release of data on individuals.

[20] See, for example, "Channeling Change: Making Collective Impact Work," Fay Hanleybrown, John Kania, & Mark Kramer, in Stanford Social Innovation Review, January 26.2012.

> San Francisco developed a "Shared Youth Database" to enable tracking in real time of at-risk youth through the city's health, child welfare, and juvenile probation systems. It enabled the city to identify at-risk youth, assess their needs, and identify early what services might be used to divert them from future problems. When attempting to expand the program the city attorney halted the plan until concerns over confidentiality could be resolved.[21]

As noted in Section Two, the "Big Data" movement can help agencies incorporate and analyze data from multiple sources if issues of confidentiality can be worked out. In doing so, it will better enable agencies to identify patterns of services relevant to particular customer demographic groups that are associated with successful or unsuccessful outcomes.

[21] "Getting Big Data to the Good Guys," Stephen Goldsmith and Christopher Kingsley, April 9, 2013, Data – Smart Ci-ty Solutions newsletter, Ash Center, Harvard Kennedy Center.

MAKING THESE USES MORE EFFECTIVE

U7: Enable managers and staff to access timely, up-to-date performance data at any time during the year.

Many current performance measurement systems primarily provide information on program outcomes simultaneously with regularly scheduled performance reports, even if only on a quarterly or annual basis. However, managers need to get performance feedback throughout the year to help them identify the need for corrective actions as soon as possible after problems arise. In addition, managers need the capability to quickly obtain performance information at any time during the year and to make mid-course corrections. Important events can occur throughout the year; circumstances change. Managers and their staff are periodically called upon to respond to higher-level officials and the media. The ability to respond quickly and in an accurate, reliable way with relevant performance data is highly desirable.

> Throughout the year, managers need to have the latest performance information available to them.

Unfortunately, many current performance measurement efforts are primarily focused on year-end reports, often prepared primarily for accountability purposes. Such information has limited use for managing programs.

For some performance indicators new raw data will come in at various times throughout the year and can be integrated into the database so as to yield continually updated performance indicator values. For example, data can be made available at any time on the number of incidents that have occurred thus far in the year for a number of services (e.g., the number of crimes, fires, traffic accidents, and child abuse cases).

Current data can potentially be updated at any time on response times, such as for customer calls for a service (e.g., calls for assistance to Social Security or Internal Revenue Service offices). To the extent that basic information, such as the time of receipt and time the request was resolved, is recorded and entered electronically into an organization's database, response times and updated performance indicators (e.g., average response times, median response times, and number of requests that took more than 24-hours to resolve) can be calculated and made available for use by managers.

Fortunately, modern IT technology is making it increasingly possible for a manager or staff members to quickly pull up the latest available data on individual performance indicators. The manager might be provided with an interactive dashboard that would enable retrieval of the latest values for any performance indicator at any time.

As part of the performance measurement system, procedures should be designed so that managers themselves can readily extract the latest performance indicator values and important breakout values. For example, a manager might want to identify and compare the recidivism rates for certain age and race/ethnicity groups for a particular time period. Such systems are not widely available at present, but technology should enable them to become so in the near future.

Another problem: Some performance data may not become available to managers for many months, if not years, especially some federal and state performance data. It might occur, for example, because large data sets are involved and final calculations need to wait until the last pieces of data come in. The availability of decennial census data is a classic example of such a delay, but this problem occurs with many large federal data sets held until the data can be "cleaned."

Tip: An agency can provide more timely feedback by breaking up large samples currently administered once a year into smaller samples administered more frequently. For example, an agency might be sampling 500 customers each April. It could instead sample 125 customers at 3-month intervals. (Even sample sizes as small as 100 can provide reasonably accurate information if the sample is representative of the customer population and the extent of the uncertainty is understood.) Users of the information need to be alerted to the level of uncertainty, as with any survey data.

In doing so, an agency would provide managers with quarterly, seasonal data. The information from each quarter can be aggregated so that by the end of the year, the combined data from the four quarters would provide more accuracy. The added cost, if any, of the more frequent sampling is likely to be small.

Current technology should help by reducing the turnaround time for data processing and reporting. Another option available for some indicators is not to wait for the last pieces of data to arrive. At least for internal management purposes, *determine if it is sufficient to provide preliminary and/or incomplete findings to managers*.

An example: Local hospitals in the South Central Pennsylvania Alliance were concerned about the age of hospital readmission data: "Recognizing that the data reported can be as much as 12-18 months old, the Alliance worked with the local hospitals to create a system to provide ambulatory care and extended care providers with much more timely data on preventable admissions to make it more feasible for providers to prevent further readmissions." ("Lessons Learned in Public Reporting," a brief prepared by the Center for Health Care Quality, George Washington University Medical Center School of Public Health and Health Services, Washington DC, May 2011.)

U8: Provide relevant performance information regularly to first-line staff, not just supervisory personnel.

Providing first-line staff with regular information relevant to the performance of their work is likely to be both of interest and have motivational value to many public employees. Moreover, it can assist first-line staff in identifying problems and encourage them to devise solutions.

As with managers and supervisors, the data relevant to the worker should be readily and easily available, such as on their computer or hand-held devices so they can access the latest performance data at any time. The data provided should typically include the performance information on their team, their whole program, and even higher-level outcomes such as the performance data on the entire agency. Data on each individual's own performance should probably be provided only to that individual and supervisors.

Front-line staff is likely to need some training on interpreting the data and how they might use data relevant to their work. The intent here is not to transform them into analysts, rather it is to inform them as to what the data mean so they can do their jobs better.

U9: When making decisions, consider not only the aggregated data but also disaggregated outcome data, categorized by key customer and service characteristics.

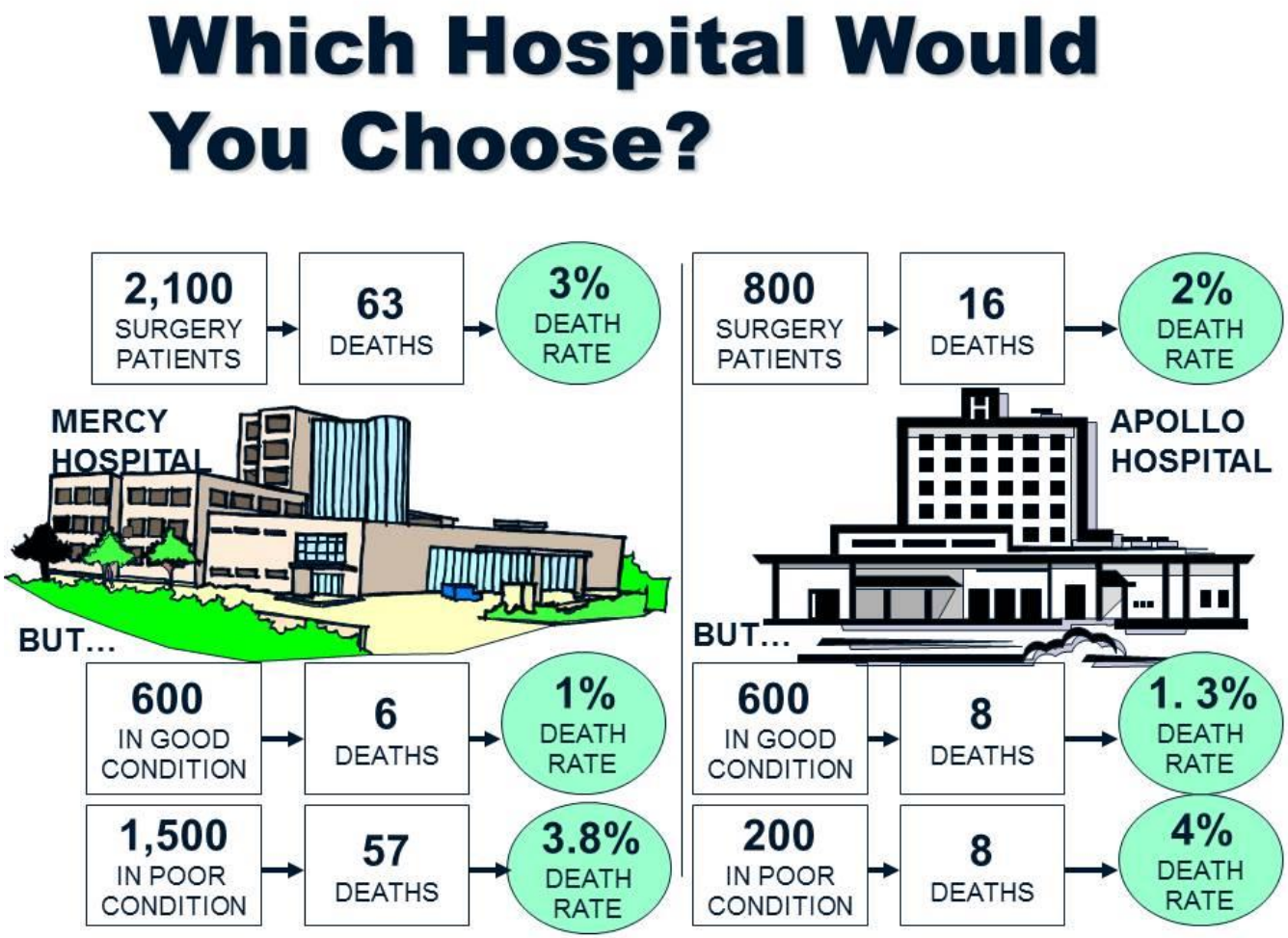
As discussed in A2, disaggregated outcome data can provide highly useful information, such as detecting where problems exist, identifying problem providers, and highlighting promising practices. Although it may seem an obvious use of performance data, many agencies have not taken advantage of their performance measurement systems to make such comparisons. One reason for this is the concern that it may lead to destructive competition or that the comparisons will not be fair.

How comparison findings are used is key to making them constructive and not destructive. If the approach is to use such comparisons to blame and punish rather than for improvement and learning opportunities, problems can arise.

Exhibit 6 illustrates the use of disaggregating the performance values by key characteristics. In this example outcomes are disaggregated by service provider (different hospitals), and by patients' condition at entry (a proxy for "client difficulty"). It illustrates the danger of not disaggregating outcome data by such characteristics. Disaggregating the performance data can help managers avoid poor decisions. Here the choice of hospitals changes when, instead of relying solely on aggregate outcome data, a key patient characteristic is also considered: patients' condition at entry.

Formal reports would contain the disaggregated data in only selected circumstances to avoid overload of information. However, managers should be able to readily access the disaggregated data when needed.

Exhibit 6


Which Hospital Would You Choose?
2,100 SURGERY PATIENTS
63 DEATHS
3% DEATH RATE
MERCY HOSPITAL
BUT…
600 IN GOOD CONDITION
6 DEATHS
1% DEATH RATE
1,500 IN POOR CONDITION
57 DEATHS
3.8% DEATH RATE
800 SURGERY PATIENTS
16 DEATHS
2% DEATH RATE
APOLLO HOSPITAL
BUT…
600 IN GOOD CONDITION
8 DEATHS
1. 3% DEATH RATE
200 IN POOR CONDITION
8 DEATHS
4% DEATH RATE

U10: Before making decisions, such as on data indicating unexpectedly poor performance, consider not only the performance data but also explanatory information.

It may seem obvious that managers would consider not only performance data but also explanatory information. However, in many instances this does not appear to happen, at least not as a standard, systematic activity. The GPRA legislation (both the GPRA Modernization Act of 2010 and the original GPRA of 1993) requires agencies to identify why performance goals were not met in their annual performance reports. However, it is unclear that in practice such explanations have actually been provided. More often, accompanying material discusses factors that *can* affect the program's outcomes, but this does not explain why the particular outcome or output deficiency occurred.

Section Two (D11) recommended that agencies require that explanations be provided for unexpectedly poor or unexpectedly good values as part of their performance measurement systems. This would enable the agency and program to determine what actions need to be taken. In some cases, the information can be used by the agency to identify specific actions that can and should be done quickly to correct observed problems.

> For example, in New York City work crews are dispatched to correct street cleanliness problems when the city's trained observers identify specific problems or rate particular locations as "poor." Trained observer ratings might, for example, provide supplemental information identifying specific hazards such as broken glass that need quick correction. Such information can be transmitted in real-time to generate work orders.

U11. Make use of information from the performance measurement system in in-depth, ad hoc studies, such as program evaluations.

Performance measurement systems can provide information that is useful for special studies, such as program evaluations. Providing outcomes disaggregated by customer demographic characteristics and by type and amount of service can provide an excellent starting point for an evaluation. For example, the Federal Highway, Railroad, and Aviation Administrations examine in depth the reasons for accidents so that interventions can be designed to address them. Performance measurement systems that routinely track accidents by the circumstances under which they occurred provide helpful information for decision makers.

Furthermore, the information obtained from qualitative explanatory information included in the performance measurement process (as discussed above in A4) can provide added starter information for in-depth evaluations.

Evaluators might also be able to *use the mechanics of the performance measurement system to provide new information needed by the evaluators*. For example, evaluators might add questions to an existing regularly administered customer survey that would provide needed information for the evaluation.

> Performance measurement is the "entry-level" evaluation approach. It provides basic data on progress in achieving outcomes but is highly limited for determining why those outcomes occurred.

U12: Provide training, technical assistance, and/or mentoring to managers and their staff in accessing, interpreting, and using performance information.

Even the best data and analysis will not be of much use if managers and their staff do not know how to use the information or how to readily access it. The interpretation and use of performance data is not every manager's cup of tea; many managers and staff members may be "people-people" and not "data-people."

Some managers and staff will have little background in interpreting performance data. Many will have significant trouble accessing the data they want, such as making use of the software to generate special tabulations of outcomes by various customer and service characteristics. Genuinely user-friendly software is yet to be widely available. Training, technical assistance, and/or mentoring will likely be needed to enable managers and staff to feel comfortable accessing, interpreting, and then using the available performance data.

U13: Track improvements that have resulted, at least in part, from use of the performance measurement system. Periodically evaluate your performance measurement and performance management systems.

A major gap in performance measurement and performance management is the lack of evidence that the benefits warrant the time and cost involved. Public agencies should periodically undertake evaluations of their systems, including seeking evidence of resulting beneficial improvements. This will help persuade public officials and the general public of their value.

Such evaluations of the use of performance measurement information need not be an arduous task. However, we do not know of any public agencies that are reporting such information on a regular basis. Agencies might ask their program managers to keep track of *program and policy changes* that were triggered by the performance data. Another source is the summary notes from any data-driven performance reviews undertaken within the agency, which presumably track actions taken after the performance review sessions.

It is more difficult to obtain information on *changes to outcome data* that can be at least partly attributed to the performance data. In some instances this may be relatively easy, especially if the data led to changes in intermediate outcomes such as response times to customer requests. Attributing outcomes, especially end outcomes, may not be feasible in many cases.

Agencies should also periodically evaluate their performance measurement and performance management processes. This would enable a public agency to identify weaknesses and encourage continuous improvement in those systems. The assessment would include examples of service improvements triggered by these systems, providing evidence to public officials and the public of their value.

Section Five

Implementation Issues

Most of the following recommendations are not likely to require significant expenditure if the organization has a basic performance measurement process and IT capability. Most expenditures are likely to be for start-up costs; on-going implementation costs will likely be minor. Some government agencies are already undertaking at least some of the steps recommended below.

The major added cost of these recommendations is that for undertaking regular surveys of customers after they have left services, as described in D5. Without such feedback from customers, many programs will have to settle for proxy indicators. While data for the proxy indicators may be readily collected, the information is likely to fall far short of providing information on highly important outcomes. For many human service programs this may be a brand new activity; suggestions on a number of ways to keep surveying costs down to an acceptable level are also described in D13.

Another potential issue is that of obtaining outcome data from other programs or agencies. As briefly discussed above in U6, barriers need to be overcome in order to protect the confidentiality of citizens.

Recruiting talented analysts to make life considerably easier for managers, as discussed above in Section Two (A1), can be a problem for some organizations. However, such talented individuals can likely be found in most public service agencies and even NPOs.

Another concern is the availability of software and training to facilitate the use of performance information, especially for undertaking the performance indicator disaggregations that provide considerably enhanced information for managers. As noted above, low-cost software is becoming available to enable such information to be readily generated. Nevertheless, for organizations with limited IT resources, introducing and keeping up-to-date the software needed to make full use of the performance data can be a challenge.

FINAL NOTE

An agency's performance measurement system has great potential to help managers manage for results. In the 20th century, considerable progress was made in gaining widespread acceptance for performance measurement. However, its potential for providing really useful information for performance management remains largely untapped.

The information obtainable from a performance measurement system can, and should, provide managers with considerably better, more rigorous, and more useful information for improving services to the public. The recommendations included in this report attempt to use the emerging technology and basic ideas from the program evaluation world to provide public managers with more credible and useful information to fulfill that potential.

The recommendations are meant to be used selectively and appropriately by individual public service organizations. Many governments and nonprofit organizations will likely have already implemented some version of at least some of the recommended actions.

Moreover, the recommendations in this report are not expected to be the last word. New experiences, innovative thinking, and technology development will inevitably mean that further improvements in performance measurement and performance management systems will occur.[22] The long-range goal of both performance measurement and performance management is to continually lead toward improved, more effective services for citizens in the 21st century.

[22] Two worthwhile readings that broadly address many of these same points in highly readable style are: Jonathan Walters, "Measuring Up 2.0," Governing Books, Washington, DC, 2007; and Mario Morino, "Leap of Reason," Venture Philanthropy Partners, Washington, DC, 2011.

Appendix

List of Recommendations

Recommendations on What Information Should be Collected and Data Collection Procedures

What Performance Information Should Be Collected?

D1: Seek input from representatives of key stakeholder groups to ensure identification of appropriate outcomes.

D2: First, consider indicators for which you already collect data. Then, consider indicators needed to measure outcomes for which you do not currently have data.

D3: Use "logic models" ("outcome sequence charts") as a tool to help identify performance indicators.

D4: Include outputs, intermediate outcome, and end outcome indicators in performance measurement systems—and distinguish which is which.

D5: Collect outcome data that identify the outcome at a time AFTER the customer has completed the program's services.

D6: Include indicators for outcomes over which you have some responsibility even though only limited control. Perhaps identify in external performance reports those indicators over which your organization has only highly limited influence.

D7: Track efficiency but do not settle for output efficiency. Focus when possible more on outcome efficiency: measuring efficiency in producing outcomes.

D8: Include "milestone" indicators for programs with substantial start-up steps.

D9: Provide key disaggregations of performance data routinely by major customer characteristics, such as by demographic and risk characteristics.

D10: Provide key disaggregations of performance data routinely by major service characteristics, such as by provider and by service type and amount

D11: Require agency programs to provide explanations for unexpectedly poor and unexpectedly good outcomes as a standard part of the performance measurement system.

Data collection issues

D12: Identify opportunities to obtain data from other programs or other agencies.

D13: Regularly survey your customers and former customers to obtain key outcome information.

D14: Avoid excessive precision and rigor when designing data collection.

RECOMMENDATIONS ON ANALYZING PERFORMANCE MEASUREMENT INFORMATION

Analyzing the information

A1: Assign staff to analyze the performance data.

A2: Compare findings to a variety of benchmarks.

A3: Provide software that enables mangers and their staff themselves to "drill down" to obtain performance information easily.

A4: Identify explanations for unexpectedly low or high outcome levels. Categorize and tabulate these explanations.

A5: Take advantage of mapping software to examine relationships among different outcomes or among outcome data and other program-related information that might have important location relationships.

A6: When an index is used to combine the values for several outcome indicators into a single consolidated number, also provide index users ready access to the values for each of the individual outcome indicators that comprise the index and provide information on how the weights used to combine the indicators were derived.

A7: Consider using "mini" experiments to test alternative service delivery procedures and provide more rigorous evidence.

Reporting the performance information

A8: Shortly after each performance report is produced, provide for its summarization and highlighting.

A9: Examine and report the extent of uncertainty in the measured outcome values.

A10: Make sure reports are clear, understandable, and meaningful—internal as well as external reports.

A11: Incorporate relevant findings from completed program evaluations into the performance measurement reports.

RECOMMENDATIONS ON USE OF PERFORMANCE INFORMATION

Basic uses for performance information

U1: Hold regular data-driven performance reviews with staff, using data from the performance measurement system as a starting point for such meetings.

U2: Encourage the use of regular data-driven performance reviews at lower levels of management, not only at the top levels of an organization.

U3: Use the performance measurement system outcome data as a major basis for developing and justifying policy choices, including budgets and strategic plans.

U4: Use performance information to motivate employees and contractors—but with caution.

U5: Use performance information as the basis of reports to persons outside the organizations, such as legislative bodies, the media, and citizens.

U6: Link up with other agencies to tackle complex issues involving multiple agencies and programs. Enable performance data sharing across agencies while protecting confidentiality.

Making these uses more effective

U7: Enable managers and staff to access timely, up-to-date, performance data at any time during the year.

U8: Provide relevant performance information regularly to first-line staff, not just supervisory personnel.

U9: When making decisions, consider not only the aggregated data but also disaggregated outcome data, categorized by key customer and service characteristics.

U10: Before making decisions, such as on data indicating unexpectedly poor performance, consider not only the performance data but also explanatory information.

U11: Make use of the information from the performance measurement system in in-depth, ad hoc studies, such as program evaluations.

U12: Provide training, technical assistance, and/or mentoring to managers and their staff in accessing, interpreting, and using of performance information.

U13: Track improvements that have resulted, at least in part, from use of the performance measurement system. Regularly publicize them