Robert D. Behn
Harvard University

Why Measure Performance? Different Purposes Require Different Measures

Performance measurement is not an end in itself. So why should public managers measure performance? Because they may find such measures helpful in achieving eight specific managerial purposes. As part of their overall management strategy, public managers can use performance measures to evaluate, control, budget, motivate, promote, celebrate, learn, and improve. Unfortunately, no single performance measure is appropriate for all eight purposes. Consequently, public managers should not seek the one magic performance measure. Instead, they need to think seriously about the managerial purposes to which performance measurement might contribute and how they might deploy these measures. Only then can they select measures with the characteristics necessary to help achieve each purpose. Without at least a tentative theory about how performance measures can be employed to foster improvement (which is the core purpose behind the other seven), public managers will be unable to decide what should be measured.

Everyone is measuring performance.[1] Public managers are measuring the performance of their organizations, their contractors, and the collaboratives in which they participate. Congress, state legislatures, and city councils are insisting that executive-branch agencies periodically report measures of performance. Stakeholder organizations want performance measures so they can hold government accountable. Journalists like nothing better than a front-page bar chart that compares performance measures for various jurisdictions—whether they are average test scores for the city's schools or FBI uniform crime statistics for the state's cities. Moreover, public agencies are taking the initiative to publish compilations of their own performance measurements (Murphey 1999). A major trend among the nations that comprise the Organisation for Economic Cooperation and Development, concludes Alexander Kouzmin (1999) of the University of Western Sydney and his colleagues, is "the development of measurement systems which enable comparison of similar activities across a number of areas," (122) and which "help to establish a performance-based culture in the public sector" (123). "Performance measurement," writes Terrell Blodgett of the University of Texas and Gerald Newfarmer of Management Partners, Inc., is "(arguably) the hottest topic in government today" (1996, 6).

Why Measure Performance?

What is behind all of this measuring of performance? What do people expect to do with the measures—other than use them to beat up on some underperforming agency, bureaucrat, or contractor? How are people actually using these performance measures? What is the rationale that connects the measurement of government's performance to some higher purpose? After all, neither the act of measuring performance nor the resulting data accomplishes anything itself; only when someone uses these measures in some way do they accomplish something. For what purposes do—or might—people measure the performance of public agencies, public programs, nonprofit and for-profit contractors, or the collaboratives of public, nonprofit, and for-profit organizations that deliver public services?[2]

Why measure performance? Because measuring performance is good. But how do we *know* it is good? Because business firms all measure their performance, and everyone knows that the private sector is managed better than

Robert D. Behn *is a lecturer at Harvard University's John F. Kennedy School of Government and the faculty chair of its executive program Driving Government Performance. His research focuses on governance, leadership, and performance management. His latest book is* Rethinking Democratic Accountability *(Brookings Institution, 2001). He believes the most important performance measure is 1918: the last year the Boston Red Sox won the World Series.* ***Email: redsox@ksg.harvard.edu.***

the public sector. Unfortunately, the kinds of financial ratios the business world uses to measure a firm's performance are not appropriate for the public sector. So what should public agencies measure? Performance, of course. But what kind of performance should they measure, how should they measure it, and what should they do with these measurements? A variety of commentators offer a variety of purposes:

- Joseph Wholey of the University of Southern California and Kathryn Newcomer of George Washington University observe that "the current focus on performance measurement at all levels of government and in nonprofit organizations reflects citizen demands for evidence of program effectiveness that have been made around the world" (1997, 92).
- In their case for performance monitoring, Wholey and the Urban Institute's Harry Hatry note that "performance monitoring systems are beginning to be used in budget formulation and resource allocation, employee motivation, performance contracting, improving government services and improving communications between citizens and government" (1992, 604), as well as for "external accountability purposes" (609).
- "Performance measurement may be done annually to improve public accountability and policy decision making," write Wholey and Newcomer, "or done more frequently to improve management and program effectiveness" (1997, 98).
- The Governmental Accounting and Standards Board suggests that performance measures are "needed for setting goals and objectives, planning program activities to accomplish these goals, allocating resources to these programs, monitoring and evaluating the results to determine if they are making progress in achieving the established goals and objectives, and modifying program plans to enhance performance" (Hatry et al. 1990, v).
- Municipalities, notes Mary Kopczynski of the Urban Institute and Michael Lombardo of the International City/County Management Association, can use comparative performance data in five ways: "(1) to recognize good performance and to identify areas for improvement; (2) to use indicator values for higher-performing jurisdictions as improvement targets by jurisdictions that fall short of the top marks; (3) to compare performance among a subset of jurisdictions believed to be similar in some way (for example, in size, service delivery practice, geography, etc); (4) to inform stakeholders outside of the local government sector (such as citizens or business groups); and (5) to solicit joint cooperation in improving future outcomes in respective communities" (1999, 133).
- Advocates of performance measurement in local government, observes David Ammons of the University of North Carolina, "have promised that more sophisticated measurement systems will undergird management processes, better inform resource allocation decisions, enhance legislative oversight, and increase accountability" (1995, 37).
- Performance measurement, write David Osborne and Peter Plastrik in *The Reinventor's Fieldbook,* "enables officials to hold organizations accountable and to introduce consequences for performance. It helps citizens and customers judge the value that government creates for them. And it provides managers with the data they need to improve performance" (2000, 247).
- Robert Kravchuk of Indiana University and Ronald Schack of the Connecticut Department of Labor do not offer a specific list of purposes for measuring performance. Nevertheless, imbedded in their proposals for designing effective performance measures, they suggest a number of different purposes: planning, evaluation, organizational learning, driving improvement efforts, decision making, resource allocation, control, facilitating the devolution of authority to lower levels of the hierarchy, and helping to promote accountability (Kravchuck and Schack 1996, 348, 349, 350, 351).

Performance measures can be used for multiple purposes. Moreover, different people have different purposes. Legislators have different purposes than journalists. Stakeholders have different purposes than public managers. Consequently, I will focus on just those people who manage public agencies.

Eight Managerial Purposes for Measuring Performance

What purpose—exactly—is a public manager attempting to achieve by measuring performance? Even for this narrower question, the answer isn't obvious. One analyst admonishes public managers: "Always remember that the intent of performance measures is to provide reliable and valid information on performance" (Theurer 1998, 24). But that hardly answers the question. What will public managers do with all of this reliable and valid information? Producing reliable and valid reports of government performance is no end in itself. All of the reliable and valid data about performance is of little use to public managers if they lack a clear idea about how to use them or if the data are not appropriate for this particular use. So what, exactly, will performance measurement do, and what kinds of measures do public managers need to do this? Indeed, what is the logic behind all of this performance measurement—the causal link between the measures and the public manager's effort to achieve specific policy purposes?

Hatry offers one of the few enumerated lists of the uses of performance information. He suggests that public managers can use such information to perform ten different

tasks: to (1) respond to elected officials' and the public's demands for accountability; (2) make budget requests; (3) do internal budgeting; (4) trigger in-depth examinations of performance problems and possible corrections; (5) motivate; (6) contract; (7) evaluate; (8) support strategic planning; (9) communicate better with the public to build public trust; and (10) improve.[3] Hatry notes that improving programs is the fundamental purpose of performance measurement, and all but two of these ten uses—improving accountability and increasing communications with the public—"are intended to make program improvements that lead to improved outcomes" (1999b, 158, 157).

My list is slightly different. From the diversity of reasons for measuring performance, I think public managers have eight primary purposes that are specific and distinct (or only marginally overlapping[4]). As part of their overall management strategy, the leaders of public agencies can use performance measurement to (1) evaluate; (2) control; (3) budget; (4) motivate; (5) promote; (6) celebrate; (7) learn; and (8) improve.[5]

This list could be longer or shorter. For the measurement of performance, the public manager's real purpose—indeed, the only real purpose—is to improve performance. The other seven purposes are simply means for achieving this ultimate purpose. Consequently, the choice of how many subpurposes—how many distinct means—to include is somewhat arbitrary. But my major point is not. Instead, let me emphasize: The leaders of public agencies can use performance measures to achieve a number of very different purposes, and they need to carefully and explicitly choose their purposes. Only then can they identify or create specific measures that are appropriate for each individual purpose.[6]

Of the various purposes that others have proposed for measuring performance, I have not included on my list: planning, decision making, modifying programs, setting performance targets, recognizing good performance, comparing performance, informing stakeholders, performance contracting, and promoting accountability. Why not? Because these are really subpurposes of one (or more) of the eight basic purposes. For example, planning, decision making, and modifying are implicit in two of my eight, more basic, purposes: budgeting and improving. The real reason that managers plan, or make decisions, or modify programs is to either reallocate resources or to improve future performance. Similarly, the reason that managers set performance targets is to motivate, and thus to improve. To compare performance among jurisdictions is—implicitly but undeniably—to evaluate them. Recognizing good performance is designed to motivate improvements. Informing stakeholders both promotes and gives them the opportunity to evaluate and learn. Performance contracting involves all of the eight purposes from evaluating to improving. And, depending upon what people mean by accountability, they may promote it by evaluating public agencies, by controlling them, or by motivating them to improve[7] (table 1).

Table 1 Eight Purposes that Public Managers Have for Measuring Performance

The purpose	The public manager's question that the performance measure can help answer
Evaluate	How well is my public agency performing?
Control	How can I ensure that my subordinates are doing the right thing?
Budget	On what programs, people, or projects should my agency spend the public's money?
Motivate	How can I motivate line staff, middle managers, nonprofit and for-profit collaborators, stakeholders, and citizens to do the things necessary to improve performance?
Promote	How can I convince political superiors, legislators, stakeholders, journalists, and citizens that my agency is doing a good job?
Celebrate	What accomplishments are worthy of the important organizational ritual of celebrating success?
Learn	Why is what working or not working?
Improve	What exactly should who do differently to improve performance?

Purpose 1. To Evaluate: How Well Is This Government Agency Performing?

Evaluation is the usual reason for measuring performance. Indeed, many of the scholars and practitioners who are attempting to develop systems of performance measurement have come from the field of program evaluation. Often (despite the many different reasons cited earlier), no reason is given for measuring performance; instead, the evaluation purpose is simply assumed. People rarely state that their only (or dominant) rationale for measuring performance is to evaluate performance, let alone acknowledge there may be other purposes. It is simply there between the lines of many performance audits, budget documents, articles, speeches, and books: People are measuring the performance of this organization or that program so they (or others) can evaluate it.

In a report on early performance-measurement efforts under the Government Performance and Results Act of 1993, an advisory panel of the National Academy of Public Administration (NAPA) observed, "Performance measurement of program outputs and outcomes provides important, if not vital, information on current program status and how much progress is being made toward important program goals. It provides needed information as to whether problems are worsening or improving, even if it cannot tell us why or how the problem improvement (or worsening) came about" (NAPA 1994, 2). These sentences do not contain the words "evaluation" or "evaluate," yet they clearly imply the performance measurements will furnish some kind of assessment of program performance.

Of course, to evaluate the performance of a public agency, the public manager needs to know what that agency

is supposed to accomplish. For this reason, two of the ten performance-measurement design principles developed by Kravchuk and Schack are to "formulate a clear, coherent mission, strategy, and objectives," and to "rationalize the programmatic structure as a prelude to measurement." Do this first, they argue, because "performance measurement must begin with a clear understanding of the policy objectives of a program, or multiprogram system," and because "meaningful measurement requires a rational program structure" (1996, 350). Oops. If public managers have to wait for the U.S. Congress or the local city council to formulate (for just one governmental undertaking) a clear, coherent mission, strategy, and objectives combined with a rationalized program structure, they will never get to the next step of measuring anything.[8]

No wonder many public managers are alarmed by the evaluative nature of performance measurement. If there existed a clear, universal understanding of their policy objectives, and if they could manage within a rational program structure, they might find performance measurement less scary. But without an agreement on policy objectives, public managers know that others can use performance data to criticize them (and their agency) for failing to achieve objectives that they were not pursuing. And if given responsibility for achieving widely accepted policy objectives with an insane program structure (multiple constraints, inadequate resources, and unreasonable timetables), even the most talented managers may fall short of the agreed-upon performance targets.

Moreover, even if the performance measures are not collected for the *explicit* purpose of evaluation, this possibility is always *implicit*. And using performance data to evaluate a public agency is a tricky and sophisticated undertaking. Yet, a simple comparison of readily available data about similar (though rarely identical) agencies is the most common evaluative technique. Hatry (1999a) notes that intergovernmental comparisons of performance "focus primarily on indicators that can be obtained from traditional and readily available data sources." This is the common practice, he continues, because "the best outcome data cannot be obtained without new, or at least, substantially revised procedures" (104).

Often, however, existing or easily attainable data create an opportunity for simplistic, evaluative comparisons. Hatry writes that those who collect comparative performance data, as well as "the public, and the media must recognize that the data in comparative performance measurement efforts will only be roughly comparable" (1999a, 104). But will journalists, who must produce this evening's news or tomorrow's newspaper under very tight deadlines, recognize this, let alone explain it? And will the public, in their quick glance at an attractive bar chart, get this message? Hatry, himself, is not completely sanguine:

> The ultimate question of comparative data is whether publication does more harm than good. More harm can occur if many of the measurements contain errors or are otherwise unfair, so that low performers are unfairly beaten up by the media and have to spend excessive amounts of time and effort attempting to explain and defend themselves.… On the other hand, if the data seem on the whole to encourage jurisdictions to explore why low performance has occurred and how they might better themselves, then such efforts will be worthwhile, even if a few agencies are unfairly treated." (Hatry 1999a, 104).

Whether the scholars, analysts, or managers like it, almost any performance measure can and will be used to evaluate a public agency's performance.

Purpose 2. To Control: How Can Public Managers Ensure Their Subordinates Are Doing the Right Thing?

Yes. Frederick Winslow Taylor *is* dead. Today, no manager believes the best way to influence the behavior of subordinates is to establish the one best way for them to do their prescribed tasks and then measure their compliance with this particular way. In the twenty-first century, all managers are into empowerment.

Nevertheless, it is disingenuous to assert (or believe) that people no longer seek to control the behavior of public agencies and public employees, let alone seek to use performance measurement to help them do so.[9] Why do governments have line-item budgets? Today, no one employs the measurements of time-and-motion studies for control. Yet, legislatures and executive-branch superiors do establish performance standards—whether they are specific curriculum standards for teachers or sentencing standards for judges—and then measure performance to see whether individuals have complied with these mandates.[10] After all, the central concern of the principle–agent theory is how principles can control the behavior of their agents (Ingraham and Kneedler 2000, 238–39).

Indeed, the controlling style of management has a long and distinguished history. It has cleverly encoded itself into one of the rarely stated but very real purposes behind performance measurement. "Management control depends on measurement," writes William Bruns in a Harvard Business School note on "Responsibility Centers and Performance Measurement" (1993, 1). In business schools, accounting courses and accounting texts often explicitly use the word "control."[11]

In their original explanation of the balanced scorecard, Robert Kaplan and David Norton note that business has a control bias: "Probably because traditional measurement systems have sprung from the finance function, the systems have a control bias. That is, traditional performance measurement systems specify the particular actions they

want employees to take and then measure to see whether the employees have in fact taken those actions. In that way, the systems try to control behavior. Such measurement systems fit with the engineering mentality of the Industrial Age" (1992, 79). The same is true in the public sector. Legislatures create measurement systems that specify particular actions they want executive-branch employees to take and particular ways they want executive-branch agencies to spend money. Executive-branch superiors, regulatory units, and overhead agencies do the same. Then, they measure to see whether the agency employees have taken the specified actions and spent the money in the specified ways.[12] Can't you just see Fred Taylor smiling?

Purpose 3. To Budget: On What Programs, People, or Projects Should Government Spend the Public's Money?

Performance measurement can help public officials to make budget allocations. At the macro level, however, the apportionment of tax monies is a political decision made by political officials. Citizens delegate to elected officials and their immediate subordinates the responsibility for deciding which purposes of government action are primary and which ones are secondary or tertiary. Thus, political priorities—not agency performance—drive macro budgetary choices.

Performance budgeting, performance-based budgeting, and results-oriented budgeting are some of the names commonly given to the use of performance measures in the budgetary process (Holt 1995–96; Jordon and Hackbart 1999; Joyce 1996, 1997; Lehan 1996; Melkers and Willoughby 1998, 2001; Thompson 1994; Thompson and Johansen 1999). But like so many other phrases in the performance-measurement business, they can mean different things to different people in different contexts.[13] For example, performance budgeting may simply mean including historical data on performance in the annual budget request. Or it may mean that budgets are structured not around line-item expenditures (with performance purposes or targets left either secondary or implicit), but around general performance purposes or specific performance targets (with line-item allocations left to the managers of the units charged with achieving these purposes or targets). Or it may mean rewarding units that do well compared to some performance targets with extra funds and punishing units that fail to achieve their targets with budget cuts.

For improving performance, however, budgets are crude tools. What should a city do if its fire department fails to achieve its performance targets? Cut the department's budget? Or increase its budget? Or should the city manager fire the fire chief and recruit a public manager with a track record of fixing broken agencies? The answer depends on the specific circumstances that are not captured by the formal performance data. Certainly, cutting the fire department's budget seems like a counterproductive way to improve performance (though cutting the fire department's budget may be perfectly logical if the city council decides that fire safety is less of a political priority than educating children, fixing the sewers, or reducing crime). If analysis reveals the fire department is underperforming because it is underfunded—because, for example, its capital budget lacks the funds for cost-effective technology—then increasing the department's budget is a sensible response. But poor performance may be the result of factors that more (or less) money won't fix: poor leadership, the lack of a fire-prevention strategy to complement the department's fire-fighting strategy, or the failure to adopt industry training standards. Using budgetary increments to reward well-performing agencies and budgetary decrements to punish underperforming ones is not a strategy that will automatically fix (or even motivate) poor performers.

Nevertheless, line managers can use performance data to inform their resource-allocation decisions. Once elected officials have established macro political priorities, those responsible for more micro decisions may seek to invest their limited allocation of resources in the most cost-effective units and activities. And when making such micro budgetary choices, public managers may find performance measures helpful.

Purpose 4. To Motivate: How Can Public Managers Motivate Line Staff, Middle Managers, Nonprofit and For-Profit Collaborators, Stakeholders, and Citizens to Do the Things Necessary to Improve Performance?

Public managers may use performance measures to learn how to perform better. Or, if they already understand what it takes to improve performance, they may use the measures to motivate such behavior. And for this motivational purpose, performance measures have proven to be very useful.

The basic concept is that establishing performance goals—particularly stretch goals—grabs people's attention. Then the measurement of progress toward the goals provides useful feedback, concentrating their efforts on reaching these targets. In his book *The Great Ideas of Management,* Jack Duncan of the University of Alabama reports on the startling conclusion of research into the impact of goal setting on performance: "No other motivational technique known to date can come close to duplicating that record" (1989, 127).

To implement this motivational strategy, an agency's leadership needs to give its people a significant goal to achieve and then use performance measures—including interim targets—to focus people's thinking and work and to provide a periodic sense of accomplishment. Moreover,

performance targets may also encourage creativity in evolving better ways to achieve the goal (Behn 1999); thus, measures that motivate improved performance may also motivate learning.[14]

In New York City in the 1970s, Gordon Chase used performance targets to motivate the employees of the Health Services Administration (Rosenthal 1975; Levin and Sanger 1994). In Massachusetts in the 1980s, the leadership of the Department of Public Welfare used the same strategy (Behn 1991). And in the 1990s in Pennsylvania, the same basic approach worked in the Department of Environmental Protection (Behn 1997a). But perhaps the most famous application of performance targets to motivate public employees is Compstat, the system created by William Bratton, then commissioner of the New York Police Department, to focus attention of precinct commanders on reducing crime (Silverman 2001, 88–89, 101).

Purpose 5. To Promote: How Can Public Managers Convince Political Superiors, Legislators, Stakeholders, Journalists, and Citizens that Their Agency Is Doing a Good Job?

Americans suspect their government is both ineffective and inefficient. Yet, if public agencies are to accomplish public purposes, they need the public's support. Performance measures can contribute to such support by revealing not only when government institutions are failing, but also when they are doing a good or excellent job. For example, the National Academy of Public Administration's Center for Improving Government Performance reports that performance measures can be used to "validate success; justify additional resources (when appropriate); earn customer, stakeholder, and staff loyalty by showing results; and win recognition inside and outside the organization" (NAPA 1999, 7).

Still, too many public managers fail to use performance measures to promote the value and contribution of their agency. "Performance-based measures," writes Harry Boone of the Council of State Governments, "provide a justification for the agency's existence," yet "many agencies cannot defend their effectiveness in performance-based terms" (1996, 10).

In a study, "Toward Useful Performance Measures," a National Academy of Public Administration advisory panel (1994) asserts that "performance indicators can be a powerful tool in communicating program value and accomplishments to a variety of constituencies" (23). In addition to "the use of performance measurement to communicate program success and worth" (9), the panel noted, the "major values of a performance measurement system" include its potential "to enhance public trust" (9). That is, the panel argues, performance measurement can not only directly establish—and thus promote—the competence of specific agencies and the value of particular programs; it also can indirectly establish, and thus promote, the competence and value of government in general.

Purpose 6. To Celebrate: What Accomplishments Are Worthy of the Important Organizational Ritual of Celebrating Success?

All organizations need to commemorate their accomplishments. Such rituals tie people together, give them a sense of their individual and collective relevance, and motivate future efforts. Moreover, by achieving specific goals, people gain a sense of personal accomplishment and self-worth (Locke and Latham 1984, 1990). Such celebrations need not be limited to one big party to mark the end of the fiscal year or the completion of a significant project. Small milestones along the way—as well as unusual achievements and unanticipated victories—provide an opportunity for impromptu celebrations that call attention to these accomplishments and to the people who made them happen. And such celebrations can help to focus attention on the next challenge.

Like all of the other purposes for measuring performance—with the sole and important exception of improvement—celebration is not an end in itself. Rather, celebration is important because it motivates, promotes, and recruits. Celebration helps to improve performance because it motivates people to improve further in the next year, quarter, or month. Celebration helps to improve performance because it brings attention to the agency, and thus promotes its competence. And this promotion—this attention—may even generate increased flexibility (from overhead agencies) and resources (from the guardians of the budget). Moreover, this promotion and attention attract another resource: dedicated people who want to work for a successful agency that is achieving important public purposes. Celebration may even attract potential collaborators from other organizations that have not received as much attention, and thus seek to enhance their own sense of accomplishment by shifting some of their energies to the high-performing collaborative (Behn 1991, 92–93).

Celebration also may be combined with learning. Rather than hold a party to acknowledge success and recognize its contributors, an informal seminar or formal presentation can realize the same purposes. Asking those who produced the unanticipated achievement or unusual victory to explain how they pulled it off celebrates their triumph; but it also provides others with an opportunity to learn how they might achieve a similar success (Behn 1991, 106–7).

Still, the links from measurement to celebration to improvement is the most indirect because it has to work through one of the other links—either motivation, budgeting, learning, or promotion. In the end, any reason for measuring performance is valid only to the extent that it helps to achieve the most basic purpose: to *improve* performance.

Purpose 7. To Learn: Why Is What Working or Not Working?

Performance measures contain information that can be used not only to evaluate, but also to learn. Indeed, learning is more than evaluation. The objective of evaluation is to determine what is working and what isn't. The objective of learning is to determine *why.*

To learn from performance measures, however, managers need some mechanism to extract information from the data. We may all believe that the data speak for themselves. This, however, is only because we each have buried in our brain some unconscious mechanism that has already made an implicit conversion of the abstract data into meaningful information. The data speak only through an interpreter that converts the collection of digits into analog lessons—that decodes the otherwise inscrutable numbers and provides a persuasive explanation. And often, different people use different interpreters, which explains how they can draw very different lessons from the same data.[15]

Moreover, if managers have too many performance measures, they may be unable to learn anything. Carole Neves of the National Academy of Public Administration, James Wolf of Virginia Tech, and Bill Benton of Benton and Associates (1986) write that "in many agencies," because of the proliferation of performance measures, "there is more confusion or 'noise' than useful data." Theodore Poister and Gregory Streib of Georgia State University call this the "'DRIP' syndrome—Data Rich but Information Poor" (1999, 326). Thus, Neves and her colleagues conclude, "managers lack time or simply find it too difficult to try to identify good signals from the mass of numbers" (1986, 141).

From performance measures, public managers may learn what *is not* working. If so, they can stop doing it and reallocate money and people from this nonperforming activity to more effective undertakings (designed to achieve the identical or quite different purposes). Or they may learn what *is* working. If so, they can shift existing resources (or new resources that become available) to this proven activity. Learning can help with the budgeting of both money and people.

Furthermore, learning can help more directly with the improving. The performance measures can reveal not only whether an agency is performing well or poorly, but also *why:* What is contributing to the agency's excellent, fair, or poor performance—and what might be done to improve the components that are performing fairly or poorly?

In seeking to learn from performance measures, public managers frequently confront the black box enigma of social science research.[16] The data—the performance measures—can reveal that an organization is performing well or poorly, but they don't necessarily reveal *why.* The performance measures can describe what is coming out of the black box of a public agency, as well as what is going in, but they don't necessarily reveal what is happening inside. How are the various inputs interacting to produce the outputs? What is the organizational black box actually doing to the inputs to convert them into the outputs? What is the societal black box actually doing to the outputs to convert them into the outcomes?[17]

Public managers can, of course, create some measures of the processes going on inside the black box. But they cannot guarantee that the internal characteristics and processes of the black box they have chosen to measure are actually the ones that determine whether the inputs are converted into high-quality or low-quality outputs. Yet, the more internal processes that public managers choose to measure, the more likely they are to discover a few that correlate well with the outputs. Such correlations could, however, be purely random,[18] or the factors that are identified by the correlations as significant contributors could merely be correlated with other factors that are the real causes. Converting performance data into an understanding of what is happening inside the black box is neither easy nor obvious.

Purpose 8. To Improve: What Exactly Should Who Do Differently to Improve Performance?

Performance " 'measurement' is not an end in itself but must be used by managers to make improvements" (NAPA 1994, 22), emphasizes an advisory panel of the National Academy of Public Administration. In fact, the word "improve" (or "improving" or "improvement") appears more than a dozen times in this NAPA report. "Ideally," the panel concludes, "performance data should be part of a continuous feedback loop that is used to report on program value and accomplishment and identify areas where performance is weak so that steps can be taken to promote improvements" (22). Yet, the panel also found "little evidence in most [GRPA pilot performance] plans that the performance information would be used to improve program performance" (8).

Similarly, Hatry argues the "fundamental purpose of performance information" is "to make program improvements" (1999b, 158). But *how?* What exactly is the connection between the measurement and the improvement? *Who* has to do *what* to convert the measurement into an improvement? Or does this just happen automatically? No, responds the NAPA panel: "measurement alone does not bring about performance improvement" (1994, 15).

For example, if the measurement produces some learning, someone then must convert that learning into an improvement. Someone has to intervene consciously and actively. But can any slightly competent individual pull this off? Or does it require a sophisticated appreciation of the

strategies and pitfalls of converting measurement into improvement? To improve, an organization needs the capacity to adopt—and adapt—the lessons from its learning.

Learning from performance measures, however, is tricky. It isn't obvious what lessons public managers should draw about which factors are contributing to the good or poor performance, let alone *how* they might modify such factors to foster improvements. Improvement requires attention to the feedback—the ability to check whether the lessons postulated from the learning have been implemented in a way that actually changes organizational behavior so that it results in the better outputs and outcomes that the learning promised. Improvement is active, operational learning.

The challenge of learning from the performance measures is both intellectual and operational. Public managers who wish to use measurement to improve the performance of their agencies face two challenges: First, they have the intellectual challenge of figuring out how to learn which changes in plans, or procedures, or personnel might produce improvements. Then, they confront the operational challenge of figuring out how to implement the indicated changes.

There are a variety of standard mechanisms for using performance measures to evaluate. There exist some such mechanisms to control and budget. For the purposes of learning and improving, however, each new combination of policy objectives, political environment, budgetary resources, programmatic structure, operational capacity, regulatory constraints, and performance measures demands a more open-ended, qualitative analysis. For performance learning and performance improvement, there is no cookbook.[19]

How does the measurement of performance beget improvement? Measurement can influence performance in a variety of ways, most of which are hardly direct or apparent. There exist a variety of feedback loops, though not all of them may be obvious, and the obvious ones may not function as expected or desired. Consequently, to measure an agency's performance in a way that can actually help improve its performance, the agency's leadership needs to think seriously not only about *what* it should measure, but also about *how* it might deploy any such measurements. Indeed, without at least some tentative theory about how the measurements can be employed to foster improvements, it is difficult to think about what should be measured.

Selection Criteria for Each Measurement Purpose

What kinds of performance measures are most appropriate for which purposes? It isn't obvious. Moreover, a measure that is particularly appropriate for one purpose may be completely useless for another. For example, "in many cases," Newcomer notes, "the sorts of measures that might effectively inform program improvement decisions may provide data that managers would not find helpful for resource allocation purposes" (1997, 8). Before choosing a performance measure, public managers must first choose their purpose.

Kravchuk and Schack note that no one measure or even one collection of measures is appropriate for all circumstances: "The search for a single array of measures for all needs should be abandoned, especially where there are divergent needs and interests among key users of performance information." Thus, they advocate "an explicit measurement strategy" that will "provide for the needs of all important users of performance information" (Kravchuk and Schack 1996, 350).

I take a similar approach. But, rather than worry about the needs of different kinds of users, I focus on the different purposes for which the users—specifically, public managers—can employ the performance measures. After all, different users want different measures because they have different purposes. But it is the nature of the purpose—not the nature of the user—that determines which characteristics of those measures will be most helpful. The usual admonition of performance measurement is, "Don't measure inputs. Don't measure processes. Don't measure outputs. Measure outcomes." But outcomes are not necessarily the best measure for all purposes.

Will a particular public manager find a certain performance measure helpful for a specific purpose? The answer depends not on the organizational position of that manager, but on whether this measure possesses the characteristics required by the manager's purpose (table 2).

Purpose 1: To Evaluate

Evaluation requires a comparison. To evaluate the performance of an agency, its managers have to compare that

Table 2 Characteristics of Performance Measures for Different Purposes

The purpose	To help achieve this purpose, public managers need
Evaluate	Outcomes, combined with inputs and with the effects of exogenous factors
Control	Inputs that can be regulated
Budget	Efficiency measures (specifically outcomes or outputs divided by inputs)
Motivate	Almost-real-time outputs compared with production targets
Promote	Easily understood aspects of performance about which citizens really care
Celebrate	Periodic and significant performance targets that, when achieved, provide people with a real sense of personal and collective accomplishment
Learn	Disaggregated data that can reveal deviancies from the expected
Improve	Inside-the-black-box relationships that connect changes in operations to changes in outputs and outcomes

performance with some standard. Such a standard can come from past performance, from the performance of similar agencies, from a professional or industry standard, or from political expectations. But without such a basis for comparison, it is impossible to determine whether the agency is performing well or poorly.

And to compare actual performance against the performance criterion requires a variety of outcome measures, combined with some input (plus environmental, process, and output) measures. The focus, however, is on the outcomes. To evaluate a public agency—to determine whether it is achieving its public purpose—requires some measure of the outcomes that the agency was designed to affect. Only with outcome measures can public managers answer the *effectiveness question*: Did the agency achieve the results it set out to produce? Then, dividing by some input measures, they can ask the *efficiency question*: Did this agency produce these results in a cost-effective way? To answer either of these evaluative questions, a manager needs to measure outcomes.[20]

Of course, the agency did not produce all of the outcomes alone. Other factors, such as economic conditions, affected them. Consequently, public managers also need to ask the *impact question*: What did the agency itself accomplish? What is the difference between the actual outcomes and the outcomes that would have occurred if the agency had not acted?

Another way of assessing an organization or program is to evaluate its internal operations. This is the *best-practice question*: How do the operations and practices of this organization or program compare with the ones that are known to be most effective and efficient? To conduct such a best-practice evaluation requires some process measures—appropriate descriptions of the organization's key internal operations that can be compared with some operational standards.

No one comparison of a single outcome measure with a single performance standard will provide a definitive evaluation. Rather, to provide a conscientious and credible picture of the agency's performance, an evaluation requires multiple measures compared with multiple standards.

Purpose 2: To Control

To control the behavior of agencies and employees, public officials need input requirements. Indeed, whenever you discover officials who are using input measures, you can be sure they are using them to control. To do this, officials need to measure the corresponding behavior of individuals and organizations and then compare this performance with the requirements to check who has and has not complied: Did the teachers follow the curricular requirements for the children in their classrooms? Did the judges follow the sentencing requirements for those found guilty in their courts? Often, such requirements are described only as guidelines: curriculum guidelines, sentencing guidelines. Do not be fooled. These guidelines are really requirements, and these requirements are designed to control. The measurement of compliance with these requirements is the mechanism of control.

Purpose 3: To Budget

To use performance measures for budgeting purposes, public managers need measures that describe the efficiency of various activities. Then, once political leaders have set macro budgetary priorities, agency managers can use efficiency measures to suggest the activities in which they should invest the appropriated funds. Why spend limited funds on some programs or organizations when the performance measures reveal that other programs or organizations are more efficient at achieving the political objectives behind the budget's macro allocations?

To use performance measures to budget, however, managers need not only data on outcomes (or outputs) for the numerator in the efficiency equation; they also need reliable cost data for the denominator. And these cost measures have to capture not only the obvious, direct costs of the agency or program, but also the hidden, indirect costs. Few governments, however, have created cost-accounting or activity-based-accounting systems that assign to each government function the full and accurate costs (Coe 1999, 112; Joyce 1997, 53, 56; Thompson 1994).

Budgeting usually concerns the allocation of dollars. But most public managers are constrained by a system of double budgeting. They must manage a fixed number of dollars and a fixed number of personnel slots. Thus, in attempting to maximize the productivity of these two constrained resources, they also need to budget their people. And to use performance measurement for this budgetary purpose, they need not only outcome (or output) measures for the numerator of their efficiency equation, but also input data in terms of people for the denominator. Public managers need to allocate their people to the activities with the highest productivity per person.

Purpose 4: To Motivate

To motivate people to work harder or smarter, public managers need almost-real-time measures of outputs to compare with production targets. Organizations don't produce outcomes; organizations produce outputs. And to motivate an organization to improve its performance, managers have to motivate it to improve what it actually does. Consequently, although public managers want to use outcome data to evaluate their agency's performance, they need output data to motivate better performance.[21] Managers can't motivate people to do something they can't do; managers can't motivate people to affect some-

thing over which they have little or no influence; managers can't motivate people to produce an outcome they do not themselves produce.

Moreover, to motivate, managers have to collect and distribute the output data quickly enough to provide useable feedback. Those who produce the agency's outputs cannot adjust their production processes to respond to inadequacies or deficiencies unless they know how well they are doing against their current performance target. Eli Silverman of the John Jay College of Criminal Justice describes Compstat as "intelligence-led policing" (2001, 182). The New York Police Department collects, analyzes, and quickly distributes to managers at all levels—from commissioner to patrol sergeants—the data about the current patterns and concentrations of crime that are necessary to develop strategic responses.

This helps to explain why society attempts to motivate schools and teachers with test scores. The real, ultimate outcome that citizens seek from our public schools is children who grow up to become productive employees and responsible citizens. But using a measure of employee productivity and citizen responsibility to motivate performance creates a number of problems. First, it is very difficult to develop a widely acceptable measure of employee productivity (do we simply use wage levels?), let alone citizen responsibility (do we using voting participation?). Second, schools and teachers are not the only contributors to a future adult's productivity and responsibility. And third, the lag between when the schools and teachers do their work and when these outcomes can be measured is not just months or years, but decades. Thus, we never could feed these outcome measures back to the schools and teachers in time for them to make any adjustments. Consequently, as a society we must resort to motivating schools and teachers with outputs—with test scores that (presumably) measure how much a child has learned. And although we cannot determine whether schools and teachers are producing productivity or responsibility in future adults, citizens expect they will convey some very testable knowledge and skills.[22]

Once an agency's leaders have motivated significant improvements using output targets, they can create some outcome targets. Output targets can motivate people to focus on improving their agency's internal processes, which produce the outputs. Outcome targets, in contrast, can motivate people to look outside their agency—to seek ways to collaborate with other individuals and organizations whose activities may affect (perhaps more directly) the outcomes and values the agency is really charged with producing (Bardach 1998; Sparrow 2000).

Purpose 5: To Promote

To convince citizens their agency is effective and efficient, public managers need easily understood measures of those aspects of performance about which many citizens personally care. And such performance may be only tangentially related to the agency's public purpose.

The National Academy of Public Administration, in its study of early performance-measurement plans under the Government Performance and Results Act, noted that "most plans recognized the need to communicate performance evaluation results to higher level officials, but did not show clear recognition that the form and level of data for these needs would be different than that for operating managers." NAPA emphasized that the needs of "department heads, the Executive Office of the President, and Congress" are "different and the special needs of each should be more explicitly defined" (1994, 23). Similarly, Kaplan and Norton stress that different customers have different concerns (1992, 73–74).

Consider a state division of motor vehicles. Its mission is safety—the safety of vehicle drivers, vehicle passengers, bicycle riders, and pedestrians. In pursuit of this mission, this agency inspects vehicles to ensure their safety equipment is working, and it inspects people to ensure they are safe drivers. When most citizens think about their division of motor vehicles, however, what is their greatest care? Answer: How long they will have to stand in line. If a state DMV wants to promote itself to the public, it has to emphasize just one obvious aspect of performance: the time people spend in line. To promote itself to the public, a DMV has to use this performance measure to convince citizens that the time they will spend in line is going down.[23]

Ammons (1995) offers a "revolutionary" approach: "make performance measurement interesting." Municipalities, he argues, ought to adopt measures "that can capture the interest of local media and the public" (43)—particularly measures that "allow meaningful comparisons that to some degree put community pride at stake" (38). Such comparisons could be to a professional or industry standard. After all, as Ammons notes, "comparison with a standard captures attention, where raw information does not" (39).[24] But he is even more attracted to interjurisdictional comparisons. For example, Ammons argues that the public pays attention to various rankings of communities (because, first, journalists pay attention to them). Thus, he wants measures that "are both revealing of operational efficiency and effectiveness and more conducive to cross-jurisdictional comparisons" (38)—measures that provide "opportunities for *interesting* and *meaningful* performance comparisons" (44).

Time spent in line is a measure that is both interesting and meaningful. But what should be the standard for comparison? Is the average time spent in line the most mean-

ingful measure? Or do people care more about the probability they will spend more than some unacceptable time (say one hour) in line?[25] Whatever time-in-line measure it chooses, a DMV may want to compare it with the same measure from neighboring (or similar) states. But will citizens in North Carolina really be impressed that they spend less time in a DMV line than do the citizens of South Carolina or North Dakota? Or will their most meaningful comparison be with the time spent in line at their local supermarket, bank, or fast-food franchise? A DMV manager who wants to promote the agency's competence to the public should compare its time-in-line performance with similar organizational performance that people experience every day. This is an easily understood performance characteristic about which citizens really care.

To do this, however, the agency must not only publish the performance data; it must also make them accessible both physically and psychologically. People must be able to obtain—perhaps not avoid—the measures; they must also find them easy to comprehend.

Purpose 6: To Celebrate

Before an agency can do any celebrating, its managers need to create a performance target that, when achieved, gives its employees and collaborators a real sense of personal and collective accomplishment. This target can be one that has also been used to motivate; it can be an annual target, or one of the monthly or quarterly targets into which an annual target has been divided. Once an agency has produced a tangible and genuine accomplishment that is worth commemorating, its managers need to create a festivity that is proportional to the significance of the achievement.

The verb "to celebrate" suggests a major undertaking—a big, end-of-the-fiscal-year bash, an awards ceremony when a most-wanted criminal is captured, or a victory party when a badly delayed project is completed on deadline. But private-sector managers celebrate lesser accomplishments; they use celebrations of significant successes to convince their employees that their firm is full of winners (Peters and Austin 1985). Public managers have used this strategy, too (Behn 1991, 103–11). But to make the strategy work—to ensure that it creates motivation and thus improvement—the agency's managers have to lead the celebrations.

Purpose 7: To Learn

To learn, public managers need a large number and wide variety of measures—measures that provide detailed, disaggregated information on the various aspects of the various operations of the various components of the agency. When seeking to learn, caution Kravchuk and Schack, public managers need to "avoid excessive aggregation of information" (1996, 357).

Benchmarking is a traditional form of performance measurement that is designed to facilitate learning (Holloway, Francis, and Hinton 1999). It seeks to answer three questions: What is my organization doing well? What is my organization not doing well? What does my organization need to do differently to improve on what it is not doing well? The organization, public or private, identifies a critical internal process, measures it, and compares these data with similar measurements of the identical (or similar) processes of organizations that are recognized as (currently) the best.[26] Any differences suggest not only that the organization needs to improve, but also provide a basis for identifying *how* it could achieve these improvements.

Benchmarking, write Kouzmin et al., is "an instrument for assessing organizational performance and for facilitating management transfer and learning from other benchmarked organizations" (1999, 121). Benchmarking, as they define it, "is a continuous, systematic process of measuring products, services and practices against organizations regarded to be superior with the aim of rectifying any performance 'gaps'" (123). Thus, they conclude, "benchmarking can, on the whole, be seen as a learning strategy" (131). Nevertheless, they caution, for this strategy to work, the organization must become a learning organization. Consequently, they conclude, "the learning effects of benchmarking are, to a very high degree, dependent on adequate organizational conditions and managerial solutions" (132).

Deciding which performance measures best facilitate learning is not easy. If public managers know what they need to do to improve performance, they don't need to learn it. But, if they don't know how they might improve, how do they go about learning it? Kravchuck and Schack note that a "measurement system is a reflection of what decision makers expect to see and how they expect to respond" (1996, 356). That is, when designing a performance-measurement system, when deciding what to measure, managers first will decide what they might see and then create a system to see it.

Real learning, however, is often triggered by the unexpected. As Richard Feynman, the Nobel Prize-winning physicist, explained, when experiments produce unexpected results, scientists start guessing at possible explanations (1965, 157). When Mendel crossed a dwarf plant with a large one, he found that he didn't get a medium-sized plant, but either a small or large one, which led him to discover the laws of heredity (Messadié 1991, 90). When the planet Uranus was discovered to be unexpectedly deviating from its predicted orbit, John Couch Adams and Urbain Le Verrier independently calculated the orbit of an unknown planet that could be causing this unanticipated behavior; then, Johan Gottlieb Galle pointed his telescope in the suggested direction and discovered Neptune (Standage 2000). When Karl Jansky observed that

the static on his radio peaked every 24 hours and that the peak occurred when the Milky Way was centered on his antenna, he discovered radio waves from space (Messadié 1991, 179). Scientific learning often emerges from an effort to explain the unexpected. So does management learning.

Yet how can public officials design a measurement system for the unexpected when they can't figure out what they don't expect? As Kravchuk and Schack write, "unexpected occurrences may not be apprehended by existing measures" (1996, 356). Nevertheless, the more disaggregated the data, the more likely they are to reveal deviancies that may suggest the need to learn. This is the value of management by walking around—or what might be called "data collection by walking around." The stories that people tell managers are the ultimate in disaggregation; one such story can provide a single deviate datum that the summary statistics have completely masked but that, precisely because it was unexpected, prods further investigation that can produce some real learning (and thus, perhaps, some real improvement).

In fact, the learning process may first be triggered by some deviance from the expected that appeared not in the formal performance data, but in the informal monitoring in which all public managers necessarily (if only implicitly) engage. Then, having noticed this deviancy—some aberration that doesn't fit previous patterns, some number so out of line that it jumps off the page, some subtle sign that suggests that something isn't quite right—the manager can create a measuring strategy to learn what caused the deviance and how it can be fixed or exploited.[27]

Failure is a most obvious deviance from the expected and, therefore, provides a significant opportunity to learn.[28] Indeed, a retrospective investigation into the causes of the failure will uncover a variety of measures that deviated from the expected—that is, either from the agency's prescribed behavior or from the predicted consequences of such prescriptions. Thus, failure provides multiple opportunities to learn (Petroski 1985; Sitkin 1992).

Yet, failure (particularly in the public sector) is usually punished—and severely. Thus, when a failure is revealed (or even presumed), people tend to hide the deviate data, for such data can be used to assign blame. Unfortunately, these are the same deviate data that are needed to learn.

As glaring departures from the expected, failures provide managers with obvious opportunities to learn. Most deviances, however, are more subtle. Thus, to learn from such deviances, managers must be intellectually prepared to recognize them and to examine their causes. They have to possess enough knowledge about the operation and behavior of their organization—and about the operation and behavior of their collaborators and society—to distinguish a significant deviance from a random aberration. And when they think they observe an interesting deviance, they need a learning strategy for probing the causes and possible implications.

Thus, Kravchuk and Schack (1996) caution, "organizational learning cannot depend upon measurement alone" (356)—that is, "performance measurement systems cannot replace the efforts of administrators to truly know, understand, and *manage* their programs" (350). Rather, they argue, the measures should indicate when the organization needs to undertake a serious effort at learning based on the "expert knowledge" (357) of its program managers and "other sources of performance information which can supplement the formal measures" (356). Thus, they suggest, "measures should be placed in a management-by-exception frame, where they are regarded as indicators that will serve to signal the need to investigate further" (357). Similarly, Neves, Wolf, and Benton write that "management indicators are intended to be provocative, to suggest to managers a few areas where it may be appropriate to investigate further why a particular indicator shows up the way it does" (1986, 129). The better the manager understands his or her agency and the political, social, and cultural environment in which it works, the better the manager is able to identify—from among the various deviances that are generated by formal and informal performance measures—the ones that are worthy of additional investigation.

Performance measures that diverge from the expected can create an opportunity to learn. But the measures themselves are more likely to suggest topics for investigation than to directly impart key operational lessons.

Purpose 8: To Improve

To ratchet up performance, public managers need to understand how they can influence the behavior of the people inside their agency (and its collaboratives) who produce their outputs and how they can influence the conduct of citizens who convert these outputs into outcomes. They need to know what is going on inside their organization—including the broader organization that consists of everything and everyone whose behavior can affect these outputs and outcomes. They need to know what is going on inside their entire, operational black box. They need inside-the-black-box data that explains how the inputs, environment, and operations they can change (influence or inspire) do (can, or might) cause (create, or contribute to) improvements in the outputs and outcomes. For example, a fire chief needs to understand how the budget input interacts (inside the fire department's black box) with people, equipment, training, and values to affect how the department's staff implements its fire-fighting strategy and its educational fire-prevention strategy—outputs that further interact with the behavior of citizens to produce the de-

sired outcomes of fewer fires and fewer people injured or killed by fires that do occur.

Unfortunately, what is really going on inside the black box of any public agency is both complex and difficult to perceive. Much of it is going on inside the brains (often the subconscious brains) of the employees who work within the organization, the collaborators who somehow contribute to its outputs, and the citizens who convert these outputs into outcomes. Moreover, any single action may ripple through the agency, collaborators, and society as people adjust their behavior in response to seemingly small or irrelevant changes made by someone in some far-off corner. And when several people simultaneously take several actions, the ripples may interact in complex and unpredictable ways. It is very difficult to understand the black box adjustments and interactions that happen when just a few of the inputs (or processes) are changed, let alone when many of them are changing simultaneously and perhaps in undetected ways.[29]

Once the managers have figured out what is going on inside their black box, they have to figure out how the few things they can do are connected to the internal components they want to affect (because these components are, in turn, connected to the desired outputs or outcomes). How can changes in the budget's size or allocations affect people's behavior?[30] How can changes in one core process affect other processes? How can changes in one strategy support or undermine other strategies? How might they influence people's behavior?

Specifically, how might various leadership activities ripple through the black box? How might frequent, informal recognition of clear, if modest, successes or public attention to some small wins activate others? How might an inspirational speech or a more dramatic statement of the agency's mission affect the diligence, intelligence, and creativity of both organizational employees and collaborating citizens? To improve performance, public managers need measures that illuminate how their own activities affect the behavior of all of the various humans whose actions affect the outputs and outcomes they seek.

Meaningful Performance Measurement Requires a Gauge and a Context

Abstract measures are worthless. To use a performance measure—to extract information from it—a manager needs a specific, comparative gauge, plus an understanding of the relevant context. A truck has been driven 6.0 million. Six million what? Six million miles? That's impressive. Six million feet? That's only 1,136 miles. Six million inches? That's not even 95 miles. Big deal—unless those 95 miles were driven in two hours along a dirt road on a very rainy night.

To use performance measures to achieve any of these eight purposes, the public manager needs some kind of standard with which the measure can be compared.

1. To use a measure to *evaluate* performance, public managers need some kind of desired result with which to compare the data, and thus judge performance.
2. To use a measure of performance to *control* behavior, public managers need first to establish the desired behavioral or input standard from which to gauge individual or collective deviance.
3. To use efficiency measures to *budget,* public managers need an idea of what is a good, acceptable, or poor level of efficiency.[31]
4. To use performance measures to *motivate* people, public managers need some sense of what are reasonable and significant targets.
5. To use performance measures to *promote* an agency's competence, public managers need to understand what the public cares about.
6. To use performance measures to *celebrate,* public managers need to discern the kinds of achievements that employees and collaborators think are worth celebrating.
7. To use performance measures to *learn,* public managers need to be able to detect unexpected (and significant) developments and anticipate a wide variety of common organizational, human, and societal behaviors.
8. To use performance measures to *improve,* public managers need an understanding (or prediction) of how their actions affect the inside-the-black-box behavior of the people who contribute to their desired outputs and outcomes.

All of the eight purposes require (explicitly or implicitly) a baseline with which the measure can be compared. And, of course, the appropriate baseline depends on the context.

The standard against which to compare current performance can come from a variety of sources—each with its own advantages and liabilities. The agency may use its historical record as a baseline, looking to see how much it has improved. It may use comparative information from similar organizations, such as the data collected by the Comparative Performance Measurement Consortium organized by the International City/County Management Association (1999), or the effort to measure and compare the performance of local jurisdictions in North Carolina organized by the University of North Carolina (Rivenbark and Few 2000).[32] Of course, comparative data also may come from dissimilar organizations; citizens may compare—implicitly or quite explicitly—the ease of navigating a government Web site with the ease of navigating those created by private businesses.[33] Or the standard may be an explicit performance target established by the legislature, by political executives, or by career managers. Even to

control, managers need some kind of Tayloristic standard to be met by those whose behavior they seek to control. Whether public managers want to evaluate, control, budget, motivate, promote, celebrate, learn, or improve, they need both a measure and a standard of performance.

The Political Complexities of Measuring Performance

Who will pick the purpose, the measure, and the performance standard? The leadership team of a public agency has both the opportunity and the responsibility. But others—elected executives and legislators, political appointees and budget officers, journalists and stakeholders, and of course individual citizens—have the same opportunity, and often the same responsibility. Consequently, the agency's managers may discover that a set of performance measures has been imposed on them.

In some ways, however, public managers have more flexibility in selecting the performance measures that will be used by outsiders than do private-sector managers. After all, investment analysts long ago settled on a specific collection of performance measures—from return on equity to growth in market share—that they use when examining a business. For public agencies, however, no such broadly applicable and widely acceptable performance measures exist. Thus, every time those outsiders—whether they are budget officers or stakeholders—wish to examine a particular agency's management, they have to create some performance measures.

Sometimes, some will. Sometimes, a legislator or a budget officer will know exactly how he or she thinks the performance of a particular public agency should be measured. Sometimes, none will. Sometimes, no outsider will be able to devise a performance measure that makes much sense. Sometimes, many will. Sometimes several outsiders—an elected executive, a newspaper editor, and a stakeholder organization—will each develop a performance measure (or several such measures) for the agency. And when this happens, these measures may well conflict.

Mostly, these outsiders use their performance measures to evaluate, control, budget, or punish. Some might say, "We need this performance measure to hold the agency accountable." By this, they really mean, "We need this performance measure to evaluate the agency and if (as we suspect) the agency doesn't measure up, we will punish it by cutting its budget (or saying nasty things that will be reported by journalists)."[34] Outsiders are less likely to use performance measures to motivate, promote, or celebrate—though they could try to use them to force improvements.

Thus, the managers of a public agency may not have complete freedom to choose their own performance measures. They may have to pay attention to measures chosen by others. Even when they must respond to measures imposed by outsiders, however, the leaders of a public agency have not lost their obligation to create a collection of performance measures that they will use to manage the agency. The leadership team still must report the measures that outsiders are, legitimately, requesting. And they may be able to use some of these measures for one or more of their own eight purposes. But even when others have chosen their own measures of the agency's performance, its leaders still need to seriously examine the eight managerial purposes for which performance measures may prove useful and carefully select the best measures available for each purpose.

The Futile Search for the One Best Measure

"What gets measured gets done" is, perhaps, the most famous aphorism of performance measurement.[35] If you measure it, people will do it. Unfortunately, what people measure often is not precisely what they want done. And people—responding to the explicit or implicit incentives of the measurement—will do what people are measuring, not what these people actually want done. This is, as Steven Kerr, now chief learning officer at Goldman Sachs, wisely observes, "the folly of rewarding A while hoping for B" (1975). Thus, although performance measures shape behavior, they may shape behavior in both desirable and undesirable ways.[36]

For a business, the traditional performance measure has been the infamous bottom line—although any business has not just one bottom line, but many of them: a variety of financial ratios (return on equity, return on sales) that collectively suggest how well the firm is doing—or, at least, how well it has done. But as Kaplan and Norton observe, "many have criticized financial measures because of their well-documented inadequacies, their backward-looking focus, and their inability to reflect contemporary value-creating actions." Thus, Kaplan and Norton invented their now-famous balanced scorecard to give businesses a broader set of measures that capture more than the firm's most recent financial numbers. They want performance measures that answer four questions from four different perspectives:

- How do customers see us? (customer perspective)
- What must we excel at? (internal business perspective)
- Can we continue to improve and create value? (innovation and learning perspective)
- How do we look to shareholders? (financial perspective)

No single measure of performance answers all four questions (1992, 77, 72).

Similarly, there is no one magic performance measure that public managers can use for all of their eight purposes.

The search for the one best measurement is just as futile as the search for the one best way (Behn 1996). Indeed, this is precisely the argument behind Kaplan and Norton's balanced scorecard: Private-sector managers, they argue, "should not have to choose between financial and operational measures"; instead, business executives need "a balanced presentation of both financial and operational measures" (1992, 71). The same applies to public managers, who are faced with a more diverse set of stakeholders (not just customers and shareholders), a more contradictory set of demands for activities in which they ought to excel, and a more complex set of obstacles that must be overcome to improve and create value.[37] Consequently, they need an even more heterogeneous family of measures than the four that Kaplan and Norton propose for business.

The leaders of a public agency should not go looking for their one magic performance measure. Instead, they should begin by deciding on the managerial purposes to which performance measurement may contribute. Only then can they select a collection of performance measures with the characteristics necessary to help them (directly and indirectly) achieve these purposes.

Notes

1. Okay, not *everyone* is measuring performance. From a survey of municipalities in the United States, Poister and Streib find that "some 38 percent of the [695] respondents indicate that their cities use performance measures, a significantly lower percentage than reported by some of the earlier surveys" (1999, 328). Similarly, Ammons reports on municipal governments' "meager record" of using performance measures (1995, 42). And, of course, people who *report* they are measuring performance may not really be *using* these measures for any real purpose. Joyce notes there is "little evidence that performance information is actually used in the process of making budget decisions" (1997, 59).
2. People can measure the performance of (1) a public agency; (2) a public program; (3) a nonprofit or for-profit contractor that is providing a public service; or (4) a collaborative of public, nonprofit, and for-profit organizations. For brevity, I usually mention only the agency—though I clearly intend my reference to a public agency's performance to include the performance of its programs, contractors, and collaboratives.
3. Although Hatry provides the usual list of different types of performance information—input, process, output, outcome, efficiency, workload, and impact data (1999b, 12)—when discussing his 10 different purposes (chapter 11), he refers almost exclusively to outcome measures.
4. These eight purposes are not completely distinct. For example, learning itself is valuable only when put to some use. Obviously, two ways to use the learning extracted from performance measurements are to improve and to budget. Similarly, evaluation is not an ultimate purpose; to be valuable, any evaluation has to be used either to redesign programs (to improve) or to reallocate resources (to budget) by moving funds into more valuable uses. Even the budgetary purpose is subordinate to improvement.

   Indeed, the other seven purposes are all subordinate to improvement. Whenever public managers use performance measures to evaluate, control, budget, motivate, promote, celebrate, or learn, they do so only because these activities—they believe or hope—will help to improve the performance of government.

   There is, however, no guarantee that every use of performance measures to budget or celebrate will automatically enhance performance. There is no guarantee that every controlling or motivational strategy will improve performance. Public managers who seek evaluation or learning measures as a step toward improving performance need to think carefully not only about *why* they are measuring, but also about *what* they will do with these measurements and *how* they will employ them to improve performance.
5. Jolie Bain Pillsbury deserves the credit for explicitly pointing out to me that distinctly different purposes for measuring performance exist. On April 16, 1996, at a seminar at Duke University, she defined five purposes: evaluate, motivate, learn, promote, and celebrate (Behn 1997b).

   Others, however, have also observed this. For example, Osborne and Gaebler (1992), in their chapter on "Results Oriented Government," employ five headings that capture five of my eight purposes: "If you don't measure results, you can't tell success from failure" (147) (evaluate); "If you can't see success, you can't reward it" (148) (motivate); "If you can't see success, you can't learn from it" (150) (learn); "If you can't recognize failure, you can't correct it" (152) (improve); "If you can demonstrate results, you can win public support" (154) (promote).
6. Anyone who wishes to add a purpose to this list should also define the characteristics of potential measures that will be most appropriate for this additional purpose.
7. But isn't promoting accountability a quite distinct and also very important purpose for measuring performance? After all, scholars and practitioners emphasize the connection between performance measurement and accountability. Indeed, it is Hatry's first use of performance information (1999b, 158).

   In a report commissioned by the Governmental Accounting Standards Board on what it calls "service efforts and accomplishments [SEA] reporting," Hatry, Fountain, and Sullivan (1990) note that SEA measurement reflects the board's desire "to assist in fulfilling government's duty to be publicly accountable and … enable users to assess that accountability" (2). Moreover, they argue, without such performance measures, elected officials, citizens, and other users "are not

able to fully assess the adequacy of the governmental entity's performance or hold it accountable for the management of taxpayer and other resources" (2–3). Indeed, they continue, elected officials and public managers have a responsibility "to be accountable by giving information that will assist the public in assessing the results of operations" (5).

But what exactly does it mean to hold government accountable? In a 1989 resolution, the Governmental Accounting Standards Board called SEA information "an essential element of accountability." Indeed, in this resolution, the agency "gave considerable weight to the concept of accountability: of 'being obliged to explain one's actions, to justify what one does'; of being required 'to answer to the citizenry—justify the raising of public resources and the purposes for which they are used'" (Hatry et al. 1990, v). But does the phrase "hold government accountable" cover only the requirements to explain, justify, and answer? Or does accountability really mean punishment?

I find the use of the word "accountability" to be both ubiquitous and ambiguous. Yet it is difficult to examine how performance measurement will or might promote accountability without first deciding what citizens collectively mean by accountability—particularly, what we mean by accountability for performance. What does it mean to hold a public agency or manager accountable for performance? Presumably, this holding-people-accountable-for-performance process would employ some measure of performance. But what measures would be most useful for this promoting-accountability purpose? And how would those measures actually be used to promote accountability? (Or to revise the logical sequence: How might we use performance measures to promote accountability? Then, what measures would be most useful for promoting this accountability?) Before we as a polity can think analytically and creatively about how we might use performance measures to promote accountability, we need to think more analytically and creatively about what we mean by accountability. For a more detailed discussion of accountability—particularly of accountability for performance—see Behn (2001).

8. Joyce (1997) makes a similar argument: "The ability to measure performance is inexorably related to a clear understanding of what an agency or program is trying to accomplish" (50). Unfortunately, he continues, "the U.S. constitutional and political traditions, particularly at the national and state levels, work against this kind of clarity, because objectives are open to constant interpretation and reinterpretation at every stage of the policy process" (60).
9. Although a controlling approach to managing superior–subordinate relations may be out of style, the same is not necessarily true for how policy makers manage operating agencies. The New Public Management consists of two conflicting approaches: *Letting* the managers manage, versus *making* the managers manage (Kettl 1997, 447–48). And while the let-the-managers-manage strategy does, indeed, empower the managers of public agencies (for example, by giving them more flexibility), the make-the-managers-manage strategy does, in some ways, seek to control the managers. Yes, under a make-the-manager-manage performance contract, the manager has the flexibility to achieve his or her performance targets; at the same time, these output targets can be thought of as output "controls." I am grateful to an anonymous referee for this insight.
10. For a discussion of the pervasive efforts of public officials to control the behavior of their subordinates, see the classic discussion by Landau and Stout (1979).
11. For example, Robert Anthony's business texts on accounting include *The Management Control Function* (1988) and (with Vijay Govindarajan) *Management Control Systems* (1998). Similarly, his equivalent book (with David W. Young) for the public and nonprofit sectors is titled *Management Control in Nonprofit Organizations* (1999).
12. "Do management information systems lead to greater management control?" Overman and Loraine (1994, 193) ask this question and conclude they do not. From an analysis of 99 Air Force contracts, they could not find any relationship between the quality, detail, and timeliness of information received from the vendors and the cost, schedule, or quality of the project. Instead, they argue, "information can symbolize other values in the organization" (194). Still, legislatures, overhead agencies, and line managers seek control through performance measurement.
13. Melkers and Willoughby (2001) report that 47 of the 50 states have some form of performance budgeting, which they define "as requiring strategic planning regarding agency mission, goals and objectives, and a process that requests quantifiable data that provides meaningful information about program outcomes" (54). Yet when they asked people in both the executive and legislative branches of state government if their state had implemented performance budgeting, they found that "surprisingly, budgeters from a handful of states (10) disagreed across the branches as to implementation of performance budget in their state" (59). A handful? Melkers and Willoughby received responses from both branches from only 32 states. Thus, in nearly a third of the states that responded, the legislative-branch respondent disagreed with the executive-branch respondent. Not only is it difficult to define what performance budgeting is, it is also difficult to determine whether it has been implemented.
14. A more controversial use of performance measurement to motivate is the linking of performance data to an individual's pay. For a discussion, see Smith and Wallace (1995).
15. Williams, McShane, and Sechrest (1994) worry that "raw data may be misinterpreted by those without statistical training" (538), while at the same time "summaries of management information based on aggregate data are potentially dangerous to decision makers" (539). Moreover, they note the differing assumptions that managers and evaluators bring to the task of interpreting data: "The administrator often makes the implicit assumption that a project or operation is fully capable of succeeding" (541), while "the evaluator is apt to see the very core of his role as a questioning of the assumptions behind the project" (541). The evaluator starts

with the assumption that the program doesn't work; the manager, of course, believes that it does.

16. Understanding what is going on inside the black box is difficult in all of the sciences. Physicists, for example, do not know what is going on inside the black box of gravity. They know what happens—they know that two physical objects attract each other, and they can calculate the strength of that attraction—but they don't understand *how* the inputs of mass and distance are converted into the output of physical attraction. Newton figured out that, to determine very accurately (in a very wide variety of circumstances) the force of attraction between two objects, you need only three measurable inputs: the mass of the first object, the mass of the second object, and the distance between them:

    $F = G \times m_1 \times m_2 / d^2$

    (where G is the universal gravitational constant)

    Unfortunately for physicists, this universal law doesn't work at the subatomic level: Here, the classical laws of gravitational and electrical attraction between physical objects do not hold. Thus, when confronted with their inability to even calculate (using an existing formula) what is happening inside one of their black boxes, physicists invent new concepts (Behn 1992)—in this case, for example, the strong force and the weak force—that they can use to produce calculations that match the behavior they observe. But this does not mean that physicists understand what creates these inside-the-black-box forces.

17. My black box of performance management differs from the one defined by Ingraham and Kneedler (2000) and Ingraham and Donahue (2000). In their "new performance model," the black box is government management, and the inputs are politics, policy direction, and resources, all of which are imbedded in a set of environmental factors or contingencies. In my conception, the black box is the agency (or, more accurately, the people who work in the agency), the collaborative (that is, the people who staff both the agency and its various partners), or society (the collection of citizens). Management and leadership are inputs that seek to improve the performance of the black box by convincing the environment to provide better inputs and by attempting to influence the diligence, intelligence, and creativity with which those inside the black box convert the other inputs into outputs and outcomes.

18. If all of the data are indeed random, analysts who use the traditional 5 percent threshold for statistical significance and who check out 20 potentially causal variables will identify (on average) one of these variables as statistically significant. If they test 100 variables, they will (on average) find five that are statistically significant.

19. Perhaps this explains why formal performance evaluation has attracted a larger following and literature than has performance learning, let alone performance improvement.

20. A note of caution: Using outcomes to evaluate an organization's performance makes sense, except that the organization's performance is not the only factor affecting the outcomes. Yet, cognitive psychologists have identified the "outcome effect" in the evaluation of managerial performance. This outcome effect causes the evaluators of a manager's decision to give more weight to the actual outcome than is warranted given the circumstances—particularly, the uncertainty—under which the original decision was made. That is, when an evaluator considers a decision made by a manager, who could only make an uncertain, probabilistic guess about the future state of one or more important variables, the evaluator will give higher marks when the outcome seems to validate the initial choice than when the outcome doesn't—even when the circumstances of the decision are precisely the same (Baron and Hershey 1988; Lipshitz 1989; Hawkins and Hastie 1990; Hershey and Baron 1992; Ghosh and Ray 2000).

21. In business, Kaplan and Norton (1992) emphasize, the challenge is to figure out how to make an "explicit linkage between operations and finance" (79). They emphasize, "the hard truth is that if improved [internal, operational] performance fails to be reflected in the bottom line, executives should reexamine the basic assumptions of their strategy and mission" (77).

    The same applies in government. Public managers need an explicit link between operations and outcomes. If they use output (or process) measures to motivate people in their agency to ratchet up performance, and yet the outcomes that these outputs (or processes) are supposed to produce don't improve, they need to reexamine their strategy—and their assumptions about how these outputs (or processes) may or may not contribute to the desired outcomes.

22. In some ways, measures that are designed to motivate internal improvements in a public agency's performance appear to correspond to the measures that Kaplan and Norton design for their internal business perspective. Such internal measures help managers to focus on critical internal operations, write Kaplan and Norton. "Companies should decide what processes and competencies they must excel at and specify measures for each." Then, they continue "managers must devise measures that are influenced by employees' actions" (1992, 74–75). That is, to motivate their employees to improve internal operations, a firm's leaders need output measures.

23. In January 2002, when he announced his campaign for state treasurer, Daniel A. Grabauskas emphasized that, as the Massachusetts registrar of motor vehicles, he had cut waiting time by over an hour (Klein 2002).

24. For the general public, NAPA's advisory panel suggests, performance measures need to be suitably summarized (NAPA 1994, 23).

25. When attempting to select a measure to promote a public agency's achievements, it is not obvious which performance measure will capture citizens' concerns. What, for example, do the citizens visiting their state division of motor vehicles really care about? They might care about how long they wait in line. They might care less about how long they wait in line if they know, when they first get in line, how long they will have to wait. Some might say they will be quite

happy to wait 29 minutes, but not 30. Or, they might not care how long they wait as long as they can do it in a comfortable chair. Thus, before selecting a measure to promote the agency's performance, the agency's leadership should make some effort—through polls, focus groups, or customer surveys—to determine what the public wants.

Polls or focus groups, however, may produce only a theoretical answer. Have people who visit the DMV only biennially really thought through what they want? A customer survey—administered as people leave the DMV (or while they wait)—might produce a better sense of how people really measure the agency's performance.

26. Actually, managers don't need to identify *the best* practice. To improve their organization's performance, they need only to identify *a better* practice.
27. To promote the division of motor vehicles with the public, managers may simply publish the *average* length of time that citizens wait in line at the DMV office (compared, perhaps, with the average wait in similar lines). Waiting time is an easily understood concept. Yet, when an organization reports its waiting time, it rarely explains that this number is the *average* waiting time because this is what people implicitly assume.

    To learn, however, the *average* wait is too much of a summary statistic. The average wait obscures all of the interesting deviances that can be found only in the disaggregated data: What branch has a wait time that is half the statewide average (and what can the division learn from this)? What day of the month or hour of the day has the longest wait time (and what might the division do about this)? From such deviances, the DMV's managers can learn what is working best within their agency and what needs to be fixed.
28. After all, if any individual had expected a major deviance, he or she presumably would have done something to prevent it. Of course, an individual might have anticipated this deviance but also anticipated that it would be labeled a minor mistake rather than a huge failure (and thus they, too, have an opportunity to learn because, although they had expected the event, they did not expect it would be called a failure.) Or, an individual may have anticipated the failure but may not have been in a position to prevent it or to convince those who could prevent it that it would really happen. Or an individual may have anticipated the failure and hoped it would occur because its consequences (which were unanticipated by others) would further the individual's own agenda. Or an individual may have anticipated the failure but gambled that the probability of the failure, combined with its personal costs, were less than the certain personal costs of exposing his or her own responsibility for the causes behind this (potential, future) failure. Some people may have anticipated the failure, but certainly not everyone did.
29. The evaluator's ideal, of course, is that "only one new strategy should be introduced in one [organization]," while the baseline strategy "would go on in a demographically similar [organization]." Public executives, however, rarely are able to conduct the evaluator's "carefully controlled field experiment" (Karmen 2000, 95). Moreover, if the manager believes that two strategies will have an synergistic effect, he or she will—quite naturally—choose to implement them simultaneously in the same jurisdiction.
30. This suggests the limitations of performance budgeting as a strategy for improving performance: How much do budget officials know about how budget allocations affect the inside-the-black-box behaviors that improve performance? Do they really know which changes in the budget inputs will create the kind of complex inside-the-black-box interactions that can create improvements in organizational outputs, and thus societal outcomes?
31. Managers could simply allocate the available funds to the existing activities that are (using strictly internal comparisons) most efficient. Without some external standard of efficiency, however, they could spend all of their appropriations on completely inefficient operations.
32. Measuring performance against similar public agencies in a way that facilitates useful comparisons among jurisdictions is not easy. Agencies and jurisdictions organize themselves differently. They collect different kinds of data. They define inputs, processes, outputs, and outcomes differently. Consequently, obtaining comparable data is difficult—sometimes impossible. To make valid comparisons, someone must undertake the time-consuming task that Ammons, Coe, and Lombardo call "data cleaning" (2001, 102).

    Still, even when perfectly similar data have been collected from multiple jurisdictions, making useful comparisons is also difficult. Is one city reporting higher performance data for a particular agency because its leadership is more inspiring or inventive, because the agency has inherited a more effective organizational structure, because its political and managerial leadership has adopted a strategy designed to focus on some outcomes and not others, because the city council established the agency's mission as a higher priority, because the city council was willing to invest in modern technology, or because more of its citizens are cooperating fully? Even comparing performance measures for such a fundamental public service as trash collection is not simple. For a more detailed discussion of why benchmarking performance results among local governments may be more difficult than theorized, see Coe (1999, 111).
33. Criticism of public-sector service delivery has increased in the last two decades because a number of the traditional process measures for public-sector services—such as the time spent in a line or the ease of navigating a Web site—can easily be compared with the same process measures for the private sector, and because many private-sector firms have made a concerted effort to improve the process measures that customers value. Once people become accustomed to a short wait at their bank or when calling a toll-free number—once they learn that it is technically possible to ensure the wait is brief—they expect the same short wait from all other organizations, both private and public.
34. For a discussion of how accountability has become a code word for punishment and how we might make it mean something very different, see Behn (2001).

35. Peters and Waterman (1982, 268) attribute it to Mason Haire.
36. For example, in business, Kaplan and Norton write, "return-on-investment and earnings-per-share can give misleading signals for continuous improvement and innovation" (1992, 71).
37. Kaplan and Norton also argue their balanced scorecard "guards against suboptimization." Because the leadership is measuring a variety of indicators of the organization's performance, people in the organization will avoid focusing on one measure (or one kind of measure) at the expense of some others; after all, an "improvement in one area may have been achieved at the expense of another." And, even if a part of the organization chooses to focus on one performance indicator and ignore the others, the leadership—because it is measuring a variety of things—is much more apt to catch this suboptimal behavior (1992, 72).

References

Ammons, David N. 1995. Overcoming the Inadequacies of Performance Measurement in Local Government: The Case of Libraries and Leisure Services. *Public Administration Review* 55(1): 37–47.

Ammons, David N., Charles Coe, and Michael Lombardo. 2001. Performance-Comparison Projects in Local Government: Participants' Perspectives. *Public Administration Review* 61(1): 100–10.

Anthony, Robert N. 1988. *The Management Control Function*. Boston, MA: Harvard Business School Press.

Anthony, Robert N., and Vijay Govindarajan. 1998. *Management Control Systems*. 9th ed. Burr Ridge, IL: McGraw-Hill/Irwin.

Anthony, Robert, and David W. Young. 1999. *Management Control in Nonprofit Organizations*. 6th ed. Burr Ridge, IL: McGraw-Hill/Irwin.

Bardach, Eugene. 1998. *Getting Agencies to Work Together: The Practice and Theory of Managerial Craftsmanship*. Washington, DC: Brookings Institution.

Baron, Jonathan, and John C. Hershey. 1988. Outcome Bias in Decision Evaluation. *Journal of Personality and Social Psychology* 54(4): 569–79.

Behn, Robert D. 1991. *Leadership Counts: Lessons for Public Managers*. Cambridge, MA: Harvard University Press.

———. 1992. Management and the Neutrino: The Search for Meaningful Metaphors. *Public Administration Review* 52(5): 409–19.

———. 1996. The Futile Search for the One Best Way. *Governing,* July, 82.

———. 1997a. The Money-Back Guarantee. *Governing,* September, 74.

———. 1997b. Linking Measurement to Motivation: A Challenge for Education. In *Improving Educational Performance: Local and Systemic Reforms*, Advances in Educational Administration 5, edited by Paul W. Thurston and James G. Ward, 15–58. Greenwich, CT: JAI Press.

———. 1999. Do Goals Help Create Innovative Organizations? In *Public Management Reform and Innovation: Research, Theory, and Application*, edited by H. George Frederickson and Jocelyn M. Johnston, 70–88. Tuscaloosa, AL: University of Alabama Press.

———. 2001. *Rethinking Democratic Accountability*. Washington, DC: Brookings Institution.

Blodgett, Terrell, and Gerald Newfarmer. 1996. Performance Measurement: (Arguably) The Hottest Topic in Government Today. *Public Management*, January, 6.

Boone, Harry. 1996. Proving Government Works. *State Government News*, May, 10–12.

Bruns, William J. 1993. Responsibility Centers and Performance Measurement. Note 9-193-101. Boston, MA: Harvard Business School.

Coe, Charles. 1999. Local Government Benchmarking: Lessons from Two Major Multigovernment Efforts. *Public Administration Review* 59(2): 110–23.

Duncan, W. Jack. 1989. *Great Ideas in Management: Lessons from the Founders and Foundations of Managerial Practice*. San Francisco, CA: Jossey-Bass.

Feynman, Richard. 1965. *The Character of Physical Law*. Cambridge, MA: MIT Press.

Ghosh, Dipankar, and Manash R. Ray. 2000. Evaluating Managerial Performance: Mitigating the "Outcome Effect." *Journal of Managerial Issues* 12(2): 247–60.

Hatry, Harry P. 1999a. Mini-Symposium on Intergovernmental Comparative Performance Data. *Public Administration Review* 59(2): 101–4.

———. 1999b. *Performance Measurement: Getting Results*. Washington, DC: Urban Institute.

Hatry, Harry P., James R. Fountain, Jr., Jonathan M. Sullivan. 1990. Overview. In *Service Efforts and Accomplishments Reporting: Its Time Has Come*, edited by Harry P. Hatry, James R. Fountain, Jr., Jonathan M. Sullivan, and Lorraine Kremer, 1–49. Norwalk, CT: Governmental Accounting and Standards Board.

Hatry, Harry P., James R. Fountain, Jr., Jonathan M. Sullivan, and Lorraine Kremer. 1990. *Service Efforts and Accomplishments Reporting: Its Time Has Come*. Norwalk, CT: Governmental Accounting and Standards Board.

Hawkins, Scott A., and Reid Hastie. 1990. Hindsight: Biased Judgements of Past Events after the Outcomes are Known. *Psychological Bulletin* 107(3): 311–27.

Hershey, John C., and Jonathan Baron. 1992. Judgment by Outcomes: When Is It Justified? *Organizational Behavior and Human Decision Processes* 53(1): 89–93.

Holloway, Jacky A., Graham A.J. Francis, and C. Matthew Hinton. 1999. A Vehicle for Change? A Case Study of Performance Improvement in the "New" Public Sector. *International Journal of Public Sector Management* 12(4): 351–65.

Holt, Craig L. 1995–96. Performance Based Budgeting: Can It Really Be Done? *The Public Manager* 24(4): 19–21.

Ingraham, Patricia W., and Amy E. Kneedler. 2000. Dissecting the Black Box: Toward a Model and Measures of Government Management Performance. In *Advancing Public Management: New Developments in Theory, Methods, and Practice,* edited by Jeffrey L. Brudney, Laurence J. O'Toole, Jr., and Hal G. Rainey, 235–52. Washington, DC: Georgetown University Press.

Ingraham, Patricia W., and Amy Kneedler Donahue. 2000. Dissecting the Black Box Revisited: Characterizing Government Management Capacity. In *Governance and Performance: New Perspectives,* edited by Carolyn J. Heinrich and Laurence E. Lynn, Jr., 292–318. Washington, DC: Georgetown University Press.

International City/County Management Association (ICMA). 1999. *Comparative Performance Measurement: FY 1998 Data Report*. Washington, DC: ICMA.

Jordon, Meagan M., and Merl M. Hackbart. 1999. Performance Budgeting and Performance Funding in the States: A Status Assessment. *Public Budgeting and Finance* 19(1): 68–88.

Joyce, Philip G. 1996. Appraising Budget Appraisal: Can You Take Politics Out of Budgeting. *Public Budgeting and Finance* 16(4): 21–25.

———. 1997. Using Performance Measures for Budgeting: A New Beat, or Is It the Same Old Tune? In *Using Performance Measurement to Improve Public and Nonprofit Programs*, New Directions for Evaluation 75, edited by Kathryn E. Newcomer, 45–61. San Francisco, CA: Jossey-Bass.

Kaplan, Robert S., and David P. Norton. 1992. The Balanced Scorecard—Measures that Drive Performance. *Harvard Business Review* 70(1): 71–91.

Karmen, Andrew. 2000. *New York Murder Mystery: The True Story behind the Crime Crash in the 1990s*. New York: New York University Press.

Kerr, Steve. 1975. On the Folly of Rewarding A, While Hoping for B. *Academy of Management Journal* 18(4): 769–83.

Kettl, Donald F. 1997. The Global Revolution in Public Management: Driving Themes, Missing Links. *Journal of Policy Analysis and Management* 16(3): 446–62.

Klein, Rick. 2002. Registry Chief Quits to Run for State Treasurer. *Boston Globe,* January 10, B5.

Kopczynski, Mary, and Michael Lombardo. 1999. Comparative Performance Measurement: Insights and Lessons Learned from a Consortium Effort. *Public Administration Review* 59(2): 124–34.

Kouzmin, Alexander, Elke Löffler, Helmut Klages, and Nada Korac-Kakabadse. 1999. Benchmarking and Performance Measurement in Public Sectors. *International Journal of Public Sector Management* 12(2): 121–44.

Kravchuk, Robert S., and Ronald W. Schack. 1996. Designing Effective Performance Measurement Systems under the Government Performance and Results Act of 1993. *Public Administration Review* 56(4): 348–58.

Landau, Martin, and Russell Stout, Jr. 1979. To Manage Is Not to Control: Or the Folly of Type II Errors. *Public Administration Review* 39(2): 148–56.

Lehan, Edward Anthony. 1996. Budget Appraisal—The Next Step in the Quest for Better Budgeting? *Public Budgeting and Finance* 16(4): 3–20.

Levin, Martin A., and Mary Bryna Sanger. 1994. *Making Government Work: How Entrepreneurial Executives Turn Bright Ideas into Real Results*. San Francisco, CA: Jossey-Bass.

Lipshitz, Raanan. 1989. Either a Medal or a Corporal: The Effects of Success and Failure on the Evaluation of Decision Making and Decision Makers. *Organizational Behavior and Human Decision Processes* 44(3): 380–95.

Locke, Edwin A., and Gary P. Latham. 1984. *Goal Setting: A Motivational Technique That Works*. Englewood Cliffs, NJ: Prentice Hall.

———. 1990. *A Theory of Goal Setting and Task Performance*. Englewood Cliffs, NJ: Prentice Hall.

Melkers, Julia, and Katherine Willoughby. 1998. The State of the States: Performance-Based Budgeting Requirements in 47 out of 50. *Public Administration Review* 58(1): 66–73.

———. 2001. Budgeters' Views of State Performance-Budgeting Systems: Distinctions across Branches. *Public Administration Review* 61(1): 54–64.

Messadié, Gerald. 1991. *Great Scientific Discoveries*. New York: Chambers.

Murphey, David A. 1999. Presenting Community-Level Data in an "Outcomes and Indicators" Framework: Lessons from Vermont's Experience. *Public Administration Review* 59(1): 76–82.

National Academy of Public Administration (NAPA). 1994. Toward Useful Performance Measurement: Lessons Learned from Initial Pilot Performance Plans Prepared under the Government Performance and Results Act. Washington, DC: NAPA.

———. 1999. *Using Performance Data to Improve Government Effectiveness*. Washington, DC: NAPA.

Neves, Carole M.P., James F. Wolf, and Bill B. Benton. 1986. The Use of Management Indicators in Monitoring the Performance of Human Service Agencies. In *Performance and Credibility: Developing Excellence in Public and Nonprofit Organizations*, edited by Joseph S. Wholey, Mark A. Abramson, and Christopher Bellavita, 129–48. Lexington, MA: Lexington Books.

Newcomer, Kathryn E. 1997. Using Performance Measures to Improve Programs. In *Using Performance Measurement to Improve Public and Nonprofit Programs,* New Directions for Evaluation 75, edited by Kathryn E. Newcomer, 5–13. San Francisco, CA: Jossey-Bass.

Osborne, David, and Ted Gaebler. 1992. *Reinventing Government*. Reading, MA: Addison-Wesley.

Osborne, David, and Peter Plastrik. 2000. *The Reinventor's Fieldbook: Tools for Transforming Your Government*. San Francisco, CA: Jossey-Bass.

Overman, E. Sam, and Donna T. Loraine. 1994. Information for Control: Another Management Proverb? *Public Administration Review* 54(2): 193–96.

Peters, Thomas J., and Robert H. Waterman, Jr. 1982. *In Search of Excellence*. New York: Harper and Row.

Peters, Tom, and Nancy Austin. 1985. *A Passion for Excellence: The Leadership Difference*. New York: Random House.
Petroski, Henry. 1985. *To Engineer is Human: The Role of Failure in Successful Design*. New York: St. Martin's Press.
Poister, Theodore H., and Gregory Streib. 1999. Performance Measurement in Municipal Government: Assessing the State of the Practice. *Public Administration Review* 59(4): 325–35.
Rivenbark, William C., and Paula K. Few. 2000. *Final Report on City Services for Fiscal Year 1998–99: Performance and Cost Data*. Chapel Hill, NC: University of North Carolina–Chapel Hill, North Carolina Local Government Performance Measurement Project.
Rosenthal, Burton. 1975. *Lead Poisoning (A) and (B)*. Cambridge, MA: Kennedy School of Government.
Silverman, Eli B. 2001. *NYPD Battles Crime: Innovative Strategies in Policing*. Boston, MA: Northeastern University Press.
Sitkin, Sim B. 1992. Learning through Failure: The Strategy of Small Losses. *Research in Organizational Behavior* 14: 231–66.
Smith, Kimberly J., and Wanda A. Wallace. 1995. Incentive Effects of Service Efforts and Accomplishments Performance Measures: A Need for Experimentation. *International Journal of Public Administration* 18(2/3): 383–407.
Sparrow, Malcolm K. 2000. *The Regulatory Craft: Controlling Risks, Solving Problems, and Managing Compliance*. Washington, DC: Brookings Institution.
Standage, Tom. 2000. *The Neptune File: A Story of Astronomical Rivalry and the Pioneers of Planet Hunting*. New York: Walker Publishing.
Theurer, Jim. 1998. Seven Pitfalls to Avoid When Establishing Performance Measures. *Public Management* 8(7): 21–24.
Thompson, Fred. 1994. Mission-Driven, Results-Oriented Budgeting: Fiscal Administration and the New Public Management. *Public Budgeting and Finance* 15(3): 90–105.
Thompson, Fred, and Carol K. Johansen. 1999. Implementing Mission-Driven, Results-Oriented Budgeting. In *Public Management Reform and Innovation: Research, Theory, and Application*, edited by H. George Frederickson and Jocelyn M. Johnston, 189–205. Tuscaloosa, AL: University of Alabama Press.
Wholey, Joseph S. 1997. Clarifying Goals, Reporting Results. In *Progress and Future Directions in Evaluation: Perspectives on Theory, Practice and Methods*, New Directions for Evaluation 76, edited by Debra J. Rog and Deborah Fournier, 95–105. San Francisco, CA: Jossey-Bass.
Wholey, Joseph S., and Harry P. Hatry. 1992. The Case for Performance Monitoring. *Public Administration Review* 52(6): 604–10.
Wholey, Joseph S., and Kathryn E. Newcomer. 1997. Clarifying Goals, Reporting Results. In *Using Performance Measurement to Improve Public and Nonprofit Programs*, New Directions for Evaluation 75, edited by Kathryn E. Newcomer, 91–98. San Francisco, CA: Jossey-Bass.
Williams, Frank P. III, Marilyn D. McShane, and Dale Sechrest. 1994. Barriers to Effective Performance Review: The Seduction of Raw Data. *Public Administration Review* 54(6): 537–42.