

PerformanceStat as a Search for Strategic Evidence

Robert D. Behn
John F. Kennedy School of Government
Harvard University
79 John F. Kennedy Street
Cambridge, MA 02138
617-495-9874 redsox@hks.harvard.edu

A Paper Prepared for the Tenth National Public Management Research Conferen
October 1-3, 2009, Columbus, Ohio

What kind of data can be analyzed in what ways to help to determine the current level of performance, to ascertain what approaches (if any) are (or have been) working and why, to diagnose key performance deficits, to nominate opportunities for the next improvements, and to suggest strategies to be pursued?

“Sometimes you don’t know if there’s a real reason
or if it’s a small sample size.”
Terry Francona, manager, Boston Red Sox¹

On June 25, 2009, Robert Dunford, Superintendent in Chief of the Boston Police Department, opened the department’s bi-weekly CompStat session with a question: “Why? What’s going on on the street?” Shootings in Boston were off, and Dunford was looking to his department for an explanation. After all, as he told the 50 top managers in the department, “If we can identify what we are going, we can replicate it.”

Dunford’s question provoked a variety of possible answers: “Maybe the drug units” were having an impact noted one officer. “There’s hardly anyone at the usual spots,” observed another. One district commander suggested that it might be “aggressive patrol.” “Quicker indictments” proposed another, explaining that the grand jury was working better so that someone arrested on Friday night would be indicted on Tuesday, not six months later.

Indeed, Dunford’s question — and his search something that was replicable — generated numerous explanations that which could be organized into three broad categories:

- (a) It was the weather. During the previous weeks, Boston had experienced a lot of rain.
- (b) It was the action of the police — particular aggressive patrol and quicker indictments that were getting the “high-impact players” off the street.
- (c) It was purely random. As one officer noted: “Crime goes up. Crime goes down.”

As he had done before, Dunford was trying to get his department's leadership team to think analytically — to examine and learn from the available data. Not everyone in the room, however, was prepared to engage in this kind of discussion.

What is PerformanceStat?

PerformanceStat is a strategy that can be employed by the leadership team of a public agency or government jurisdiction to improve its own performance. It is not an effort by auditors, evaluators, inspectors general or other outsiders to measure and critique the performance of the organization.² Rather, PerformanceStat is an internal effort to produce better results. The line managers of subunits may view the PerformanceStat staff as outsiders; but they have been established not just to critique performance but to help improve it. Even so, different people can view any specific PerformanceStat differently:

PerformanceStat can be conceived as an *accountability mechanism*: We need to hold those lazy subordinates accountable (for doing what we have figured out we want them to do). From this perspective, data are a weapon that can be used to demonstrate subordinates inadequacy or incompetence.

Or PerformanceStat can be seen as a *computerized data base*: We need information on performance that subordinates can and will use to somehow improve their performance. From this perspective, data are collected and stored somewhere — with the hope that someday someone will look at and somehow use them.

Or PerformanceStat can be designed to be a *results-producing, results-improving strategy*: We need the active, analytical use of data to figure out what results the organization is currently producing as well as how it might produce better results in the future. From this perspective, someone (perhaps several someones) needs to collect *and* analyze the data, in an effort to uncover some insight into what is working, what isn't working, and how the organization might do more of what is (or might be) working.

Unfortunately, too many of the agencies and jurisdictions that have created something they call ***Stat (or claim is somehow a philosophical or operational descendent of Compstat) are mere mimicries. They miss one or more of the core concepts, key features, or essential components of this performance strategy. Consequently, to separate out those adaptations that are apt to have an impact on performance and those that are not, I have created the following definition of PerformanceStat:

A jurisdiction or agency is employing a PerformanceStat leadership strategy if it holds an ongoing series of *regular, frequent, integrated meetings* during which the chief executive and/or the principal members of the chief executive's leadership team plus the director (and the top managers) of different subunits — in an effort to achieve specific public purposes — use *current data* to analyze specific, previously defined aspects of each unit's past *performance*, to *follow-up* on previous decisions and commitments to produce *results*, to solve *performance-deficit* problems, to achieve its next *performance targets*, and to examine the effectiveness of its overall *performance strategies*.

This definition would appear not to be all that constraining. It could apply to a public agency and its AgencyStat or to a governmental jurisdiction and its JurisdictionStat. It could apply to a variety of public-sector performance strategies that was created without any knowledge of even the existence of CompStat or CitiStat.

The Search for Data, Evidence, Insight, and Strategy

To get better — to improve performance — an organization needs some indicator of its past, current, and future performance. It needs some data to tell it how well it has done in the past, how well it is doing now, and how well it is doing when it arrives at some future — next month, next quarter, next year. . . . And it needs to be able to mine such data for evidence about the effectiveness of its existing approaches for producing results — to figure out what is working, what is not working, and why. To employ a PerformanceStat leadership strategy, a public agency or governmental jurisdiction needs people with an analytical mindset.

Managers and their analysts need analytical skills to mine the data for evidence about what is and what isn't working — about what subunits are performing well, and what ones are not (either in comparison with each other or in comparison with some established standard). They also need imagination, creativity, and flexibility necessary to generate, from this evidence, insights about *why* a subunit is performing better or worse. And, from that insight, they need an understanding of human and organizational behavior to craft some new tactics and strategies that, taking advantage of the evidence and insights, can motivate significant performance improvements.

The Availability of Data in Policing

A police department is one of those rare public agencies that *automatically* — in the normal course of conducting its regular, day-to-day business — collects data on some key aspects of its performance — data that can prove helpful in an effort to analyze improve its performance. Every day, every police department collects data on its arrests — one of its key outputs. And, every day, every police department collects data on (reported) crimes — one of the outcomes about which citizens care.

Moreover, the dimensions of these data go far beyond mere aggregate counts. A police departments collects a lot of data about every arrest that it makes: who, where, for what crime(s), based on what evidence, and (eventually) on the disposition of the case. And, depending upon the information provided by the person(s) reporting a crime, a police department collects a little or a lot of data about the nature of each reported crime: what, where, when, by whom (either by name or demographic characteristics).

All this just happens. Bratton did not need to get NYPD's 76 precinct commanders to start collecting data, or start collecting any new or different data. When NYPD created CompStat, it and every other police department in the United States. already possessed data that related to performance.

A police department is not required, however, to do anything with these data. In the U.S., it can voluntarily forward these data monthly to the FBI, which compiles them into its UCR reports.³ How much use a department makes of these data is, however, a local choice. A department can assign a clerk to enter its data into a large data base and to submit them monthly to the FBI. Or department's leadership team can charge a team of analysts with the task of examining the data in an effort to determine the department's current level of performance, to diagnose key performance deficits, to nominate opportunities for the next improvements, to suggest strategies to be pursued, and to determine what (if anything) is working?

Just because a police department collects detailed crime and arrest data is no guarantee that it makes any use of (or even looks at) these data. Still, every police department does have some potentially useful performance data. But where? In the early 1990s, NYPD's data were in the precincts.

“Accurate and Timely Intelligence”

Jack Maple was obsessed with data — accurate and timely data. Again, the inspiration came to him while sitting in Elaine’s. On this particular night, he observed Elaine, who was certainly a hands-on manager of her restaurant, checking the tape that tallied of the evening’s receipts: “She knew at all times,” recalled Maple, “exactly how well the night was going.” And, “if the receipts were down, she’d look for her waiters: Were they loitering near the coffee service area or were they out on the floor, anticipating and attending to the customers’ every need.”⁴

Then Maple, himself, made an analytical comparison: “How different business was in the NYPD, I thought. We didn’t check our crime numbers hourly, daily, or even weekly. Headquarters gathered the numbers every six months.” Moreover, what did the precinct commanders do with their own data? “I couldn’t imagine that many of our precinct commanders checked the tape on their crime numbers every day.” And, Maple wondered, if they did, how did they use their data: “If the numbers were going the wrong way, would any of them be out on field inspections to determine what was going wrong?”⁵

To Maple, the contrast between Elaine’s and NYPD’s use of data was not just glaring. From the comparison, Maple quickly generated some insights about operations and some implications for strategy. NYPD — in both the precincts and city-wide — needed current data. And both needed to analyze these data in an effort to develop new policing tactics or strategies.

Specifically, Maple wanted these data in police headquarters. Conceptually, he wanted timely data. Operationally, he wanted weekly data. So he began requiring that each precinct commander deliver its data to One Police Plaza every Friday afternoon where the data had to be transcribed into a common format that would permit some kind of comparative analysis. Soon Maple created an electronic template; then the precincts were required to deliver a disk to headquarters every Friday. This solved the data-entry problem, and facilitated electronic analyses. Today, of course, the precincts simply enter their data on to the NYPD’s servers, and the analysts headquarters have instantaneous access to the latest data.

The Availability of CitiStat Data in Baltimore

In Baltimore, Mayor O’Malley didn’t have it so easy. The city’s agencies — like most public agencies — did not collect any data that had anything to do with the results they produced. This wasn’t unusual. In 2000, many (most?) public agencies were not collecting (let alone analyzing and using) any data that related to their performance.

Consequently, O’Malley’s CitiStat team began by simply telling the city’s departments: “Bring us whatever data you have.” Most departments did not have any data — outcome data, output data, or even activity or process data — that provided any insight into the results that they were (or might have been) producing. With only a few exceptions (such as police), city agencies did not have any useful performance data.

All organizations, however, do need to keep two kinds of data. Public, private, and nonprofit organizations all have to keep financial data. And any such organization larger than a family also has to maintain (if only to comply with tax-reporting requirements) personnel data. So that’s what city agencies brought to CitiStat: financial data and personnel data. As a result, one of CitiStat’s first initiatives was to reduce overtime. (After all, if a mayor can make a significant reduction in overtime expenses, that is the same as getting a budget increase.)

O’Malley’s overarching purpose did not, however, have anything to do with overtime. He was interested in service delivery. He wanted to get all agencies in city government to be diligent in responding to every citizen’s service request — or in Baltimore’s jargon to every “SR”. And as

the 311 and CitiTrack data systems came on line, O'Malley's CitiStat team began to get data related to the performance that he wanted to produce. Specifically, CitiTrack could take the information in the 311 system and create response-time data for each specific SR — particularly the response times for priority SRs. Indeed, the CitiStat analysts and the CitiStat meetings often focused on such data:

Was the agency meeting its SR targets?

If not: What new approach might it employ to improve its response times and achieve its target?

And what resources might be needed to implement this new approach?

Or is the response-time target unrealistic, and thus should it be relaxed?

If yes: Should the target response time be changed?

Or should resources be shifted to improve other response times?

Or should some other SRs be added to the priority list?

Finally, was the quality of the SR responses satisfactory?

An agency's SR data — when compared with its SR targets — would provide a basis for a conversation about how the agency might improve.

Data for What?

What data to collect? What data to analyze? This is rarely obvious. Sometimes the choice, at least initially, will be strictly opportunistic: What data do we have? Eventually, however, the PerformanceStat team will want to search for data that reveals how well it is doing in achieving its purpose.

Unfortunately, identifying such data (let alone collecting it) may be difficult. Indeed, the easily available data may be only loosely connected to the PerformanceStat purpose. And the data that are most directly connected to that purpose may be hard (or impossible) to obtain.

Rarely does an organization have ready access to data that (after some simple analysis) can answer — both for the entire organization and for each of its subunits — six key questions:

- (a) How well are it and they currently doing?
- (b) What performance approaches (if any) are (or have been) working?
- (c) Why are the successful approaches working, and why are the unsuccessful ones not?
- (d) What performance deficits do the organization and its subunits need to fix to improve?
- (e) What opportunities for such improvements should it and they tackle first?
- (f) What new strategies might it and they pursue?

Through analysis, data can help an organization identify what it is doing well so it can replicate it, and what it is not doing well so it can terminate it.

Still there is *no perfect performance measure*. Some measures will capture on certain aspects of performance while missing others. Some measures will suggest that things are currently improving while masking the significant deterioration in other (perhaps equally important) components of performance. Some measures will capture aspects of current performance but ignore important contributors to future performance. Some measures will create incentives for people to improve their performance-measurement scores yet contribute little to (or even undermine) the organization's efforts to achieve its purposes. Any performance measure comes with unintended (and sometimes perverse) incentives and consequences.⁶

Thus, from the PerformanceStat perspective (but not necessarily from the auditing perspective) the search is never for the [nonexistent] perfect measure. Rather, the PerformanceStat staff needs to look for the measure(s) that best help everyone in the organization understand how well it is performing and what it should do next to improve its performance.

What Data Will Help How?

When compared with a public agency, a private-sector organization has a much narrower and better defined purpose. Moreover, there exists a variety of well-established measures for determining how well a private-sector organization is doing in achieving its purpose. Yet within a business firm, report Marshall Meyer of the Wharton School and Vipin Gupta of Simmons College, "the most commonly used performance measures tend to be uncorrelated with one another," which (they argue) explains why business has created so many different performance measures. To illustrate their point, Meyer and Gupta "pose a hypothetical question: Imagine that financial performance and quality management were perfectly correlated. Would we need the Baldrige award?"⁷

"Market share" is often viewed as a useful and leading indicator of financial success: If a firm can increase its market share, goes the strategic logic, it can improve its long-run profitability. As Bryan Sharp and his colleagues at the University of South Australia's Marketing Science Centre write: "It has been known for many years that there is a positive relationship between market share and profitability."⁸

Yet Denise Rousseau of Carnegie Mellon University reports that "organizational success in obtaining market share often bears little relationship to other performance indicators."⁹ Stuart Jackson of L.E.K Consulting goes further, arguing that expanding market share can actually hurt a firm's long term financial prospects: "Market share, as most companies use it, is a misleading and dangers measure," writes Jackson. "Higher market share often means lower profits and lower return."¹⁰

If, however, business executives have a difficult time identifying useful performance-indicators, public-sector executives will find this challenge even more daunting. The task is to identify some performance data that are (a) correlated in some way with the organization's purpose, and (b) relatively easy to obtain. Finding data with both of these characteristics is rarely easy.

Still, an organization may occasionally discover a piece of data that is both easy to obtain and quite helpful in identifying potential performance problems.

Early Warning Indicators

On any given day, the New York City Department of Correction is responsible for the safe keeping of between 13,000 and 18,000 inmates — detainees who have been accused of a crime and are awaiting trial, or individuals who have been convicted of a crime and sentenced to less than a year in jail (and thus are not transferred to a state prison). Despite its name, the department is not in the correction business; inmates are not detained long enough for the department have any ability to improve their behavior when then reenter society. Rather, the department is in the housing (and transportation¹¹) business.

Thus, the department's primary responsibility is to keep its inmates, its staff, and the city's citizens safe from any violence. Indeed, on its Web site, the department lists (after its head-count data on inmate population admissions, and average length of stay) data on incidents of stabbings and slashings.¹² Thus, the Department of Correction has some advantages similar to those of police department. Everyone — particularly everyone who works in the prisons — agrees on which way its primary performance indicators should go (down not up). Moreover, during the normal course of doing its business, the department collects data on stabbings and slashings as well as other forms inmate-on-inmate, inmate-on-staff, and staff-on-inmate violence.

The Department of Correction calls its version of PerformanceStat T.E.A.M.S., for "Total Efficiency Accountability Management System." Think of it as CorrectionStat. And in the department's "Primary Indicator Report" which is prepared just before the monthly T.E.A.M.S. meeting, the first several pages of data are devoted to measures of "Security" — everything from stabbings and slashings, to weapons recovered (including shanks/shivs and razors), to visitor arrests.

These are after-the-fact data. They reveal how dangerous the department's facilities were last month or last year. Such aggregate data about past results have, however, little predictive power. Still they do provide the basis for further, more detailed analysis. If analysts can disaggregate the data — breaking them down to focus on particular inmates, or particular wards, or particular shifts, or even particular guards — they may discover some micro-patterns that suggest some small tactical modifications or major strategic shifts that will reduce some specific segments of the past violence.

As for predicting the future, the Department of Correction has uncovered a type of easily available and quite predictive indicator of potentially dangerous event: data on commissary sales to inmates. A slashing injures only one person (or maybe a few people). But a prison riot can harm many more people. And who would know when some group of inmates is planning a riot? Other inmates. And so what do these inmates do? They know that after the riot, they will all be locked in their cells. They won't be able to go to the prison's commissary. Not a happy time. So what do the inmates do? They stock up on whatever they would like to have in their cell during the post-riot lock-down.

For anyone managing a prison, a very useful indicator of future performance problems — a hint (though not proof) that some inmates may be planning a riot — is a spike in commissary sales.

Outputs vs. Outcomes

The mantra of performance measurement is well known: "Don't measure inputs. Don't measure process. Don't measure activities. Don't measure outputs. Only measure outcomes." Theoretically, this makes a lot of sense. The objective is to produce outcomes. And the causal connection between inputs and outcomes is, at best, uncertain. More inputs do not guarantee better outcomes. More cops does not guarantee less crime.

The same applies to processes, activities, and outputs — though as you move along the value chain from inputs, to processes and activities, to outputs, you ought to get a closer connection to organization's outcomes. Still, improved outputs does not ensure improve outcomes. The police can increase their outputs — more cops patrolling the streets; more detectives investigating each crime, more people arrested — and not reduce crime (let alone improve the citizenry's sense of safety).

Unfortunately, outcomes are difficult to measure. Moreover, even when you have outcome data, you still front two problems: the timeliness problem and the attribution problem.

First, there is the problem of timeliness. Maple wanted timely data. Delays reduce the value of the data. Yet, often the outcomes don't happen until years or decades later. Yes, we might be able to use the data on such outcomes to learn what strategies had worked and what hadn't. We could not, however, do that until much later. But how could we use that evidence to improve performance. The use of outcome data can create an unacceptably long delay in the feedback loop designed to improve performance now.

Second, there is the problem of attribution. A outcome is the result of an complex interaction between the outputs of a public agency (or agencies) and the actions of society. If outcomes improve, how much of that improvement can be attributed to the actions of government and how much to other, societal factors? Even the best research design often has a difficult time separating out the impact of influences that government agency does not control.

Both the timeliness and attribution problems confound any effort to determine the effectiveness of elementary and secondary education. The purpose of the public schools is (I believe) to help children grow up to be productive employees and responsible citizens.¹³ Unfortunately, it is quite difficult to determine whether a school district is doing this. Of course, there is the problem of deciding how to measure whether an adult is a "productive employee" let alone a "responsible citizen." But even if we could solve this problem and create some kind of easily collectable and widely accepted measure for these two outcomes, we would still confront the timeliness and attribution problems. First, we would be unable to collect these outcome data for years, indeed decades, after the (now former) children have left school. Moreover, if a student had grown up to be a notably productive employee and an especially responsible citizen, how much this individual's economic and civic contribution could we attribute to his or her schools and how much to family, or community, or church, or a single influential mentor.

To avoid the measurement, timeliness, and attribution problems, we don't even try to obtain any *outcome* data on schools. Instead, we create and collect output data: standardized test scores. And obviously, such output data have some deficiencies; the causal connection between these output data and the outcomes that schools seek to achieve is not direct, obvious, or proven. Specifically, it is not obvious that a child who scores well on such tests will grow up to be a productive employee, let alone a responsible citizen. The child could simply grow up to be a very clever con artist.

Performance Data for Public Welfare

In 1998, the New York City Human Resources Administration (HRA) created JobStat, its PerformanceStat strategy for its welfare-to-work effort. Similarly (though with no collaboration; indeed, without any interaction), the Los Angeles County Department of Public Social Services (DPSS) created, in 2005, DPSSTATS (for "DPSS Total Accountability, Total Success").¹⁴ Both of these agencies have, as an important purpose, helping families move from dependence on welfare to economic and psychological independence.

To some, this purpose might be controversial or even wrong. But both of these organizations accept this purpose. In New York, HRA emphasizes its “commitment to move cash assistance recipients to employment” observing that it “provides temporary help to individuals and families with social service and economic needs to assist them in reaching self-sufficiency.”¹⁵ In Los Angeles, DPSS states that its “mission” is: “To provide effective services to individuals and families in need, which both alleviate hardship and promote personal responsibility and economic independence. . . .”¹⁶

But what to measure? How would — how could — the leadership of either agency collect data that could suggest whether or not it was doing a good, better, improving, or deteriorating job at achieving this mission? What kind of data could either agency collect and then analyze to help it determine its current level of performance, to ascertain what approaches (if any) are (or have been) working and why, to diagnose its key performance deficits, to nominate opportunities for the next improvements, to suggest strategies to be pursued? What kind of data might either agency find somewhat (somehow) useful to analyze? It isn’t obvious.

After all, what outcome data could you collect and then analyze in a way that could provide useful feedback to improve performance? Almost any kind of outcome data would present the agency with both the problem of delay and the problem of attribution. Consequently, both NYC’s HRA and LA’s DPSS are forced to rely on something other than outcome data.

Moreover, from the perspective of citizens and voters, there is a big difference between a police department and a social service department. For the police, the public cares about crime. Citizens are concerned about when what crimes are committed (and the newspapers report such data), and how many people are arrested for these crimes. That is, for a police department, the citizens personally care about both the outputs and the outcomes. For a social service department, however, there exists no simple, obvious and useful data on outcomes (how many people’s lives are improved by how much?), and the output data (how many people received how many benefits?) is not something that affects the average voter either directly or daily.

For both police and social service departments, citizens and legislators care about the rules — or, at least, about whether too many rules were violated. So, to stay out of trouble, both of these kinds of public agencies need to monitor how faithfully they comply with the rules.

In addition, both NYC’s HRA and LA’s DPSS administer programs established by the U.S. national government. These means that they are required to report to the Feds a variety of input, activity, process, and output data. Moreover, if some of these data are outside the permissible range, the agency can be penalized financially. Consequently, both NYC’s HRA and LA’s DPSS are directly and immediately concerned about any federally prescribed criteria for input, activity, process, or output data — regardless of whether such data provide any information about how well they are doing in achieving their mission.

In 2009, HRA’s JobStat sessions covered 35 different measures, which are listed on the monthly JobStat Report in five different categories: Application Process, Employment Process, Case Management Process, Fair Hearing, and Placements/Participation (which is the category into which any output, though not outcome, data would fall). The first measure under Placements/Participation is the number of “qualified reported placements” of public assistance beneficiaries into employment. This is the JobStat’s primary output (though not outcome) measure. For this placement measure, each Job Center has a target, that depends upon the size of its caseload.

The other data that HRA lists in its JobStat Report include a variety of processing rates such as the “Child Support (OCSE [Office of Child Support Enforcement]) Referral Rate,” the “Food Stamp Application Timeliness Rate,” and the “Cash Assistance Fair Hearing Comply Rate.” All of these have a target rate, as well as a minimally acceptable rate. For example, for the “Cash Assistant Fair Hearing Comply Rate,” the target is 95% and the minimally acceptable rate is 80%.¹⁷

Indeed, all of the other data on the JobStat Report are activity, process, or administrative measures.¹⁸

In Los Angeles County, the Department of Public Social Services collects and publicizes a variety of data. Its quarterly report on “Caseload Characteristics” runs over 450 pages and describes those receiving assistance by age (under 1, 1-2, 3-5 . . . over 65), by primary language (Armenian, Cambodian, . . . Tagalog, Vietnamese), gender, citizenship status, and ethnic origin. Moreover, DPSS reports each such datum for the entire county, plus by different geographic subdivisions (including departmental district, service planning area, congressional district, state senate district, assembly district, and supervisorial district).¹⁹ DPSS also prepares and publishes a monthly “Statistical Report” that includes the same county-wide caseload information, plus data on application processing and terminations. In addition, this report includes a variety of graphs showing ten-year trends in monthly data for total caseload as well as “persons aided” for different programs (CalWORKS, General Relief, . . . In-Home Supportive Services) as well as the county’s monthly unemployment rate.²⁰ Finally, the department reports, “DPSSTATS maintains an inventory of more than 160 different performance measures.”²¹

That’s a lot of data to collect, let alone to analyze, digest, and discuss. Obviously, the department does not discuss all 160 performance measures at every DPSSTATS meeting. Yet, any organization that has easily available so much data — can there be anything such as too much data? — has to figure out how to focus. What is important? What is irrelevant? And, most difficult, what data are, while very important, not quite as important as the core data that provide real information about how well the organization is doing in achieving its purpose, and what it should or might be doing differently — and thus better?

DPSS addresses this problem in two ways. First, to ensure that its local offices concentrate on some core activities, it collects data on eight activity measures covering customer service (“seen in 20 minutes” plus “participant satisfaction”), application progressing (for CalWORKS, General Relief, Food Stamps, and Medi-Cal), Medi-Call Redetermination Processing, and the Food Stamp Error Rate. If an office earns a satisfactory score on all of its measures, it earns a “star” that is prominently displayed on its name plaque at the next DPSSTATS meeting.²²

For example, one of these eight measures is “Food Stamp Applications Processed Timely.” The target is 90 percent processed within 30 days. And to dramatize how well the department, its four divisions, and its 26 offices are doing, the DPSSTATS staff creates a variety of bar charts. For example, for each division, there is a bar chart showing, for each office in the division, the number of applications that have been pending for 31-to-45 days, 46-to-90 days, 91-to-180 days, and 181+ days.

In addition, two months before each upcoming DPSSTATS meeting, the staff gets together with Phil Ansel, the department’s assistant director for program and policy. The purpose is to select — from the list of 160 different types of data that it tracks — the measures on which to focus. Essentially, the staff is deciding on the “performance deficits” (my words, not theirs) on which it and the managers of the divisions and local offices need to concentrate next: From all of the many different ways to think about (and collect data on) the department’s performance, what does it need to fix next?

Still, for public welfare agencies, the data that are available in a timely manner are often limited to input, activity, process, and outcome data. Indeed, this is true of many public agencies. And, unfortunately, the causal connection between such operational data and the outcomes that the agency is attempting to produce is rarely direct, obvious, or proven.

The Need For Output-to-Outcome Theory

In the public sector, the best direct connection with which I am familiar between an agency's operational output and the desired societal outcome is the vaccination of children to immunize them against measles. Why? Because the vaccination output is closely and causally connected to the immunization outcome. As the Centers for Disease Control reports, the production of the output has a greater than 99 percent chance of producing the desired outcome: "Studies indicate that more than 99 percent of persons who receive two doses of measles vaccine (with the first dose administered no earlier than the first birthday) develop serologic evidence of measles immunity."²³ Not too many public agencies can report that one of their outputs has a 99 percent chance of producing the intended outcome.

Moreover, even if a child is the one-in-a-hundred for whom the vaccine does not produce the desired immunization, he or she is better off because of the government's immunization effort. For this un-immunized child is much less likely to ever come in contact with another child who does have the measles. This additional benefit has been called the "herd effect."²⁴

Thus, if public health agency sets a performance target to vaccinate 100 percent of the children living within its jurisdiction — or even just 90 percent of them — and if it achieves its target, the agency has had a significant impact on the immunization and thus health of these children.

Of course, public-health agencies benefit from the research of the public-health professionals. And for some specific interventions, such as vaccination protocols for different diseases, public-health professionals can collect (or even generate) a lot of data that can reveal whether government's operational output (vaccination) generates society's desired outcome (immunization).

In most circumstances, however, government isn't so fortunate. Not only is government's output only one factor affecting the desired outcome. It is also difficult to determine how much — if anything — of the societal outcome can be attributed to the government's output.

When a public agency can measure (in a timely way, at least) only its outputs but not its outcomes, it needs a cause-and-effect theory — a theory that explains why its outputs are producing, or at least contributing to the desired outcomes. In public health, it is possible to test such theories rigorously, and thus create quite specific theories about cause and effect — theories that include well documented probabilities.

For most public agencies, however, the theory will remain a theory. It will be possible to collect data that generates evidence suggesting that the theory has some value. But it will rarely be able to prove the theory. Still, the data can be analyzed to determine how much evidence there is behind the theory.

Compared With What?

Any analysis involves a comparison. *What*, however, should be compared with *what*? Deciding what kind of data would be most helpful in improving performance, still leaves the key question: "Compared with what?" Indeed, the choice of data may depend as much on the opportunities for meaningful comparisons as for the ability of the data to directly answer the six key questions. In fact, any data can help answer any of these questions only if they can be compared with other similar data.

For a basis of comparison, there exist numerous possibilities. Current performance could be compared with the organization's own historical performance to see if the organization is

improving or not. Current performance could be compared with the performance of similar agencies or jurisdictions, to see if the organization is doing better or worse than the rest of the world. Or current performance could be compared with some preestablished performance target, to see if the organization is living up to its commitments. The choice isn't obvious.

After all, sometimes there are data problems, sometimes there are comparison problems, and sometimes there are both. The organization may not have collected data on its past performance; if so, there exists no way to compare current performance with historical performance. Undoubtedly there exist similar organizations, yet none of them have precisely the same problems, circumstances, or resources; thus any differences found in the data could be attributed to differences in performance, differences in operational circumstances, or political support. Any organization seeking to improve performance will, eventually, need to set some specific targets. Nevertheless, if the organization achieves its targets, is this because it improved or because its target was set too low. Conversely, if the organization fell short of its target was this because it failed to perform adequately or because the target was set unrealistically high.

For any comparative differences among these kinds of data, there exist always multiple, alternative (and competing) interpretations.

The Gold Standard of Comparison

Riding to the rescue comes the gold standard for comparisons: The double-blind, randomized, placebo-controlled experiment. If done properly, this experimental approach can eliminate all but one of the alternative explanations.

The classic example is the experiment conducted in the 1954 to determine whether the polio vaccine developed by Jonas Salk was effective. Some children were given the vaccine (and only the vaccine); others were given an identical looking placebo (a salt solution). The choice about who would receive the vaccine and who would not was done randomly; no human could influence (even unconsciously or indirectly) the choice. Moreover, neither the children, their parents, the individuals who administered the vaccine, nor the doctors who examined the children knew whether the child received the vaccine or the placebo. As *The American Journal of Public Health* observed in an editorial, the experiment demonstrated that the "vaccine was notably harmless and the benefit undoubtedly significant, especially against the paralytic form of the disease."²⁵

The polio field trials were, however, quite controversial.²⁶ Salk had created a "killed-virus vaccine," which some worried might actually give children polio. Moreover, the field trials were originally to be based on a "observed control" design. Children in the second grade would receive the vaccine; children in the first and third grades would not, and the incidence of polio in the two groups would be compared. In the end, the actual trials included both designs. The observed-control approach was employed in 33 states; over a million first, second, and third graders in 127 test areas participated this party of the study. The placebo-control design was used in eleven states; the participants included 750,000 first, second, and third graders in 84 test areas.²⁷

The randomized, placebo-control design was clearly superior. Half the children got the vaccine; the other half were denied it. Because the assignment of who got the vaccine and who got the placebo was made purely randomly, no human bias — subtle or subconscious — could influence who got the vaccine and who could not.²⁸ No one else had access to the vaccine, eliminating the possibility that some parents, fearing that their child had only received the placebo, could serendipitously obtain a vial of the vaccine. Moreover, none of the children, and none of those who had contact with them, knew who got which; this eliminated the possibility that children or parents or medical professionals who knew who did and did not receive the vaccine would, though some subtle, psychological influences stimulate physiological reactions in the children or prejudice any physician's diagnosis. Finally, each child in the experiment received either the

vaccine or the placebo, but no other intervention, thus isolating the vaccine as the only treatment and permitting the experiment to compare just two (each very well defined) alternatives.

The experiment required an elaborate record-keeping system to keep track of which children got the vaccine and which got the placebo. Still, if the experiment discovered that those who received the vaccine and those who got the placebo had a statistically different incidence of contracting polio, there could be only one possible explanation for this difference: the vaccine.

The double-blind, placebo-controlled experiment eliminates multiple interpretations of any difference in the outcomes for the treatment and control groups. This is the “gold standard” that every policy researcher would like to employ to determine whether some policy intervention or management strategy is really working.

The Advantage of AgencyStat (compared with JurisdictionStat)

For this task of comparing performance, a public agency employing an AgencyStat strategy has a distinct advantage over a governmental jurisdiction employing a JurisdictionStat approach. The agency has several (maybe many) subunits all of which have been given the same public purpose, have been assigned the same operational responsibilities, and collect the same data concerning their performance. Thus, there exists an obvious answer to the compared-with-what? question. The agency’s leadership team and the AgencyStat staff possess a built-in basis for comparing, analyzing, appraising, and diagnosing the performance of their subunits. (Also, these subunits can learn from each other.)

In contrast, for a JurisdictionStat, each subunit is different: It has an entirely different mission, produces entirely different results, and thus has an entirely different basis for analyzing, appraising, and diagnosing its performance. The jurisdiction’s leadership team and the JurisdictionStat staff cannot compare the performance of the police department with the performance of the fire department.

This is why Baltimore’s CitiStat needed performance targets for each agency. At the city-wide level, Baltimore a basis of comparison for each department. And although many city departments do have geographic subunits, none of them come close to having 76 such subunits. Thus, to answer the compared-with-what? question, Baltimore created service-delivery performance targets. Each department would have its performance — its ability to response to citizen service requests — compared with a specific target, a deadline by which time each citizens request for a specific service needed to be completed. And, because every department would get many such requests — particularly for its priority SRs — the CitiStat staff could aggregate the data into meaningful summary statistics: *average* time to complete the SR (compared with the target); *percent* completed by the target date; and *number* of SRs not completed by the target date.

Actually, even some agencies employing the PerformanceStat strategy can confront the problem of creating a basis of comparison for some its units. For the subunits with line (operational) responsibilities, there exists a clear basis of comparison; that is the performance of the other (similar) line units. That makes it easy (or at least easier) to use AgencyStat to drive their performance of line units. For their staff units, however, there is no such basis of comparison; there is only one budget office, only one personnel office, only one legislative-liaison office. Thus these staff units can come to any AgencyStat session armed with a common, yet not-easily-counteracted, excuse about why their performance is adequate: “We’re different.”

At the Los Angeles County Department of Public Social Services, the DPSSTATS staff collects a wide variety data concerning the performance of different units, and the department’s leadership conducts monthly meetings for both line units and staff units. And, reports Karen Kent, the director of DPSSTATS, her department’s version of PerformanceStat works much better

for line units than staff units, precisely because there is no obvious basis of comparison for the various staff units.

Maps, Dots, Cops, and Analysis

Among the four reasons why police departments would be more likely than other public agencies to invent a PerformanceStat-like strategy is availability of a simple, first-order analytical tool. Putting the dots of crimes on a map.

Putting dots on a map is not, however, a revolutionary analytical technique. Indeed, putting dots on a map wasn't even invented by the police. It started in public health. John Snow is famous for his 1854 mapping of the location of cholera deaths in London, and to using that the concentration of deaths on a map to demonstrate the local source of the disease: the Broad Street pump.²⁹

Indeed, even within policing, putting dots of crimes on maps did not originate with CompStat. When Bratton was a young police lieutenant in Boston, he had "requisitioned gigantic maps" of his district, "papered all four walls with them," and then "put up dots: red dots for burglaries, blue dots for robberies" — earning him the nickname "Lord Dots."³⁰

The beauty of the maps is, of course, that you don't need a degree in econometrics to use the map to see where the dots — and thus the crimes — are clustered. From there it is a simple logical step to Maple's mantra: "Map the crime and put the cops where the dots are.' Or more succinctly: 'Put cops on dots.'"³¹

Today, of course, the maps and the dots are computerized. In the Los Angeles Police Department, these maps are quite sophisticated. LAPD doesn't just use circular dots. The department displays triangles for robbery, squares for aggravated assault, circles for burglary, stars for auto theft auto, and diamonds for burglary theft from motor vehicle. A small circle with a smaller diamond inside indicates that a shot was fired, and if this was a gang crime, its also gets a six-pointed star. Finally, each of these various "dots" is colored coded by watch: red for midnight to 6:00 a.m.; orange for 6:00 a.m. to noon; green for noone to 6:00 p.m., and blue for 6:00 p.m. to midnight.³²

Still, the basic analytical and strategic insight remains the same: Put the cops on the dots.

Putting cops on the dots does not, however, work for every crime. Mapping white color crimes would not be of much help. If the New York Police Department mapped securities fraud and discovered that most of its was located in the Wall Street area of lower Manhattan, what would it do? Have a bunch of cops hang out at 20 Broad Street in front of the New York Stock Exchange. That would neither deter nor catch many security-fraud criminals.

Yet, when other public officials create their own JurisdictionStat or AgencyStat for a different kind of public agency, they often mindlessly copy the mapping strategy. Even if the purposes they are trying to achieve and the performance deficits they are trying to eliminate or mitigate will not be affected by putting anything on the any collection of dots, they nevertheless create maps.

Still, this mimicry is quite explicable. After all, the PerformanceStat strategy first evolved within police departments, and almost all police departments that have employed CompStat (and many others too) have used maps with dots. Moreover, almost all (non-police) organizations that have created their own PerformanceStat have learned the strategy either directly or indirectly from a police department's Compstat. Finally, most governments now possess the technological capacity (using GIS) to create their own maps.

The maps may look nice. They may wow people impressed by high-tech gadgetry. But will they actually help? The maps are only valuable to the extent that they help reveal patterns and suggest new or better targeted strategies.

Disaggregating the Data

Analysis requires the disaggregation of the available data. The summation of a city's total crime reveals very little. Yes, when compared with last month's or year's summation of total crime, this aggregated data will reveal whether crime is up, down, or unchanged. But such totals cannot suggest what to do. Only by disaggregating the data can the analyst identify patterns, and perhaps causes, and thus maybe strategies for dealing with the causes of the patterns?

Along what dimensions, however, should the data be disaggregated. This is rarely obvious. Again, however, the police have an advantage. You don't have to pass the civil service test for detective to recognize the value of disaggregating the data by type of crime. If homicides are down but auto theft is up, the police have an initial clue about how to allocate their resources.

Crime maps with dots are also useful first-order analytic tool. They disaggregate the data on crime by location. This permits anyone looking at the map to observe the geographic patterns and, then, to deploy the police to the high-crime locations. And if the maps also disaggregate the data by type of crime (robbery, aggravated assault burglary, . . .) and by time of day, they provide even more detailed "picture" (literally) of the patterns and thus suggest more nuanced strategy.

Still, a map is only one analytic tool. Often, it will not suggest a quick, simple, and effective strategy for dealing with a problem — or even suggest what the problem really is. More often, it will only suggest the need to apply some additional, more detailed analytic approach to understand the nature of the performance deficit. The map is only one analytic tool. It is certainly not the only analytic tool that can be employed to identify and understand specific performance deficits let alone to suggest specific strategies for improving performance in all policy and management areas.³³

The Search for Useful Analytic Tools

How should a PerformanceStat's analytic staff conduct its analysis? In particular, what analytical techniques should they use?

These questions have no answer. There is no officially certified PerformanceStat analytic methodology, nor should there be.³⁴ There exists no single computational routine, no scientific procedure, no mathematical tool, no statistical template, no one best analytic practice that will reveal an organization's every performance deficit and simultaneously suggest the perfect strategy for eliminating each of them. Even within a specific policy or programmatic area there exist no one best analytic practice.

Still some analytical approaches might have proven more useful than others. What are they? Again, this question has no definitive answer. In fact, the most honest and (perhaps?) least helpful response is the ubiquitous: It all depends. Unfortunately, it is never obvious on what exactly it does or might depend.

An analytical methodology might appear to illuminate patterns and still provide absolutely no hint about what to do, no guidance about possible approaches, not even a clue as to the underlying nature cause of the patterns, and thus little information about the real performance deficit(s). Some education in analytical techniques might prove useful by suggesting what

approaches might provide some insight. At the some time, the formal analytical techniques learned in graduate school might produce a variety of numbers but no real understanding of the performance problem — or even a suggestion that the performance problem might be. Each specific analytic technique is designed to understand (and, perhaps, help solve) specific problems. But if the nature of the problem is unknown — indeed, if it is not even known if a problem exists, let alone what that problem might be — it is difficult to know which technique to select from one’s analytical toolbox.

Fortunately, there does exist one analytical technique that every PerformanceStat analyst knows and that can, moreover, prove useful in a wide variety of circumstances. Indeed, this analytic technique is some elementary that it is known by every public manager too. It is so elementary that every PerformanceStat analyst and every PerformanceStat executive learned this technique in elementary school. This most valuable analytic technique is: Long Division.

Long division provides a basis for comparing two numbers: Outputs with inputs. Outputs over time. The outputs of organization A with the outputs of organization B. Long division creates ratios that can be undeniably revealing or insipidly uninteresting. Long division might produce a number this is astonishingly big or startlingly small. Or it might produce a number that is very close what everyone had already guessed. Unfortunately, before you actually do the calculation, you don’t know which. If you did, you won’t have to bother to do the math.

But what to divide by what? Again — an unfortunately — it all depends. Yet, it is never obvious. Any organization whose performance is worth analyzing comes a large array of data, with a large variety of numbers. These numbers include data that the organization routinely collects, data that exist but that no one has thought to connect, plus data that could be collected if someone made the effort to establish the necessary mechanisms. And the analyst — or the manager — can divide any one of these numbers any other number. But so what? What does — or might — that reveal. Who knows?

In some cases, experience may help the PerformanceStat analyst by suggesting what might be usefully compared with what. After all, if you have worked in the world of policing you have accumulated a variety of intuition, insights, and instincts that permit you to discern what to go looking for and how to find it. Experience ensures that the analyst approaches the search for insights into the nature of the organization’s performance deficit with both knowledge about how the organization (and the individuals within it) behave as well as hunches about where to go looking for new intelligence about what is going on.

Experience, however, also creates blinders. Thus naivete may help too. The innocent neophyte will ask questions that the experienced professional knows aren’t worth asking. The neophyte might create ratios that the pros assume will never be informative. Naivete ensures the analyst approaches the search for insights into the nature of the organization’s performance deficit with both a lack of subconscious prejudices about how the organization and its people are behaving as well as few preconceptions about where to do go looking for new intelligence about what is going on.

Whoever is doing the analysis — be it the experienced professional or the naive neophyte — needs to possess what my Kennedy School colleague Malcolm Sparrow calls the “habits of mind” and “patterns of thought” that lead them to “slice and dice” the data, “molding and testing different problem-definitions and specifications.”³⁵ They have to be prepared to immerse themselves in the details — to look at the existing data from a variety of perspectives and to go looking for other data that might (but only might) reveal something interesting.

Thus, in addition to not being obvious what to divide by what, it is also not obvious to whom to give the long-division assignment. Ideally, a PerformanceStat staff would consist to two people (or, if it had more than two people, two types of people): the experienced professional who

knows exactly where to go looking for useful ratios, and the innocent neophyte who applies long division to a variety of numbers with no preconception about what the resulting ratio might reveal.

Sparrow, is found of saying: “Identify important problems and then fix them.” The operation challenge is to get the organization to fix the identified problems. But first comes the analytical challenge: Getting someone — analyst or manager — to identify an important problem. And the task of identifying problems may come with fewer suggestions for analytical approaches than the task of fixing them comes with suggestions for operational approaches. In a large, complex organization, identifying and understanding truly important problems — problems that are worth fixing — may be similar to looking for a needle in a haystack — except that you don’t know what the needle actually looks like (are you looking for a number, a virus, or an asteroid?), or what the haystack looks like (is it disguised as baseball park?) or even where this haystack might be located.

Analysis as a Search for Patterns and Strategies

One of the analytical choices concerns the level at which to disaggregate the data. Should the analyst look at the most macro data for the jurisdiction? Or should the analyst disaggregate the data into its most micro components? Once again, the answer is rarely obvious. Sparrow writes about “the art of picking the right levels of aggregation” and suggests that this should be somewhere in between “broad, general phenomena” and the “minutia” of individual events.³⁶

For example, at the end of the year, a police department could collect and report the annual city-wide data for each type of crime. From such a report, any analyst (indeed, anyone) could determine whether *total* crime (and every individual type of crime) went up or down during the year. But the analyst could never discern from such a macro-level aggregation of the data anything about the jurisdiction’s patterns of crime. Such summary data would provide no hint about the department’s performance let alone how they might seek to reduce them. Annual reports contain little or no analysis of the data that can be employed to improve performance.

Alternatively, the police department’s analyst could disaggregate the data at a very fine level in both geography and time. For example, the analyst could break the data down by every street address and by every hour (or minute) of every day. But this nano-level of analysis would contain almost exclusively zeros (no crime committed at this address during this hour). Thus this would be, again, completely uninformative — suggesting little about the nature of the jurisdiction’s patterns of crime. Neither examining a jurisdiction’s crime by observing it from 35,000 feet with the naked eye, nor examining its crime at the street level with a microscope are apt to be particularly revealing about the nature of the jurisdiction’s crime problem let alone what particular types of crime would be susceptible to what kinds of crime-control strategies.

Obviously, an informative level of analysis will lie somewhere in between the completely comprehensive macro and the finely differentiated nano. Unfortunately, once again, the appropriate level of disaggregation will rarely be obvious. Someone will have to guess. The experienced pro and the naive neophyte may make different initial guesses about what level to employ. Moreover, both guesses could prove valuable. For it is not preordained or required that one single level of disaggregation will provide all of the insights. Every level of disaggregation could produce a useful understanding of one of the organization’s performance (many?) deficits.

The Luck of “That’s Funny”

In Boston in the summer of 2009, a CompStat analyst happened to notice four different cases of seafood — valued from \$400 for five boxes of shrimp, to \$2,3000 for eight cases of lobster

and 90 lbs of shrimp, up to \$1,150 for five cases of shrimp — that had been stolen from delivery trucks near Fanueil Hall over a two-month period. The analyst's computer had not been programmed to search for stolen seafood. Indeed, no one in the department had been the least bit concerned about seafood larceny. Gang violence, not disappearing shrimp, are the dominant focus of CompStat discussions. Still, when the analyst saw the second case, he must have said to himself "that's funny" and went looking for similar crimes that might suggest a pattern.

Isaac Asimov, the biochemist, popular science writer, and author of numerous books of science fiction, wasn't interested in helping public executives recognize patterns in their world. Rather, he was interested in how scientists produced new discoveries. Asimov's insight into the process of discovery in science is captured in his oft quoted saying: "The most exciting phrase to hear in science, the one that heralds the new discoveries, is not 'Eureka!' (I found it!) but 'That's funny . . .'"³⁷ Asimov didn't mean "that's funny, ha-ha." He meant, "that's funny, strange." And when someone with the right "habits of mind" and "patterns of thought" comes across two different cases of seafood stolen from trucks in the same location, he is apt to say to himself, "that's funny" . . . and then go looking to see whether this is just a coincidence or whether there exists a more significant underlying pattern.

The analyst may be lucky, simply stumbling across some data that looks funny. Yet, this analyst is more than lucky, for the analyst recognizes this luck. In an 1854 lecture at the University of Lille, Louis Pasteur remarked: "dans les champs de l'observation le hasard ne favorise que les esprits préparés."³⁸ Or, in one of the many translations: "In the field of observation, chance only favors the prepared mind."

The search for patterns and for the underlying performance deficits that a pattern may reveal, often begins with the simple but puzzling exclamation: "That's funny." Then, if the mind is prepared, it can go looking for an evaluation of the current level of performance; or for some judgment about what approaches have or have not been working; or for some insights into previously unidentified, underappreciated, or misunderstood performance deficit, or for some assessment of the opportunities for improvement, or for some appraisal of strategies that could be pursued.

"It's an Evidence-Based World Now"³⁹

Today, all professions seek to be "evidence-based." (What's the alternative? To be superstition-based?) And what better evidence for this proposition than the invasion of the "evidence-based" phrase into the popular culture via Gary Trudeau's *Doonesbury* cartoon?

The medical profession can claim to have created today's evidence based world. Indeed, as the field trials for the Salk vaccine illustrate, the medical profession has some built in advantages when it comes to generating evidence about what treatments work and what don't — evidence that will be broadly (though never universally) accepted. The medical profession can design — and implement — the gold standard for comparing the results from a single, specific treatment with the results from a placebo. It can assign — on a purely random basis — some people to get the treatment and some people to get the placebo. It can deny the treatment to those who receive the placebo. It can limit the treatment, isolating it to a single factor. And it prevent those who will get the treatment or the placebo — and everyone with whom they will come in contact — from knowing who got which. Most other professions — and the management profession in particular — can rarely replicate all of these conditions.

Nevertheless, many other professions are adopting the evidence-based language, if not necessarily the identical approach.⁴⁰ We have don't just have evidence-based medicine. We also have evidence-based pediatrics, evidence-based ophthalmology [including a journal with that title], evidence-based dentistry, and evidence-based nursing. We also have evidence-based social

policy,⁴¹ evidence-based social work,⁴² evidence-based child welfare,⁴³ evidence-based education and evidence-based teaching,⁴⁴ evidence-based design in architecture,⁴⁵ We even have “evidence-based librarianship.”⁴⁶

There is also a consulting firm named Evidence-Based Research, Inc.⁴⁷ If you don’t already have your own evidence-based Web site and research institute, it’s probably too late. Catch the next wave.

Evidence-Based Management

But what makes something (anything?) “evidence-based”? In medicine, the widely accepted definition was developed by David Sackett, now at the University of Toront, and his colleagues:

Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients . . . integrating individual clinical expertise with the best available external clinical evidence.

Sackett and his colleagues note that “evidence based medicine is not ‘cookbook’ medicine,” nor is it “restricted to randomised trials and meta-analyses.” Rather, they emphasize that “good doctors use both individual clinical expertise and the best available external evidence, and neither alone is enough.”⁴⁸

The same would apply to evidence-based management. Indeed, Jeffrey Pfeffer and Robert Sutton, two of the most prominent advocates of evidence-based management, argue that evidence-based medicine is “a model for evidence-based management.” Specifically, Pfeffer and Sutton argue:

Evidence-based medicine and evidence-based management require a mind-set with two critical components: first, willingness to put aside belief and conventional wisdom — the dangerous half-truths that many embrace — and instead hear and act on the fact; second, an unrelenting commitment to gather the facts and information necessary to make more informed and intelligent decisions, and to keep pace with new evidence and use the new facts to update practices.⁴⁹

In her presidential address to the Academy of Management, Denise Rousseau of Carnegie Mellon University explored the “promise” of evidence-based management. To Rosseau, “evidence-based practice is not one-size-fits-all; it’s the best current evidence coupled with informed expert judgment” with six features:

- learning about *cause-effect* connections in professional practices;
- isolating the variations that measurably affect desired outcomes;
- creating a culture of evidence-based decision making and research participation;
- using information-sharing communities to reduce overuse, underuse, and misuse of specific practices;
- building decision supports to promote practices the evidence validates, along with techniques and artifacts that make the decision easier to execute or perform (e.g., checklists, protocols, or standing orders); and

- having individual, organizational, and institutional factors promote access to knowledge and its use.

Rousseau argues that evidence-based management offers “a guide to closing the research-practice gap,” which she views as quite “large.”⁵⁰

In the title of her address, Rousseau asked a question: “Is there such a thing as evidence-based management?” Her answer: “no—at least not yet.” After all, as she notes, “managers tend to work in settings that make valid learning difficult.” For example, Rousseau observes that “sadly, there is poor uptake on management practices of known effectiveness (e.g., goal setting and performance feedback).”⁵¹ Yet, at the core of an effective PerformanceStat strategy is precisely this management practice: the setting of targets followed up with feedback on the level of performance.

Indeed, PerformanceStat has the potential (but only the potential) to give a public agency or governmental jurisdiction the opportunity to practice evidence-based management.

Creating the Necessary Analytical Capacity

Unfortunately, few large organizations today are managed by people with the “mind-set” that Pfeffer and Sutton value. They are not populated by the kind of managers whom Rousseau respects, those who have acquired “a systematic understanding of the principles governing organizations and human behavior.”⁵²

Instead, today’s large organizations (public, private, and nonprofit) are built to follow standard operating procedures and organizational routines. If everyone is analyzing data, if everyone is entitled to extract his or her own lessons from these data, and if everyone is authorized to act on the lessons that he or she has derived from the data, then the standardization and uniformity that large organizations seek is compromised. The deviants will try to create their own innovative “skunkworks,” but if the guardians of standardization and uniformity discover these deviants, they will quickly exterminate them.

Large organizations seek to eliminate inequities, to prevent mistakes, and to control confusion by imposing organization-wide consistency in everything from accounting forms, to personnel practices, to operating procedures. In such an organization, central headquarters has also seized the responsibility for producing results — for deciding not just what results to produce but also *how* to produce them. Headquarters creates standard operating procedures for producing results, and as long as a subunit follows these SOPs, it can’t get in trouble because it is not achieving the desired results. After all, this subunit did what headquarters told it to do.

Yet, if headquarters decides that results are important, while accepting that it doesn’t know how best to produce them, and if (as a consequence) it decides to decentralize responsibility for results, then individual units have a duty to analyze their data, to learn from these data, and to act on their learning — perhaps even experimenting with strategies that might be slightly, or even significantly, deviant from the organization’s existing standard operating procedures.

Of course, the task of collecting and analyzing the data can be delegated to a central unit, staffed with a few whiz-kid analysts who can figure out what, within the organization, is working, what isn’t, and what should be done in the future. But, again, with this arrangement, who is responsible for performance? The decentralized line units? Or the centralized analytical staff?

The good news — from the perspective, that is, of the guardians of organization-wide consistency — is that few middle managers are trained or able to analyze their own data. Young people do not choose to enter the policing profession because they want to analyze data to determine what crime-fighting strategies are most effective for which crimes. They do not enter

the teaching profession because they want to analyze data to determine what teaching strategies are most effective for which students. Most people tend to choose a profession to do it, not to analyze it.

Moreover, it is their doing of it — if admittedly by following the standardized procedures — that gives them personal satisfaction plus organizational promotion. Who becomes the fire chief? The best fire fighter. People tend to be promoted for their demonstrated professional abilities, not necessarily for their skills at management or leadership, let alone analysis.

Consequently, when the leadership team of a public agency or governmental jurisdiction decides to create a PerformanceStat strategy, it discovers that most of its middle managers are good at following procedures and getting others to do so too. Often, however, they lack both a performance perspective and analytical skills. They don't have Sparrow's "habits of mind" and "patterns of thought." They have not been expected to engage in data analysis let alone to exercise performance leadership. And they have given little training in either.

This is a core problem for those seeking to employ a PerformanceStat strategy: How do they identify within their (rather bureaucratic) organization people who have the leadership and analytical skills necessary to make the strategy work?

Of course, it would be possible to augment the leadership team of each unit that has the decentralized responsibility for producing results by permitting it hire its own analyst. But unless the unit's analyst is able to explain the operational implications of the data in non-technical terms, and unless the leadership team is able to understand, appreciate, and accept both the analysis and its implications, and unless the leadership team is also willing and able to act strategically on these implications, an analyst would be of little use to the unit.

In performance-focused government, basic analytical skills — a fluency and comfort with numbers and the ability to extract information from data — is becoming as important as sophisticated interpersonal skills. (Today, even baseball teams need managers with the analytical ability to think about whether their abundance of performance data have some real and significant implications for their team's strategy.) At the moment, however, people who possess both sophisticated interpersonal and analytical skills are rare.

This paper is the working draft of chapter five of a forthcoming book tentatively titled: *The PerformanceStat Potential: A Leadership Strategy for Producing Results*. Consequently, the author would appreciate all comments and criticisms.

Notes

1. Quoted in, Amalie Benjamin, "Major move: Matsuzaka will start Tuesday," *The Boston Globe*, September 12, 2009, p. C3.

2. Over the past several decades, much of government's attention to performance has consisted primarily of auditing and evaluation — efforts by independent units (usually within government but outside the unit being audited or evaluated) to determine whether some other agency of government is performing well, adequately, or poorly. Michael Power of the London School of Economics and Political Science has called this "the audit explosion" and critiqued "the audit society."

Michael Power, *The Audit Explosion* (London: Demos, 1994)

Michael Power, *The Audit Society: Rituals of Verification* (Oxford, U.K.: Oxford University Press, 1999)

Michael Power, "The Audit Society — Second Thoughts," *International Journal of Auditing*, vol. 4, no. 1 (March 2000), pp. 111-119.

In contrast (perhaps based on the assumption that a little KITA [see Herzberg] from these outsiders will motivate agencies to improve) much little effort has been put into actually helping the leaders of public agencies to improve their performance.

3. There is, of course, nothing particularly "uniform" about the FBI's Uniform Crime Reports. The FBI has definitions for each crime, and provides guidelines for classifications. But the decisions about how to classify a crime are made by nearly 17,000 local police departments — often by the officer who goes to the scene and submits a report. And there little reason to believe that all police officers across the country are collectively — or, indeed, individually — consistent in how they classify identical or merely similar crimes.

4. Jack Maple (with Chris Mithcell), *The Crime Fighter* (New York: Doubleday, 1999), p. 30.

5. Jack Maple (with Chris Mithcell), *The Crime Fighter* (New York: Doubleday, 1999), p. 30.

6. The literature on the unintended and perverse consequences is long and critical. One such example with a survey of this literature is:

Sandra van Thiel and Frans L. Leeuw, "The Performance Paradox in the Public Sector," *Public Performance & Management Review*, vol. 25, no. 3 (March 2002), pp. 267-281.

7. Marshall W. Meyer and Vipin Gupta, "The Performance Paradox," in *Research in Organizational Behavior*, vol. 16, Barry M. Staw and L. L. Cummings (eds.) (Greenwich, Conn.: JAI Press, 1994), pp. 310 and 322.

Is this lack of correlation among indicators of organizational performance mirrored by a similar lack of correlation among indicators of individual performance? Seashore, Indik, and Georgopolous report a lack of correlation, while Viswesvaran, Schmidt, and Ones do find what they call "a general factor" in the rating of individual job performance.

Stanley E. Seashore, Bernard P. Indik, and Basil S. Georgopolous, "Relationships among criteria of job performance," *Journal of Applied Psychology*, vol. 44, no. 3 (June 1960), pp. 195-202.

Chockalingam Viswesvaran, Frank L. Schmidt, and Deniz S. Ones, "Is There a General Factor in Ratings of Job Performance? A Meta-Analytic Framework for Disentangling Substantive and Error Influences," *Journal of Applied Psychology*, vol. 9, no. 1 (January 2005), pp. 108-131.

8. Byron Sharp, Erica Riebe, John Dawes and Nick Danenberg, "A Marketing Economy of Scale--Big Brands Lose Less of their Customer Base than Small Brands," *Marketing Bulletin*, vol. 13 (May 2002), pp. 1-8. See also, Ajay K.Kohli, N. Venkatraman, and John H. Grant, "Exploring the Relationship Between Market Share and Business Profitability," in *Research in Marketing*, vol. 10, Jagdish N. Sheth (ed.) (Greenwich, Conn.: JAI Press, 1990), pp. 113-33.

For a meta-analysis 276 studies of the relationship between market share and profitability, see David M. Szymanski, Sundar G. Bharadwaj, and P. Rajan Varadarajan, "An Analysis of the Market Share-Profitability Relationship," *The Journal of Marketing*, vol. 57, no. 3 (July, 1993), pp. 1-18.

9. Denise M. Rousseau, "Organizational Behavior in the New Organizational Era," *Annual Review of Psychology*, vol. 48 (1997), p. 525.

10. Stuart E. Jackson, *Where Value Hides: A New Way to Uncover Profitable Growth for Your Business* (New York: Wiley, 2006), pp. 15 & 16.

11. "On a typical weekday, the Department logs more than 3,000 miles transporting inmates to courts in the five boroughs and to medical and other jail or prison facilities throughout the city and state." http://www.nyc.gov/html/doc/html/about/facilities_overview.shtml (September 13, 2009)

12. http://www.nyc.gov/html/doc/html/stats/doc_stats.shtml (September 13, 2009)

13. Robert D. Behn, "Linking Measurement and Motivation: A Challenge for Education," in *Advances in Educational Administration: Improving Educational Performance: Local and System Reforms*, vol. 5, Paul W. Thurston and James G. Ward (eds.) (Greenwich, Conn.: JAI Press, 1997), pp. 15-58.

14. DPSS explains that "DPSSTATS, implemented in April 2005, is an innovative management tool that creates a forum for Executive Management and key managers across the Department to review current and comprehensive data so that swift, effective decisions could be made to enhance the Department's operations to better serve the more than 2 million residents of Los Angeles County that are served by this Department."

<http://dpss.lacounty.gov/dpss/dpsstats/default.cfm> (September 14, 2009)

15. http://www.nyc.gov/html/hra/html/about/about_hra_dss.shtml (September 16, 2009)

16. http://dpss.lacounty.gov/dpss/about_dpss/dpss_mission.cfm (September 16, 2009)

17. Some of the measures are less obvious to outsiders. For example, to ensure that HRA can check to see if it needs to adjust the benefits of a client who has been placed in a job, HRA requires the vendor that places a client in a job report that placement (using form FIA-3A) within one week of when the participant started working. For 2009, each of HRA's 28 Job Centers had a 90 percent completion-rate target, with 80 percent being the minimally acceptable level.

18. For example, see the July 2009 JobStat Report for the Dekalb Job Center in Brooklyn at:

<http://www.nyc.gov/html/hra/downloads/pdf/dekalb.pdf> (September 17, 2009)

19. The latest such report is: Bureau of Contract and Technical Services, *Caseload Characteristics*, (Los Angeles County Department of Public Social Services, Los Angeles, June 2009). It can be found at: http://www.ladpss.org/dpss/ISS/pdf/2009/caseload_char_june_2009_reduced.pdf

20. The latest such report is: Bureau of Contract and Technical Services, *Statistical Report*, (Los Angeles County Department of Public Social Services, Los Angeles, July 2009). It can be found at: http://www.ladpss.org/dpss/ISS/pdf/2009/stat_0709_reduced_combined_link.pdf.

Access to the latest and past Caseload Characteristics Reports, and Statistical Reports can be found at: http://www.ladpss.org/dpss/ISS/ISS_Section.cfm.

21. DPSSTATS "Fact Sheet" dated July 23, 2008. Available at

http://dpss.lacounty.gov/dpss/dpsstats/dpsstats_process_fact_sheet/dpsstats_process_fact_sheet_07_08.pdf (Accessed September 14, 2009)

The full description of "performance measures" on this fact sheet is:

DPSSTATS maintains an inventory of more than 160 different performance measures." That's a lot of data to collect, let alone to analyze and digest. The measures come from a variety of sources including, Performance Counts!, Department Head Goals, managers' MAPP goals, State or federal standards, or current Departmental priorities. New measures are constantly being developed and explored through the STATS process. Divisions may recommend measures for themselves or for other operations in the Department. The measures chosen for a particular meeting are determined by the Chief Deputy and the Assistant Directors.

While the measures reviewed at STATS often change, a set of core measures for Line offices (Customer Service, all Application Processing, Food Stamp Error Rate, and Medi-Cal Redetermination Processing) are reviewed prior to each Line meeting and presented in a

separate document at the start of the meeting. District Directors that meet all of the applicable core measures are recognized with stars on their nameplates for each meeting where they have achieved the established targets.

22. Earning a “star” might not appear to be a significant reward. Nevertheless, it establishes a performance hierarchy among the local offices, and thus (to many) is both prized and pursued.

23. William Atkinson, Jennifer Hamborsky, Lynne McIntyre, Charles Wolfe (eds.), *Epidemiology and Prevention of Vaccine-Preventable Diseases*, 10th edition (Atlanta, Ga: The Centers for Disease Control and Prevention, 2008), p. 139.

24. T. Jacob John and Reuben Samuel, “Herd immunity and herd effect: new insights and definitions,” *European Journal of Epidemiology*, vol. 16, no. 7 (June 2000), pp. 601-606.

25. “The Salk Poliomyelitis Vaccine,” *American Journal of Public Health*, vol. 45, no. 5 (May 1955), p. 676.

26. Marcia Meldrum, “‘A calculated risk’: the Salk polio vaccine field trials of 1954,” *MJB*, vol. 317, no. 7167 (October 31, 1998), pp. 1233-1236.

27. Thomas Francis, Jr., Robert F. Korn, Robert B. Voight, Morton Boisen, Fay Hemphill, John A. Napier, and Eva Tolchinsky, “An Evaluation of the 1954 Poliomyelitis Vaccine Trials,” *American Journal of Public Health*, vol. 45, no. 5, (May 1955, pt. 2). See also: Paul Meier. “The Biggest Public Health Experiment Ever: The 1954 Field Trial of the Salk Poliomyelitis Vaccine” in *Statistics, A Guide to the Unknown*, Judith M. Tanur, Frederick Mosteller, William H. Kruskal, Richard F. Link, Richard S. Pieters, Gerald R. Rising (eds.) (San Francisco: Holden-Day, 1975), pp. 2-13.

28. For a description of how this randomized was done, see Francis, et al., “An evaluation of the 1954 Poliomyelitis Vaccine Trials,” pp. 4-5.

29. Steven Johnson, *The Ghost Map: The Story of London’s Most Terrifying Epidemic and How It Changed Science, Cities, and the Modern World* (New York : Riverhead Books, 2006).

For a dissent from the significance of Snow’s role, see:

Howard Brody, Michael Russell Rip, Peter Vinten-Johansen, Nigel Paneth, Stephen Rachman, “Map-making and myth-making in Broad Street: the London cholera epidemic, 1854,” *Lancet*, vol. 356, no. 9223 (July 1, 2000), pp. 64-68.

Actually, the mapping of the incidence of disease occurred at least half a century earlier. See:

Lloyd G. Stevenson, “Putting Disease on the Map: The Early Use of Spot Maps in the Study of Yellow Fever,” *Journal of the History of Medicine*, vol. 20, no. 3 (July 1965), pp. 226-261.

Frank A. Barrett, “Finke’s 1792 map of human diseases: the first world disease map?” *Social Science & Medicine*, vol. 50, nos. 7-8 (April 2000), pp. 915-921.

30. William Bratton (with Peter Knobler), *Turnaround: How America’s Top Cop Reversed the Crime Epidemic* (New York: Random House: 1998), p. 99.

31. Jack Maple with Chris Mitchell, *The Crime Fighter: Putting the Bad Guys Out of Business* (New York: Doubleday, 1999), p. 128.

32. Today, even citizens have access to on-line crime maps of their municipality. The National Institute of Justice provides links to the crime maps posted by numerous police departments: <http://www.ojp.usdoj.gov/nij/maps/links.htm>. Several organizations outside of government — CrimeReports.com (at <http://crimereports.com>), and The Omega Group (at <http://www.crimemapping.com>), and SpotCrime (at <http://www.spotcrime.com>) — have established relationships with local police departments to the municipality’s crimes on the Web. I can even get a map of the crime at my university by going to <http://ucrime.com/ma/harvard+university>.

33. Even when using the wonderful, first-order analytic tool of the map, the analyst needs to think carefully about the appropriate level of disaggregation. Mark Monmonier of Syracuse University observes that for a choropleth map — one that uses different colors, shades, or patterns to indicate the level of some variable (such as the crime rate) for different subareas — the level of disaggregation can be important. Of course, if the map simply consists of one dot for each incident — a crime or a case of cholera — the disaggregation is complete. But if the map maker decides to aggregate the data into a choropleth map (because, for example, he or she believes that there are simply too many individual dots to provide a useful appreciation of the phenomena) the choice of boundaries of the subareas can be important.

Monmonier takes Snow's map of the cholera dots and aggregates them using three different ways of drawing the boundaries using the streets near the Broad Street pump. Different aggregations produced different impressions, but all three "diluted the Broad Street cluster." Monmonier notes that "aggregation involves not only areal units but also time, disease classification, and demography" and how those choices are made affects what the aggregation reveals or conceals. "Clearly one map is not sufficient," emphasizes Monmonier, "although one good map can signal the need for a more detailed investigation." Indeed, that is the objective of any first-cut analysis: to find something that suggests "the need for a more detailed investigation."

Mark Monmonier, *How to Lie with Maps* (Chicago: University of Chicago Press, 1991), pp. 22-23, 141-143.

34. One of the worst things that could happen to the PerformanceStat strategy would be for it to be kidnaped by the ISO crowd. For the International Organization for Standardization gets hold of this innovative leadership strategy, they will standardize the hell out of it. There will be nothing left but the PerformanceStat template.

35. Malcolm Sparrow, *The Character of Harms: Operational Challenges in Control* (Cambridge, U.K.: Cambridge University Press, 2008), pp. 8, 11, 9.

36. Malcolm Sparrow, *The Character of Harms: Operational Challenges in Control* (Cambridge, U.K.: Cambridge University Press, 2008), pp. 14, 9.

37. Read Montague, *Why Choose This Book? How We Make Decisions* (New York: Dutton, 2006), p. 108

38. Louis Pasteur, *La Vie De Pasteur*, Rene Vallery-Radot, ed. (Paris: Librairie Hachette, 1900), p. 88.

39. In the January 30, 2009 edition of the comic strip *Doonesbury*, Gary B. Trudeau has Joan tell Clyde "It's an evidence-based world now."

http://www.doonesbury.com/strip/dailydose/index.html?uc_full_date=20090130

40. For links to the evidence-based movement in other professions, see:

<http://www.evidence-basedmanagement.com/movements/> (September 16, 2009)

41. Sandra M. Nutley, H. T. O. Davies, Peter C. Smith (eds.), *What Works?: Evidence-Based Policy and Practice in Public Services* (Bristol, U.K.: The Policy Press, 2000). And the "Coalition for Evidence-Based Policy" is a nonprofit that focuses on "social programs that work": <http://evidencebasedprograms.org/wordpress/>

42. The last several years have produced numerous books on "evidence-based social work": Albert R. Roberts and Kenneth R. Yeager, *Foundations of Evidence Based Social Work Practice* (New York: Oxford University Press, 2006). Peter Sommerfeld and Prisca Herzog (eds.), *Evidence-based Social Work: Towards a New Professionalism?* (New York: Peter Lang, 2005). Tony Newman, Alice Moseley, Stephanie Tierney, *Evidence-Based Social Work: A Guide for the Perplexed* (Dorset, U.K.: Russell House, 2005). Mel Gray, Debbie Plath, and Stephen A. Webb, *Evidence-based Social Work: A Critical Stance* (London: Routledge, 2009).

In addition, there is a journal titled *Evidence Based Social Work*.

43. At least one book covers evidence-based child welfare: Duncan Lindsey and Aron Shlonsky (eds.), *Child Welfare Research Advances for Practice and Policy* (New York: Oxford University Press, 2008), "Part II. Evidence-Based Practice in Child Welfare." In addition, the California Evidence-Based Clearinghouse for Child Welfare has a Web site, <http://www.cachildwelfareclearinghouse.org>. And the National Association of Public Child Welfare Administrators has published a "Guide for Child Welfare Administrators on Evidence Based Practice," <http://www.chadwickcenter.org/Documents/Guide-for-Evidence-Based-Practice.pdf>. And the U.S. Government's Child Welfare Information Gateway is a Web site that provides information "About Evidence-Based Practice." http://www.childwelfare.gov/systemwide/service/improving_practices/about.cfm

44. Books on evidence-based teaching and/or education include: Geoff Petty, *Evidence Based Teaching: A Practical Approach* (Cheltenham, Gloucestershire, U.K.: Nelson Thornes 2006); David Mitchell, *What Really Works in Special and Inclusive Education: Using evidence-based teaching strategies* (London: Taylor & Francis, 2007); Carol A. Angell (Eric Paul Hartwig ed.), *Evidence-Based Education: Examining Today's Research for Tomorrow's Decisions* (Horsham, Penna.: LRP Publications, 2006); Richard Pring and Thomas, *Evidence-Based Practice in Education* (Maidenhead, Berkshire, U.K.: Open University Press, 2004); Kathleen R. Stevens, and Virginia R. Cassidy (eds.), *Evidence-Based Teaching: Current Research in Nursing Education* (Sudbury, Mass.: Jones & Bartlett Publishers, 1999).

Not that either evidence-based teaching or education is noncontroversial. See, for example, Scott Webster, "How evidence-based teaching practices are challenged by a Deweyan approach to education," *Asia-Pacific Journal of Teacher Education*, vol. 37, no. 2 (May 2009), pp. 215-227. And then, what is it: The Council for Exceptional Children reports: "While the law requires teachers to use evidence-based practices in their classrooms, the special education field has not yet determined criteria for evidence based practice nor whether special education has a solid foundation of evidence-based practices." http://www.cec.sped.org/AM/Template.cfm?Section=Evidence_based_Practice&Template=/TaggedPage/TaggedPageDisplay.cfm&TPLID=24&ContentID=4710

45. D. Kirk Hamilton and David H. Watkins. *Evidence-Based Design for Multiple Building Types* (Hoboken, N.J.: Wiley, 2008); Linda Nussbaumer, *Evidence-Based Design for Interior Designers* (New York: Fairchild Publications, 2009).

46. Jonathan D. Eldredge, "Evidence-based librarianship: an overview," *Bulletin of the Medical Library Association*, vol. 88, no. 4 (October 2000), pp. 289-302. See also, Andrew Booth, "Counting what counts: performance measurements and evidence-based practice," *Performance Measurement and Metrics*, vol. 7, no. 2 (2006), pp. 63-74.

47. <http://www.ebrinc.com/> (September 15, 2009)

48. David L. Sackett, William M. C. Rosenberg, J. A. Muir Gray, R. Brian Haynes, and W. Scott Richardson, "Evidence based medicine: what it is and what it isn't". *BMJ*, vol. 312, no. 7023 (January 13, 1996), pp. 71-72.

49. Jeffrey Pfeffer and Robert I. Sutton, *Hard Facts, Dangerous Half-Truths and Total Nonsense: Profiting from Evidence-Based Management* (Boston, Mass.: Harvard Business School Press, 2006), pp 13, 14. Pfeffer and Sutton also have a Web site on evidence-based management:

<http://www.evidence-basedmanagement.com/>

50. Denise M. Rousseau, "Is There Such a Thing as 'Evidence-Based Management?'" *Academy of Management Review*, vol. 31, no. 2 (April, 2006), pp. 256, 267, 259-260, 260, & 261.

51. Denise M. Rousseau, "Is There Such a Thing as 'Evidence-Based Management?'" *Academy of Management Review*, vol. 31, no. 2 (April, 2006), pp. 256, 258, 261, & 258.

52. Denise M. Rousseau, "Is There Such a Thing as 'Evidence-Based Management?'" *Academy of Management Review*, vol. 31, no. 2 (April, 2006), p. 261