Measuring the Performance of Law Enforcement Agencies

Edward R. Maguire, PhD, Associate Professor, Administration of Justice Program, George Mason University, Fairfax, VA

Part 1 of a two-part article appearing in the *CALEA Update*

Introduction

This is the first segment of a two-part article on measuring the performance of law enforcement agencies. It is written for a policing audience and draws in part on my discussions with members of CALEA's Performance Measurement Subcommittee and those who have attended my training workshops at the last two CALEA Conferences. In this first segment, I introduce the general concepts, terminology, and history of comparative performance measurement in policing. The second segment will show you how to develop, pilot-test, and implement comparative performance measurement in your agency. This article is one small part of a larger effort by CALEA to explore the feasibility and utility of agency-level performance measurement in policing. That journey is just beginning, and will proceed slowly, but it is expected to be a worthwhile one.

From corporate boardrooms to elementary school classrooms, performance measurement is everywhere. Our children are required to take standardized tests designed to ensure that school performance is up to snuff.[1] When a patient dies, physicians must now wonder whether the event will inflate the hospital's "risk adjusted mortality rates"[2] beyond established performance benchmarks. Even the Internal Revenue Service has not escaped the movement toward performance measurement.[3] With performance measurement appearing in such a diverse array of organizational settings, it is not surprising that it is now becoming a hot topic in policing.

Performance measurement is at the heart of nearly every innovative management fad or organizational development strategy in the past two decades. It is an essential component of zero-based budgeting and management by objectives,[4] reinventing government,[5] re-engineering the corporation,[6] total quality management,[7] benchmarking,[8] balanced scorecards,[9] and organizational learning.[10] Despite its popularity, *performance measurement* is an inherently ambiguous term. It is used in various ways to refer to the performance of individuals, products and services, subunits, projects, and organizations. Yet the methods and data used to measure performance at these different levels can vary significantly. This report discusses some options for measuring the performance of police organizations. Moreover, it focuses on *comparative* performance measures: those that can be used to compare units over space or time. Comparative performance measures can be used to compare the performance of two or more organizations, or they can be used to compare a single organization's performance at multiple points in time.

This article provides a brief review of comparative performance measurement in policing. It is written with practical application in mind, alerting readers to

the many issues that arise in performance measurement, and suggesting some concrete steps that CALEA and its members can follow if they choose to implement a performance measurement system. The next section provides a brief history of police performance measurement followed by a description of what I have called a "Golden Thread," a theme that is woven throughout a story, linking together its disparate parts. In this case, that theme is very simple, yet very powerful: police performance is multidimensional. This idea, as simplistic as it might seem, is the foundation of effective performance measurement. The last section reviews some of the dimensions of police performance that have been examined in the past, offering some practical suggestions for those who are thinking about generating their own lists. The next segment of this article will feature a number of additional sections that explore how to implement performance measurement, both nationally, and within your agency.

A Brief History of Police Performance Measurement

In this section, I provide a brief overview of comparative performance measurement in policing. The review is brief in spite of a large and growing body of academic and professional literature on the topic. I begin by discussing the role of performance measurement in the early part of the 20th century, with particular focus on the 1930s. I then skip ahead to the 1960s, 1970s, and beyond, assessing the level of progress that has been made in the development and implementation of comparative performance measurement. I finish by briefly discussing the influence of the community policing movement on police performance measurement.

Police organizations have been collecting data about their performance since the birth of modern policing in the mid-19th century.[11] Most of these efforts were primarily local and intended to demonstrate the inputs, activities, and outputs of individual police agencies. The idea of comparative performance measurement began to take root in the early 20th century, shortly after the birth of the International Association of Chiefs of Police (IACP) in 1894. In 1927, the IACP created a Committee on Uniform Crime Records to develop a standardized system for collecting crime data from police agencies throughout the nation. The Committee created the architecture for the Uniform Crime Reports (UCR), and in 1930, Congress authorized the Attorney General to begin collecting UCR data, a task he assigned to the Federal Bureau of Investigation. During its first year, the UCR program collected data from 400 police agencies in 43 states. By 1998, it was routinely collecting data from more than 17,000 police departments in all 50 states.[12] The UCR has now become the primary foundation for comparative performance measurement of police agencies in the United States.

The 1930s saw several significant milestones in the history of police performance measurement. In 1930, Donald Stone, Director of Research for the International City Managers' Association, proposed two measures of police effectiveness: "the number of cases cleared and the value of stolen property recovered."[13] Both proposed measures were later criticized; though, in practice they continue to be used by both police and academics. In 1935, Arthur Bellman, a protégé of August Vollmer, created an extensive instrument designed to measure the quality of police service.[14] Containing 685 specific items, the instrument was designed to be completed by expert police analysts asked to render a professional judgment on each item. With its vast array of standards, Bellman's scale looked curiously like an accreditation

checklist. Bellman's approach to police performance measurement was criticized on three primary grounds. First, it was based on "conformity to current notions of good administrative practice" and, therefore, was poorly equipped to accommodate innovations and improvements in policing.[15] Second, echoing a theme to which we will return at the end of this article when we discuss "weighting," Bellman's rating system treated each of the indicators equally. According to critics, the additive nature of Bellman's system "resulted in mixing significant and petty issues indiscriminately."[16] Finally, Bellman's approach focused exclusively on internal measures relating to policies, practices, and equipment. It completely neglected the processes, outputs, and outcomes of police agencies.[17]

In 1938, responding to problems with Bellman's system, Spencer Parrat proposed an alternative performance measurement system involving the use of citizen surveys to measure public confidence in the police. Parrat's recommendation has been adopted in many jurisdictions throughout the nation; however, there is little research to demonstrate how much time elapsed before the idea took root. Citizen surveys were a crucial component of the research done by the President's Commission on Law Enforcement and the Administration of Justice in the late 1960s in response to the disorder and civil unrest of that rebellious period.[18] The 1970s saw the blossoming of citizen surveys as a standard research tool for police researchers. By the late 1990s, nearly one-third of police agencies reported having conducted citizen surveys within the past year.[19] Nonetheless, the proliferation of citizen surveys has done little to move the policing field closer to the use of comparative performance measures since such surveys are usually designed and administered locally. The Bureau of Justice Statistics and the Community Oriented Policing Services Office recently completed a study of victimization experiences and satisfaction with the police among citizens in 12 cities. The results demonstrated important intercity variation in citizen experiences and perspectives—valuable information for police managers in these cities.[20]

Starting in 1939, the International City Managers' Association (now called the International City/County Management Association or ICMA) began collecting data from police organizations as part of its Municipal Yearbook series.[21] The Municipal Yearbooks include data on a variety of city government features, with police data only one small part of a much larger data collection effort that inquires about form of government, salaries of local officials, personnel practices, technology, economic development, and other related topics.[22] It is unknown to what extent this data collection series was used as a platform for comparing the performance of police organizations in the 1930s. Anecdotal evidence suggests that it continues to be used today in spite of two limitations: (1) it is one of the only databases on police performance that must be purchased; nearly all others are available for free in various archives and (2) a recent review of surveys of police organizations found that response rates in the ICMA surveys were among the lowest of all the surveys examined in the review.[23] Low response rates led social scientists to wonder whether a sample is biased—whether those agencies represented in the ICMA databases are representative of all police agencies, especially those that refused or otherwise failed to complete the ICMA surveys.

In summary, the 1930s saw a mix of ambitious activities and proposals for measuring the performance of police agencies. A national system, the Uniform Crime Reports, was developed to collect "official" statistics on crime and arrests.

This was followed almost immediately by proposals about how the data ought to be used for comparative performance measures. The ICMA instituted its Municipal Yearbook series containing data that continues to be collected today. Bellman created an exhaustive list of performance standards containing mostly internal features and inputs. Parrat criticized Bellman's approach, recommending instead subjective indicators of public confidence and satisfaction derived from surveys of citizens. As I will show throughout this article, although many people now recognize the need for alternative performance measures, many of the issues that warranted discussion and debate in the 1930s are still with us today.

Throughout the next three decades, "traditional" measures of police agency performance became entrenched within the policing profession with little debate or fanfare. Crime rates, arrests and citations, clearances, and response times all played a key role in measuring police performance at multiple levels, from the individual police officer to the organization as a whole. According to Geoff Alpert and Mark Moore, these "generally accepted accounting practices became enshrined as the key measures to evaluate police performance."[24]

During the 1960s, several themes converged to cast light upon these traditional performance measures. Passionate discontent about the military action in Vietnam, the civil rights movement, and other social forces led a generation of youth to rebel against the conventions of mainstream society.[25] Since police are the gatekeepers of mainstream society, much of the civil unrest of this period brought the police face-to-face with citizens expressing various forms of protest, from peaceful civil disobedience to violent rebellion and rioting.[26] Police use of force and mistreatment of minority citizens became a prominent theme during the 1960s. Research conducted during that period showed that many police officers held racist attitudes toward minorities.[27] Several of the riots that engulfed American cities occurred in the aftermath of police actions such as shootings, traffic stops, or raids occurring in minority neighborhoods.[28] The National Advisory Commission on Civil Disorders (1968) found that "deep hostility between police and ghetto communities" was a primary determinant of the urban riots that it studied. The U.S. Supreme Court, under Chief Justice Earl Warren, began to scrutinize closely the activities of the police. In several "landmark" cases, the Court restricted the powers of the police to conduct searches, obtain confessions, or prevent detainees from consulting with an attorney. While civil libertarians praised this "due process revolution," others complained loudly that these new rules interfered with the ability of the police to fight crime.[29] All of these factors combined to produce an epidemic crisis of legitimacy for the American police. From 1968 to 1971, three national commissions recommended sweeping reforms intended to improve the relationships between police and communities, reduce the levels of racism, limit the use of force, and encourage lawful behavior by the police. All of these themes pointed rather forcefully to the need for alternative measures of police performance.

With these themes in mind, many critics pointed out that police departments that excel at controlling crime; generating arrests, citations, and clearances; and responding quickly to calls-for-service might still perform poorly in many other ways. They might have low morale, poor relationships with communities, problems with corruption or brutality, or an undeveloped capacity to deal with large-scale civil disturbances. Furthermore, numerous observers began to note that a substantial proportion of police work is unrelated to crime:

> No longer can we group police noncriminally related public services into a "miscellaneous" category which composes 70% of recorded police activities, but must refine our measurement of this group of activities and develop performance measurements and criteria relating to the adequacy and quantity of these services . . .[30]

Therefore, a comprehensive suite of performance measures needs to account for a broader spectrum of the work that police do, not just that part of their work related to issuing citations and arresting offenders. If police are supposed to prevent crime and motor vehicle accidents, solve community problems, reduce disorder, and build lasting community relationships, then performance measures should reflect their success in producing these and other valuable outcomes.[31]

Research in the 1960s and 1970s revealed not only that police performance measures needed to be broader and more inclusive; it also pointed out severe flaws in existing traditional measures. Below I highlight some of the criticisms that have been leveled at four traditional measures of police performance: (1) crime rates, (2) arrests and citations, (3) clearances, and (4) response time.

Crime Rates

Most policing scholars argue that there is no single "bottom line" in policing.[32] Like other public agencies, police departments have multiple, perhaps even competing goals; therefore, to focus exclusively on one goal at the expense of the others is to invite poor performance on alternative goals. William Bratton, Chief of the Los Angeles Police Department and former Commissioner of the New York City Police Department, disagrees vehemently with this notion. Under his administration of the NYPD, "crime statistics [became] the Department's bottom line, the best indicator of how police are doing, precinct by precinct and citywide."[33] Elsewhere, he wrote that "crime reduction is to a police department what profit is to a private company—the bottom line."[34] Critics were quick to illuminate the dangers inherent in this perspective. For instance, as Mark Moore, Professor at Harvard's Kennedy School of Government, points out, "We have learned through the efforts of pioneering police chiefs that there are ways of operating police departments that reduce crime and enhance security without harming civil liberties or community satisfaction."[35] Criminologist George Kelling has argued that "measuring police performance solely by crime statistics simply ignores consequential values . . . [such as] justice, integrity, fear reduction, citizen satisfaction, protection and help for those who cannot protect or help themselves, and many others."[36]

Regardless of one's perspective on the relative importance of crime rates as measures of police success, there are two primary problems with using "unadjusted" crime rates as performance measures for police.[37] First, police are not the only factor that influences crime rates. Crime is the product of a complex array of social, economic, and political forces. Research demonstrates clearly that police departments can have a substantial impact on some types of crime.[38] Some crimes, like open-air drug markets, are more visible, preventable, and suppressible than others. Other crimes may be more difficult for police to reduce. Sometimes crime is reduced through the efforts of police, while in other instances, it is reduced through factors having nothing to do with the police. Similarly, when a variety of social factors coalesce to increase crime, it is inappropriate to blame the police for factors beyond their

control. Police can have an effect on crime, but so can many other factors. Before using crime rates as measures of police performance, it is necessary to "adjust" them statistically to account for these other factors. Later in this report, I will describe some methods for performing these adjustments.

Second, reported crime rates often have as much to do with how local police departments process the information they receive as they do with the "true" level of crime. Crime rates derived from police data have been referred to as "organizational outcomes." In other words, they are as much a product of the police department that produced them as they are of the community or situation in which the alleged offenses took place.[39] Police departments can influence crime rates in any number of ways: departmental policies or structures that inhibit or encourage reporting, the behavior of a call-taker or police officer toward an alleged victim, or outright manipulation of crime statistics.[40]

Finally, not all crime is reported to the police; therefore, it makes sense to supplement "official" crime data with victimization surveys that indicate the extent of unreported crime. Through such surveys, police agencies might be able to identify high-risk populations that do not routinely ask for police assistance when needed. For instance, when immigrants fear deportation, domestic abuse victims fear retaliation from their attackers, or teens fear that reporting a crime will hurt their reputation among their peers, unknown pockets of crime will exist. Knowing they exist will enable police to design potentially effective responses. Some research has combined "official" crime data with victimization data to compute composite performance measures.[41]

Arrests and Citations

Arrest represents one of the most visible measures of police output. Furthermore, it is one of the few output measures collected from most police agencies in the country. The Federal Bureau of Investigation has been collecting arrest data from American police departments since 1930 as part of its Uniform Crime Reports program.[42] Data is available for 29 general offense categories. Although this measure might appear on its face to be clear, research has shown that the legal definition of arrest varies widely across agencies.[43] One study concluded that "differing arrest definitions make productivity comparisons between agencies impossible."[44] Research findings suggested that state Uniform Crime Reporting agencies failed to ensure quality control of arrest data and that, in some cases, they failed to understand the rules themselves. The study concluded that "the regulation of arrest statistics is inadequate and that UCR arrest statistics cannot be used to evaluate police performance."[45]

In addition to these measurement problems, arrests are also conceptually ambiguous. As George Kelling has suggested, arrests are not effective measures of police performance because sometimes they represent a failure by police to adopt other, more useful solutions.[46] Herman Goldstein, the architect of problem-oriented policing, views arrests as just one tool in the police officer's toolkit.[47] It is often the right tool to use, but sometimes other solutions may be more effective. Simply counting arrests, therefore, produces a figure of unknown value.

Unlike arrests, there is no national data on citations issued by police agencies. Police departments traditionally maintain their own records on citations and have historically paid close attention to citation productivity. Citations are one of the

basic outputs of police agencies, used much more frequently than arrests. Of the estimated 19.3 million drivers who were pulled over by police at least one time in 1999, about 54% received a traffic citation, about 26% received a warning, and only about 3% were arrested.[48] Research has shown that there is substantial interagency variation in traffic citations for moving violations.[49] Traffic tickets are not the only kind of citation used by police agencies. Many jurisdictions now rely on citations in lieu of arrest for certain misdemeanors. For instance, many states authorize the use of citations for possession of small amounts of marijuana.[50] The use of field citations played a role in the well-publicized changes instituted in the New York City Police Department in the 1990s. William Bratton, former Commissioner of the New York City Police Department, derided the use of "Desk Appearance Tickets" (DATs), a form of field citation used in lieu of arrest, in which people accused of minor offenses were given a court date and released. Bratton directed his agency to curtail its use of DATs in favor of making more arrests: "No more DATs. If you peed in the street, you were going to jail."[51] Implicit in Bratton's statement is a judgment that citations are a less effective pretrial strategy than arrest. Little is known about whether this assumption is valid. We do know, however, that the number of arrests under Bratton's tenure rose dramatically, suggesting that he was able to mobilize his officers to reduce their use of citations in favor of arrests. Citation data may be useful for individual police organizations to keep track of how officers are spending their time or, as in the example with regard to former Commissioner Bratton, to ensure that the organization is producing outputs in the manner prescribed by the chief executive. This data is not available nationally; therefore, it cannot be used to compare police departments nationally.

Furthermore, arrests and citations are "output" measures. They demonstrate the extent to which organizations engage in certain activities, but they say nothing about whether these activities were effective in producing something of value for communities. In other words, they are not "outcomes." When police departments cite the number of arrests they make or citations they issue, it is the equivalent of a carpenter boasting about how many board feet of lumber he cut or how many nails he sank. Certainly these are some of the activities we expect of our police officers and our carpenters. These measures clearly show that the police officer and the carpenter were busy, but they do not demonstrate that the community is safer or happier or that the home has been well built. This is not to say that arrest and citation data should not play any role in performance measurement. Rather, it is a challenge to police executives to think creatively about what these measures represent and how they might contribute to a more comprehensive performance measurement scheme.

Clearances

Like the arrest rate, the clearance rate, which is the proportion of reported crimes solved by the police, is another measure of police output that is collected widely and frequently by police agencies around the nation.[52] Despite numerous conceptual and technical problems with clearance rates, they are "the most common measure of investigative effectiveness" used by police.[53] Some critics have argued that clearance rates are beset with measurement problems.[54] For example, in his classic 1966 book, *Justice Without Trial*, Jerome Skolnick demonstrated how clearance rates are sometimes manipulated by detectives who deem certain offenses as "unfounded" due to suspicious circumstances. According to one supervisor Skolnick interviewed,

". . . we're an honest department. All these other departments that have fancy clearance rates—we know damned well they're stacking the cards. It's easy to show a low crime rate when you have a category like suspicious circumstances to use as a waste basket . . . ." Another study described how detectives manipulated clearance rates to inflate their performance measures. If they arrested a suspect, sometimes they would use the arrest to clear other similar offenses, even when the evidence that the cases were related was slim. Furthermore, the detectives demonstrated a profound disinterest in whether the "cleared" cases resulted in court convictions; they viewed their job as generating the clearance regardless of the consequences.[55]

Similarly, Gary Cordner has argued that both the numerator (cases cleared) and denominator (total reported offenses) used in computing the clearance rate are "susceptible to manipulation and measurement error."[56] A 1985 report on the future of the Uniform Crime Reporting program listed a number of problems with clearance rates that reduce their utility for measuring police performance.[57] Despite these problems with the measurement of clearance rates, they are reported routinely by police departments, and they are used routinely by researchers.[58] Nationally, clearance rates for homicide have been falling almost linearly over the past 4 decades, dropping from 92% in 1960 to 66% in 1997.[59] If clearance rates do represent investigative effectiveness, then this trend illustrates a substantial decline.

Clearance rates can be useful measures. As with arrest and citation measures, important concerns have been raised about the quality of the data, particularly when used to compare different agencies. In a later section, I will discuss some methods for ensuring quality in clearance rate data.

Response Times

The standard response to calls for service in most police departments has historically been to dispatch a sworn police officer, who responds quickly. Yet, research and experience have shown that not everybody who calls the police requires, or even necessarily expects, a rapid response. Police agencies facing resource shortages have often been able to streamline their existing resources and improve both efficiency and effectiveness by implementing some form of alternative response strategy. Collectively, these alternative responses have come to be known as "differential police response" (DPR) strategies. The development and diffusion of DPR strategies in American police agencies was informed by several influential research findings. Several studies showed that rapid response to reports of serious crimes led to an arrest less than 5% of the time.[60] For offenses in which there are no witnesses and no evidence, citizens are often willing to file a police report over the telephone.[61] A series of studies also showed that the single most important factor in citizen satisfaction with police response was whether the response time matched citizen expectations, even if the response time was lengthy. In other words, providing citizens with an accurate estimate of the response time is often more important than providing a rapid response.[62]

What challenges do response rates raise as comparative performance measures? First, community standards vary widely. Some communities demand a different police response than others. Second, communities differ in geography, topography, traffic patterns, and other features that make it difficult to compare response rates. As we will show later, it may be possible to adjust for these factors, but the scientific

foundation for generating accurate adjustment procedures is still incomplete. Third, rapid response can sometimes be a less efficient, less effective response strategy than alternative approaches. Rapid response to nuisance calls is sometimes wasteful and may detract from more important police priorities. It is possible to compute a comparative performance measure that is based on response times, but it would require careful thought. It would mean developing a uniform definition of calls requiring a rapid response and measuring the response times for only those calls. Furthermore, even these measures would require statistical adjustments (which I describe later in this report) to render them meaningful across communities of different sizes and types. Response time is important, but using it as a comparative performance measure invites several challenges. One more feasible alternative to using actual response times is to use customer satisfaction with response times as a performance measure.[63]

Toward a New Conception of Police Performance

With the evolution of community policing, police reformers have recommended an entirely new way of viewing police performance measurement.[64] The community policing reform literature suggests important changes in the way we measure police performance. Police departments and communities are urged to engage in the philosophical and conceptual work of identifying the goals that they expect the police to produce. This exercise will help the police in any community clarify their mission and expand beyond the traditional performance measures reviewed. Certainly maintaining safe communities with a good quality of life will play a role in any thoughtful analysis of the goals of policing. As I demonstrate in the next section, however, there are many more goals worth pursuing. In addition, these goals need to provide an accurate reflection of the work that police actually do. If police spend a large amount of time on traffic safety functions or maintaining community order, for instance, then those functions should play some role in the list of the goals of policing. Evaluating police departments only on their prowess in apprehending offenders ignores the vital importance of all the other work that they do. Furthermore, it relieves them of accountability for performing equally well in all of their other work. In the next section, I explore the multidimensional nature of police performance in much more detail and provide some ideas about how to specify the appropriate dimensions.

Finally, the community policing reform literature suggests that police agencies need to adopt outside-the-box thinking when generating performance measures. Police are accustomed to thinking about performance measures that exist already within the many data sets available to them. Yet, many alternatives exist. Once those interested in developing performance measures have established a list of general goals, they must then initiate the work of turning these into performance measures. Implicit in any goal is a series of more specific outcomes that reflect the general goal and can be translated into specific performance measures. For instance, suppose one of the goals is "citizen satisfaction with police." A number of more specific performance measurements might issue from this single goal. For instance, police agencies might determine the proportions of victims, witnesses, and drivers who are satisfied with the police. Perhaps citizen complaints could be used as a proxy for citizen satisfaction (though this measure is often problematic).[65] Perhaps different kinds of satisfaction might be parsed out (e.g., satisfaction with the call-taker, the response time, and the effort provided by the patrol officer or detective on the scene).

These specific measures should extend beyond the traditional measures outlined earlier. Furthermore, the methods used to collect them should vary widely; general surveys of residents, "contact" surveys with those who have had recent contact with the police, employee surveys, direct observation of community conditions or police-citizen encounters, administrative data collected by the police department, or data collected by other agencies are all permissible and can be mixed in a variety of ways. The goal is to assemble information from a wide variety of data sources that can be used to generate knowledge useful for organizational learning.

The Multi-Dimensional Nature of Police Performance: A Golden Thread

Police agencies provide a variety of public services to their communities. The nature of these services varies widely, from educating citizens about crime prevention and responding to automobile accidents, to investigating crimes and apprehending offenders. It is this variety in the day-to-day tasks that police perform that makes measuring their performance so difficult. Some agencies might do a terrific job at maintaining positive and interactive relationships with their communities but fail to be adequately prepared for critical incidents. Others may take advantage of the newest investigative and information-processing technologies while still relying on outmoded or inefficient patrol deployment strategies. In other words, police agency performance is multidimensional. Those police agencies that concentrate only on one or a handful of performance dimensions to the exclusion of others do so at their peril.

The idea that police agencies might be very successful in some ways but less successful in others is not unique to the police. It is an axiom among public organizations that performance is multidimensional. Fire departments need to excel at responding quickly to emergency situations, yet they must do so without getting into an automobile accident en route or running over a pedestrian along the way. They must rescue citizens in danger, while at the same time not incurring serious injury or deaths among the firefighters. They must manage the scene, often in concert with other agencies. They must excel in the various technical aspects of their duties, from putting out fires to administering emergency medical aid. It is not difficult to imagine a fire department that excels in one of these dimensions but performs less adequately in others. A fire department that excels at the technical aspects of putting out fires, without paying much attention to the safety of pedestrians, drivers, and its personnel will eventually find itself in crisis.

Think long enough about an organization and what its various constituencies expect of it, and it becomes rapidly apparent that performance is multidimensional in virtually every organizational setting. Even among corporations, who have a readily available measure of performance—the bottom line, or corporate profits—performance can still be thought of as multidimensional. Corporations can be rated on a variety of measures outside of the profits they generate. For instance, they are rated on their "greenness": the extent to which they engage in environmentally responsible practices.[66] Each year, *Business Ethics* selects the 100 Best Corporate Citizens, a distinction that is measured based on corporate service to seven primary stakeholders: "stockholders, employees, the community, the environment, overseas stakeholders, minorities and women, and customers."[67] They are also rated on consumer confidence, responsiveness to customers, and customer satisfaction using the *American Customer Satisfaction Index*.[68] *Consumer Reports* and other outlets routinely rate the quality of products and services offered by companies. As recent events in corporate America have demonstrated so aptly, a corporation that puts profits ahead of all other dimensions of performance, such as maintaining fair and accurate accounting and employment practices, places itself and its investors at significant risk. These risks range from poor profits to civil and criminal penalties leveled at both the organization and its leaders. Clearly, there are crucial differences between corporations and local government agencies like the police, yet even the famed bottom line is often not the only important measure of corporate performance.

One way of thinking about the dimensions of performance in organizations of any type is to consider the three *E*s: (1) equity, (2) effectiveness, and (3) efficiency.[69] Equity, also referred to as rectitude, refers to the fairness, or the ethical standards by which the organization operates.[70] It is one generic dimension of performance. Effectiveness refers to how well the organization meets its goals. This dimension can sometimes be broken down into multiple sub-dimensions since organizations often have multiple goals, which may even conflict with one another. Efficiency is a ratio of outputs or outcomes to inputs. If one firm is able to build the same bridge as another firm for half the cost, the former is twice as efficient as the latter.

These three generic dimensions are helpful for beginning to think about some of the ways that organizations might vary in terms of their performance. It is not difficult to think of departments that might score highly on some dimensions but not others. An agency might embrace fair practices throughout and produce an optimal level of public safety but require a substantial level of funding that is out of range when compared to its peer agencies. In this case, it would score highly on equity and effectiveness but lower on efficiency. These dimensions are also useful for thinking about the normative decisions a community must make about public safety. As David Bayley has pointed out in his book *Police for the Future*, hiring a cop to stand on every corner would probably reduce crime substantially, but at what cost?[71]

While the three *E*s have some value as a thought exercise, it is often more useful in practice to measure performance using dimensions that conform closely to the specific industry in question. One reason for this is that effectiveness is itself typically multidimensional. Effective police agencies might be those that produce low crime rates, low rates of revictimization, high quality of life, feelings of safety, and high clearance rates. Equity too might refer to various kinds of fairness and rectitude: to officers and other employees within the organization, to victims, to arrestees, and to those they encounter on the street. Efficiency, as a ratio in which the denominator is the measure of resources, can be expressed in different ways: per dollar, per officer, per employee, or per hour.

Rather than relying on a set of generic dimensions for measuring performance in any kind of organization, it makes a lot of sense to focus on the dimensions of performance within a particular industry. The purpose of this report is *not* to suggest the dimensions on which performance might be measured. Instead, the idea is to clearly demonstrate that whatever performance measurement scheme is selected must, first and foremost, recognize that performance is multidimensional. That sets up the rather daunting task of deciding upon the relevant dimensions for policing.

CALEA, with its ready access to forward-thinking police executives from around the nation, is situated ideally to engage in this process.

The next section reviews some of the measures contained in previous research on performance measurement in policing. Some of these dimensions relate to conditions internal to the police organization, such as structure, management, and policy. Others relate to the way the organization interacts with its environment.

The Dimensions of Performance

Dimensions are independent categories or domains of a characteristic or property. I have found that a useful example or metaphor for thinking about the dimensions of police performance worth measuring is the idea of intelligence testing. Researchers who study human intelligence have debated the dimensions of intelligence for many years. Some theorists argue that intelligence is comprised of a single dimension known as "g" or general intelligence, a unitary phenomenon.[72] This reasoning is what led to the development of standardized IQ tests, which result in a single overall score that measures intelligence. Others believe that intelligence is a multidimensional phenomenon and that there are really multiple kinds of intelligence. For instance, psychologist Howard Gardner argues that there are seven dimensions: (1) linguistic, (2) logical, (3) spatial, (4) musical, (5) kinesthetic, (6) interpersonal, and (7) intrapersonal.[73] Furthermore, these seven dimensions are independent of one another; therefore, it is possible for a person to exhibit high intelligence in one area (such as logical) but low or moderate performance in another (such as musical). Other "multiple-intelligence" theorists eschew the idea of an overall general intelligence but disagree with Gardner on the number or nature of the dimensions.

The debate over the dimensions of human intelligence is a helpful metaphor for thinking about the performance of police departments. Treating performance as a unidimensional phenomenon means that "good" departments are good at all aspects of policing, while poor departments are poor in all aspects. As I proposed in the previous section, one of the most useful methods for thinking about police performance is to avoid the tendency to view it as unidimensional. Police performance is multidimensional; the number and nature of those dimensions is a matter for speculation and debate. In this section, I review a handful of methods outlined in the past for categorizing the dimensions of police performance.

In 1980, Michael O'Neill and his colleagues developed the Police Program Performance Measurement system. It was a modular performance measurement system "into which each locally based organization could plug its own goals and objectives." As part of this exercise, the authors prepared a "model structure of police objectives" containing five dimensions.[74]

O'Neill's Five Dimensions

1. Crime prevention
2. Crime control
3. Conflict resolution
4. General service
5. Police administration

Within these five dimensions were 46 specific outcomes that were operationalized into 65 performance measures. This effort, like all of the others presented here, has not been institutionalized widely. It is now part of the historical record of police performance measurement.

Another system, devised by Harry Hatry and his colleagues at the Urban Institute and ICMA contains five dimensions of police performance.[75] This proposed system, like the one before it, has not yet led to a widespread, systematic suite of performance measures institutionalized across the nation. One important area for reflection among police executives is why such measures get so much lip service but so little action.

Hatry's Five Dimensions

1. Prevention of crime
2. Apprehension of offenders
3. Responsiveness of police
4. Feeling of security
5. Fairness, courtesy, helpfulness/cooperativeness, honesty

The most recent framework for measuring police performance was developed by Professor Mark Moore and several colleagues at Harvard's Kennedy School of Government. Their work appears in a book, entitled *Recognizing Value in Policing: The Challenge of Measuring Police Performance*, published in 2002 by the Police Executive Research Forum. Moore's framework lays a solid intellectual foundation for measuring seven dimensions of police performance.

Moore's Seven Dimensions

1. Reduce criminal victimization.
2. Call offenders to account.
3. Reduce fear and enhance personal security.
4. Guarantee safety in public spaces.
5. Use financial resources fairly, efficiently, and effectively.
6. Use force and authority fairly, efficiently, and effectively.
7. Satisfy customer demands/achieve legitimacy with those policed.

One element of policing that often gets overlooked in performance measurement is the nature of the policing process. Stephen Mastrofski has outlined a spectrum of humanistic concerns that he terms "Policing for People."[76] According to Mastrofski, traditional police performance measures ignore a fundamental element of the relationship between police and communities: the nature of police-citizen encounters. He highlights six features of these encounters that should be measured. Like other variables we have discussed, these are characteristics of individual encounters and officers, but in the aggregate, they can be used to characterize and compare police agencies over time and place. One option is to think of these as full dimensions, but a more likely solution is to think of them as subdimensions of a single broader dimension that focuses on the nature of the policing process (such as Mark Moore's seventh dimension).

Mastrofski's Six Dimensions: "Policing for People"

1. Attentiveness
2. Reliability
3. Responsiveness
4. Competence
5. Manners
6. Fairness

Research has shown that current data on policing is insufficient for either measuring performance or doing adequate comparative research on police organizations because it fails to capture the full range of work that police do.[7] Whatever dimensions one chooses, they should reflect a full and realistic range of police functions and goals. It may be that some of these functions are more important than others; this and other technical issues are discussed later.

The next segment will show how to use the principles and concepts introduced in this first segment to create comparative performance measures. It will show how such measures can be developed nationally in the policing industry, as well as how you can begin to implement them in your agency.

Endnotes

[1] Pasi, 2000
[2] Thomas & Hofer, 1999
[3] www.irs.gov/pub/irs_pdf/p3560.pdf
[4] Hoover, 1984
[5] Osborne & Gaebler, 1992
[6] Hammer & Champy, 1993
[7] Cohen & Brand, 1993
[8] Ammons, 1996
[9] Kaplan & Norton, 1996
[10] Geller, 1997
[11] Maguire & Uchida, 2000
[12] The material in this paragraph was drawn from Maguire and Uchida (2000) and Uchida, Bridgeforth, and Wellford (1986).
[13] Parks, 1971, p. 9; Stone, 1930
[14] Bellman, 1935
[15] O'Neill, Needle, & Galvin, 1980, p. 254
[16] O'Neill et al., 1980, p. 254. Also see Parrat (1937).
[17] O'Neill et al., 1980
[18] President's Commission on Law Enforcement and Administration of Justice, 1967
[19] According to the Bureau of Justice Statistics, 30% of local police conducted citizen surveys in 1997, and 27% conducted such surveys in 1999. See Hickman and Reaves (2001).
[20] Smith, Steadman, Minton, & Townsend, 1999
[21] Uchida, Bridgeforth, & Wellford, 1986
[22] Maguire, 2002; Uchida et al., 1986
[23] Maguire, 2002
[24] Alpert & Moore, 1993, p. 110

[25] Barlow & Barlow, 2000; Walker, 1980
[26] Walker, 1980
[27] Bayley & Mendelsohn, 1969; Reiss, 1971, p. 147; Westley, 1970, pp. 99-104. To be fair, research in the 1960s also showed that police were not substantially more racist than others in the communities they served (Bayley & Mendelsohn, 1969).
[28] Walker, 1980
[29] Cassell & Fowles, 1998; Leo, 1996
[30] Hoffman, 1971, pp. 172-173
[31] Maguire & Uchida, 2000
[32] Alpert & Moore, 1993; Kelling, 1996; Maguire & Uchida, 2000; Moore, 1999, 2002
[33] Bratton, 1999, p. 15
[34] Bratton, 1998; Also see Moore (1999).
[35] Moore, 1999
[36] Kelling, 1996, p. 32
[37] Unadjusted crime rates are simply those that are provided by the FBI without any transformations. Later in this article, these unadjusted measures are contrasted with "risk-adjusted" crime rates. These are crime rates that have been adjusted statistically to account for differences between communities (like income inequality or population density) that might be expected to affect crime rates.
[38] Sherman, 1998
[39] McLeary, Nienstedt, & Erven, 1982
[40] Black, 1970; McLeary et al., 1982; Seidman & Couzens, 1974
[41] Decker, 1980, 1981
[42] Maguire & Uchida, 2000; Maltz, 1999
[43] Sherman, 1980a, 1980b
[44] Sherman, 1980b, p. 468
[45] Sherman & Glick, 1984
[46] Kelling, 1992
[47] Goldstein, 1990
[48] Langan, Greenfeld, Smith, Durose, & Levin, 2001
[49] Gardiner, 1969
[50] Feeney, 1982, p. 38
[51] Bratton, 1998, p. 229
[52] Once a reported offense is designated as founded, it may be cleared in two ways: (1) by arrest or (2) by exception (FBI, 1980; Martin & Besharov, 1991). A crime is cleared by arrest when at least one person is "arrested; charged with the commission of the offense; and turned over to the court for prosecution (whether following arrest, court summons, or police notice)" (FBI, 1980, p. 394). An offense is cleared by exception when the investigation has clearly identified the offender; there is sufficient information on which to make an arrest; the offender's location is known; and extenuating circumstances outside the control of the police make it difficult to make an arrest (FBI, 1980, p. 395). One of the most common reasons for clearing an offense by exception is the unwillingness of victims and/or witnesses to testify in court. Another instance is when the offender is dead (Cordner, 1989).
[53] Cordner, 1989, p. 146
[54] Alpert & Moore, 1993; Hoffman, 1971; Riedel & Jarvis, 1999
[55] Nadel, 1978
[56] Cordner, 1989, p. 146

[57] Poggio, Kennedy, Chaiken, & Carlson, 1985; Also see Schneider and Wiersema (1990).
[58] Davenport, 1996
[59] Riedel & Jarvis, 1999, pp. 279, 301
[60] Caron, 1980; Eck & Spelman, 1987; Spelman & Brown, 1981; Sumrall, Roberts, & Farmer, 1981; Also see Bracey (1996) for an overview.
[61] Eck, 1983
[62] Pate, 1976; Percy, 1980; Tien, Simon, & Larson, 1979
[63] Hatry, 1999, p. 23
[64] Alpert, Flynn, & Piquero, 2001; Alpert & Moore, 1993; Horne, 1992; Langworthy, 1999
[65] Walker, 1998, 2001
[66] National Advisory Council for Environmental Policy and Technology, 2001
[67] Miller, 2002
[68] Anderson & Fornell, 2000; Fornell, 2001
[69] Bayley, 1994; Eck & Rosenbaum, 1994
[70] Bayley, 1994
[71] Bayley, 1994
[72] Spearman, 1927
[73] Gardner, 1983
[74] O'Neill et al., 1980, p. 257
[75] Hatry et al., 1992
[76] Mastrofski, 1999; Also see Eck & Maguire (2000) and Moore & Poethig (1999).
[77] Maguire, 2002; Maguire & Uchida, 2000

References

Alpert, G. P., Flynn, D., & Piquero, A. (2001). Effective community policing performance measures. *Justice Research and Policy, 3*(2), 79-94.

Alpert, G. P., & Moore, M. H. (1993). Measuring police performance in the new paradigm of policing. In J. J. Dilulio, Jr. (Ed.), *Performance measures for the criminal justice system* (pp. 109-142). Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

Ammons, D. N. (1996). *Municipal benchmarks: Assessing local performance and establishing community standards*. Thousand Oaks, CA: Sage.

Anderson, E. W., & Fornell, C. (2000). Foundations of the American Customer Satisfaction Index. *Journal of Total Quality Measurement, 11*(7), S869-S882.

Barlow, D. E., & Barlow, M. H. (2000). *Police in a multicultural society: An American story*. Prospect Heights, IL: Waveland.

Bayley, D. H. (1994). *Police for the future*. New York: Oxford University Press.

Bayley, D. H., & Mendelsohn, H. (1969). *Minorities and the police: Confrontation in America*. New York: The Free Press.

Bellman, A. (1935). A police service rating scale. *Journal of Criminal Law, Criminology, and Police Science, 26*, 79.

Black, D. J. (1970). The production of crime rates. *American Sociological Review, 35*(4), 733-748.

Bracey, D. H. (1996). Assessing alternative responses to calls for service. In L. T. Hoover (Ed.), *Quantifying quality in policing* (pp. 153-166). Washington, DC: Police Executive Research Forum.

Bratton, W. (with Knobler, P.). (1998). *Turnaround: How America's top cop reversed the crime epidemic*. New York: Random House.

Bratton, W. J. (1999, July). How higher expectations for police departments can lead to a decrease in crime. In R. H. Langworthy (Ed.), *Measuring what matters: Proceedings from the Policing Research Institute meetings* (pp. 11-26). Washington, DC: National Institute of Justice and the Office of Community Oriented Policing Services.

Caron, N. (1980). *Response time analysis, Vol. III: Part II crime analysis*. Kansas City, MO: Kansas City Police Department.

Cassell, P. G., & Fowles, R. (1998). Handcuffing the cops?: A thirty-year perspective on Miranda's harmful effects on law enforcement. *Stanford Law Review, 50*, 1055.

Cohen, S., & Brand, R. (1993). *Total quality management in government: A practical guide*. San Francisco: Jossey-Bass.

Cordner, G. W. (1989). Police agency size and investigative effectiveness. *Journal of Criminal Justice, 17*, 145-155.

Davenport, D. R. (1996). *Public agency performance and structure: Assessing the effects of the organizational environment*. Doctoral dissertation, Texas Tech University, Lubbock, 1990.

Deakin, T. (1988). *Police professionalism: The renaissance of American law enforcement*. Springfield, IL: Charles C. Thomas.

Decker, S. H. (1980). The effect of police characteristics on alternative measures of police output. *Criminal Justice Review, 5*(2), 34-40.

Decker, S. H. (1981). Alternate measures of police output. *American Journal of Police, 1*(1), 23-35.

Eck, J. E. (1983). *Solving crimes: The investigation of burglary and robbery*. Washington, DC: Police Executive Research Forum.

Eck, J., & Maguire, E. R. (2000). Have changes in policing reduced violent crime? An assessment of the evidence. In A. Blumstein & J. Wallman (Eds.), *The crime drop in America* (pp. 207-265). New York: Cambridge University Press.

Eck, J. E., & Rosenbaum, D. P. (1994). The new police order: Effectiveness, equity, and efficiency in community policing. In D. P. Rosenbaum (Ed.), *The challenge of community policing: Testing the promises* (pp. 3-23). Thousand Oaks, CA: Sage.

Eck, J. E., & Spelman, W. (1987). *Problem-solving: Problem-oriented policing in Newport News*. Washington, DC: Police Executive Research Forum.

Federal Bureau of Investigation (FBI). (1980). *Uniform Crime Reporting program data: Offenses known and clearances by arrest, 1980* (Codebook). Ann Arbor, MI: Inter-University Consortium for Political and Social Research.

Feeney, F. F. (1982). *The police and pretrial release*. Lexington, MA: Lexington Books.

Fornell, C. (2001). The science of satisfaction. *Harvard Business Review, 79*(3), 120-121.

Gardiner, J. A. (1969). *Traffic and the police: Variations in law enforcement policy*. Cambridge, MA: Harvard University Press.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: BasicBooks.

Geller, W. A. (1997, December). Suppose we were really serious about police departments becoming "learning organizations"? *National Institute of Justice Journal* (234), 2-8.

Goldstein, H. (1990). *Problem-oriented policing*. New York: McGraw-Hill.

Hammer, M., & Champy, J. (1993). *Reengineering the corporation*. New York: HarperCollins.

Hatry, H. P. (1999). *Performance measurement*. Washington, DC: Urban Institute.

Hatry, H. P., Blair, L. H., Fisk, D. M., Greiner, J. M., Hall, J. R., Jr., & Schaenman, P. S. (1992). *How effective are your community services? Procedures for measuring their quality* (2nd ed.). Washington, DC: Urban Institute and International City/County Management Association.

Hickman, M. J., & Reaves, B. A. (2001). *Community policing in local police departments, 1997 and 1999*. Washington, DC: Bureau of Justice Statistics.

Hoffman, R. B. (1971). Performance measurements in crime control. *Journal of Research in Crime & Delinquency, 8*(2), 165-174.

Hoover, W. J. (1984). Measuring the productivity of managers in a municipal police department. *FBI Law Enforcement Bulletin, 53*(9), 7-13.

Horne, D. G. (1992). Public opinion surveys: Implications for police organizations. *Canadian Police College Journal, 16*(4), 263-281.

Kaplan, R., & Norton, D. (1996). *The balanced scorecard*. Boston: Harvard Business School Press.

Kelling, G. L. (1992, Spring). Measuring what matters: A new way of thinking about crime and public order. *The City Journal*, 21-33.

Kelling, G. L. (1996). Defining the bottom line in policing: Organizational philosophy and accountability. In L.T. Hoover (Ed.), *Quantifying quality in policing* (pp. 23-36). Washington, DC: Police Executive Research Forum.

Langan, P. A., Greenfeld, L. A., Smith, S. K., Durose, M. R., & Levin, D. J. (2001). *Contacts between police and the public: Findings from the 1999 National Survey*. Washington, DC: Bureau of Justice Statistics.

Langworthy, R. H. (1999). *Measuring what matters: Proceedings from the Policing Research Institute meetings*. Washington, DC: National Institute of Justice.

Leo, R. A. (1996). The impact of Miranda revisited. *Journal of Criminal Law and Criminology, 86*(3), 621-692.

Maguire, E. R. (2002, Fall). Multiwave establishment surveys of police organizations. *Justice Research and Policy, 4*, 39-60.

Maguire, E. R., & Uchida, C. D. (2000). Measurement and explanation in the comparative study of American police organizations. In D. Duffee (Ed.), *Criminal justice 2000, Vol. 4: Measurement and analysis of crime and justice* (pp. 491-557). Washington, DC: National Institute of Justice.

Maltz, M. D. (1999). *Bridging gaps in police crime data: A discussion paper from the BJS Fellows Program*. Washington, DC: Bureau of Justice Statistics.

Martin, S. E., & Besharov, D. J. (1991). *Police and child abuse: New policies for expanded responsibilities. Issues and practices series*. Washington, DC: National Institute of Justice.

Mastrofski, S. D. (1999, March). *Policing for people. Ideas in American policing*. Washington, DC: U.S. Department of Justice, Police Foundation.

McLeary, R., Nienstedt, B. C., & Erven, J. M. (1982). Uniform Crime Reports as organizational outcomes: Three time series quasi-experiments. *Social Problems, 29*, 361-372.

Miller, M. (2002). 100 best corporate citizens: America's most responsible and profitable major public companies. *Business Ethics, 16*(2).

Moore, M. H. (1999, April 12). Private-sector lessons for New York's finest. *Wall Street Journal*, A22.

Moore, M. H. (2002). *Recognizing value in policing: The challenge of measuring police performance*. Washington, DC: Police Executive Research Forum.

Moore, M. H., & Poethig, M. (1999). The police as an agency of municipal government: Implications for measuring police effectiveness. In R. H. Langworthy (Ed.), *Measuring what matters: Proceedings from the Policing Research Institute meetings* (pp. 151-168). Washington, DC: National Institute of Justice and the Office of Community Oriented Policing Services.

Nadel, S. W. (1978). Measurement systems and organizational goals in a large metropolitan police department. *Police Studies, 1*(4), 39-45.

National Advisory Council for Environmental Policy and Technology. (2001). *Green dividends? The relationship between firms' environmental performance and financial performance.* Washington, DC: United States Environmental Protection Agency.

O'Neill, M. W., Needle, J. A., & Galvin, R. T. (1980). Appraising the performance of police agencies: The Police Program Performance Measures (PPPM) System. *Journal of Police Science and Administration, 8*(3), 253-264.

Osborne, D., & Gaebler, T. (1992). *Reinventing government: How the entrepreneurial spirit is transforming the public sector*. Reading, MA: Addison-Wesley.

Parks, R. B. (1971). *Measurement of performance in the public sector: A case study of the Indianapolis Police Department.* Doctoral dissertation, Pub. No. W71-1, Indiana University, Department of Political Science, Bloomington.

Parrat, S. D. (1937). A critique of the Bellman Police Service Rating Scale. *Journal of Criminal Law, Criminology, and Police Science, 28*, 895-905.

Pasi, R. (2000). The SOL: No easy answers. *Educational Leadership, 57*(5), 75-76.

Pate, T. (1976). *Police response time: Its determinants and effects*. Washington, DC: Police Foundation.

Percy, S. (1980). Response time and citizen evaluation of police. *Journal of Police Science and Administration, 8*(1), 75-86.

Poggio, E. C., Kennedy, S. D., Chaiken, J. M., & Carlson, K. E. (1985, May). *Blueprint for the future of the Uniform Crime Reporting Program: Final report of the UCR study.* Washington, DC: U.S. Department of Justice.

President's Commission on Law Enforcement and Administration of Justice. (1967). *The challenge of crime in a free society*. Washington, DC: U.S. Government Printing Office.

Reiss, A. J., Jr. (1971). *The police and the public.* New Haven, CT: Yale University Press.

Riedel, M., & Jarvis, J. (1999). The decline of arrest clearances for criminal homicide: Causes, correlates and third parties. *Criminal Justice Policy Review, 9*(3), 279-306.

Schneider, V. W., & Wiersema, B. (1990). Limits and uses of the Uniform Crime Reports. In D. L. MacKenzie, P. J. Baunach, & R. Royberg (Eds.), *Measuring crime: Large-scale, long-range efforts* (pp. 21-48). Albany, NY: SUNY Press.

Seidman, D., & Couzens, M. (1974, Spring). Getting the crime rate down: Political pressure and crime reporting. *Law and Society Review, 8*, 457-493.

Sherman, L. W. (1980a). Enforcement workshop: Defining arrest, practical consequences of agency differences (Part 1). *Criminal Law Bulletin, 16*(4), 376-380.

Sherman, L. W. (1980b). Enforcement workshop: Defining arrest, practical consequences of agency differences (Part 2). *Criminal Law Bulletin, 16*(5), 468-471.

Sherman, L. W. (1998a). Evidence-based policing. *Ideas in American Policing*, Washington, DC: U.S. Department of Justice, Police Foundation.

Sherman, L. W. (1998b). Policing for crime prevention. In L. W. Sherman, D. Gottfredson, D. MacKenzie, J. Eck, P. Reuter, & S. Bushway (Eds.), *Preventing crime: What works, what doesn't, what's promising. A report to the United States Congress*. College Park, MD: University of Maryland.

Sherman, L. W., & Glick, B. D. (1984, August). *The quality of police arrest statistics.* Washington, DC: Police Foundation.

Skolnick, J. H. (1966). *Justice without trial.* New York: Wiley and Sons.

Smith, C. B., Jr. (1923). The adequacy of police forces. *Journal of the American Institute for Criminal Law and Criminology, 13*, 266-271.

Smith, S. K., Steadman, G. W., Minton, T. D., & Townsend, M. (1999). *Criminal victimization and perceptions of community safety in 12 cities, 1998*. Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* London: Macmillan.

Spelman, W., & Brown, D. (1981). *Calling the police: Citizen reporting of serious crime.* Washington, DC: Police Executive Research Forum.

Stone, D. C. (1930, September). Can police effectiveness be measured? *Public Management, 12*, 465-471.

Sumrall, R., Roberts, J., & Farmer, M. (1981). *Differential police response strategies.* Washington, DC: Police Executive Research Forum.

Thomas, J. W., & Hofer, T. P. (1999). Accuracy of risk-adjusted mortality rate as a measure of hospital quality of care. *Medical Care, 37*(1), 83-92.

Tien, J., Simon, J., & Larson, R. (1979). *An alternative approach to police patrol: The Wilmington split-force experiment*. Cambridge, MA: Public System Evaluation.

Uchida, C. D., Bridgeforth, C., & Wellford, C. F. (1986). Law enforcement statistics: The state of the art. *American Journal of Police*, 5(1), 23-43.

Walker, S. (1980). *Popular justice*. New York: Oxford University Press.

Walker, S. (1998, September). Achieving police accountability. *Research Brief, The Center on Crime, Communities and Culture*, 3, 1-16.

Walker, S. (2001). *Police accountability: The role of citizen oversight*. Belmont, CA: Wadsworth.

Westley, W. A. (1970). *Violence and the police: A sociological study of law, custom, and morality*. Boston: Massachusetts Institute of Technology.

Part 2 of a two-part article appearing in the *CALEA Update*

Introduction

This is the second segment of a two-part article on measuring the performance of law enforcement agencies. It is written for a policing audience and draws in part on my discussions with members of CALEA's Performance Measurement Subcommittee and those who have attended my training workshops at the last three CALEA Conferences. In the first segment, I introduced the general concepts, terminology, and history of comparative performance measurement in policing. In this second segment, I show you how to develop, pilot-test, and implement comparative performance measurement in your agency. This article is one small part of a larger effort by CALEA to explore the feasibility and utility of agency-level performance measurement in policing. That journey is just beginning and will proceed slowly, but it will be a worthwhile one.

Developing Police Performance Measures

Once the goals of policing have been identified and a salient list of dimensions (and perhaps sub-dimensions) has been prepared, it is time to begin formulating specific performance measures. Many times, people start off in the middle of the process by generating a list of performance measures without having first completed the necessary and far more difficult work of thinking about the broad dimensions of police performance. The process I am recommending is a rough analogue to the deductive model of science in which we start by identifying theories and concepts and then collect data on specific measures that reflect those broader theories and concepts. The search for specific performance measures should be a liberating, unconstrained process in which participants are encouraged to think well outside of the traditional boundaries. In this section, I provide a number of suggestions about potential data sources and research methods for generating performance measures.

Traditional performance measures in policing are often derived from administrative data maintained by the police department. While this data can often be very useful and should be included, official police data should not be the only source used in a comprehensive performance measurement system. Some other options are presented below.

General Community Surveys

Nationwide, nearly one-third of local police departments conduct community surveys each year. The Bureau of Justice Statistics now makes available to police departments a free software package and guide for conducting community surveys.[78] These kinds of surveys are useful for several purposes: learning about crime, fear of crime, victimization experiences, and overall impressions about the police.[79] They are sometimes used as a crutch, however. Research on customer and client satisfaction across industry types has shown that satisfaction levels reported in response to general survey questions are routinely high and do not tend to differ greatly across organizations.[80] Other research, however, shows that the specificity and wording of the survey question can have a profound impact on satisfaction levels.[81] Therefore, police organizations can get out of a community survey what they put into it. If they want a public relations gimmick, they can ask one or two very general questions about citizen satisfaction with police. If they are interested in using the survey as a platform for organizational learning, they can ask a number of more specific questions about the quality of policing in the community. Another problem with general community surveys is that many of the respondents have not had any contact with the police; therefore, it is difficult to know whether their impressions of the police were formed through the media, through vicarious contact with friends or relatives, or through previous experiences with *other* police organizations.

Contact Surveys

Contact surveys are administered to those who have had recent contact with the police. These kinds of surveys can be very revealing, particularly when they are focused on different kinds of contacts. Surveys of victims, for instance, can be useful for learning whether the department is responding appropriately to their needs. When police in Toronto surveyed rape victims, they received numerous complaints about the uniformed officers who responded to the initial call but nearly universal praise for the department's sex crimes unit.[82] Surveys of drivers stopped and/or searched by the police can be used to learn about citizen perceptions of police practices. Even arrestee surveys can be quite useful. Although a common perception is that such surveys would not be useful because all arrestees will be dissatisfied with the outcome, research has shown that citizens are willing to accept negative outcomes if they view the process that led to the outcome as fair.[83] Arrestee surveys administered in multiple cities could be useful for learning whether a department is perceived as more or less fair than others.[84] Contact surveys could also be administered over time within a single department to learn whether certain training programs or supervisory approaches are improving citizen perceptions of police.

Employee Surveys

Employee surveys are valuable for many reasons. They can be used to gauge the perceptions of employees about certain administrative initiatives. They can be used

to assess morale issues. Employee surveys have also been used in some unique and helpful ways in recent years. For instance, researchers from the Federal Bureau of Prisons aggregated (combined) data from individual employee surveys to form composite measures of the organizational social climate in the Bureau's various prison facilities.[85] A similar approach was recently applied to the measurement of police integrity. Researchers aggregated the responses of more than 3,000 individual police officers to form an aggregate measure of the "environment of integrity" in 30 police agencies.[86] The results showed that police agencies vary widely with regard to their overall environments of integrity. This information was presumably quite useful to police executives in those agencies, particularly those who ranked at the bottom of the list.

Direct Observation

Direct observation by trained raters or coders can also be a useful method for collecting valuable performance information. For example, the Early Childhood Environment Rating Scale uses trained observers to rate the quality of childcare facilities based on direct observation of the space and furnishings, the interaction between children and teachers, and several other dimensions of performance.[87] Observers in New York City use vehicles equipped with specially designed measuring instruments to rate the "smoothness" of 670 miles of streets in 59 districts.[88] In policing and criminology, there is some precedent for using direct observation to generate data useful for performance measurement. For instance, coders can use "systematic social observation" techniques to record the volume of physical and social disorder in neighborhoods.[89] This is a useful technique for generating data, independent of police, on quality-of-life issues in the community. Using techniques developed by Mastrofski and his colleagues, trained coders can also conduct standardized observations of police-citizen encounters.[90] While direct observation can be a useful technique for gathering data on performance, it is personnel intensive, and therefore very expensive.

Independent Testing or Simulation Studies

Another alternative source for collecting data on police performance is independent testing or simulation studies. Rather than observing performance in completely "natural" settings, independent tests create artificial opportunities to measure performance. For instance, the Insurance Institute for Highway Safety uses crash tests to rank vehicle safety. The Institute's primary mission is research, and insurance companies rather than auto manufacturers fund it. Many firms hire people to pose as customers (known as "secret shoppers" or "mystery shoppers") who visit their facilities to perform checks on quality of service, cashier accuracy, ethical standards, and many other issues. Internal affairs units in large police agencies have conducted various kinds of "integrity tests" for many years. ABC News conducted independent integrity tests of police in New York and Los Angeles by turning over 40 wallets or purses to police officers chosen at random. In every case, the officers turned in the wallets and purses with contents intact.[91] The Police Complaint Center (PCC) is a Florida-based firm that conducts proactive investigations of police misconduct. The PCC videotapes its investigators in a variety of settings: being stopped by officers, trying to secure complaint forms from police agencies, and other situations. PCC investigators have videotaped numerous instances of police misconduct.

While certainly controversial, testing and simulation offer promise for collecting performance data that are truly independent of the police.

Computing Aggregate Measures

Once the performance measures have been selected and the data has been collected, the next question is what kind of analysis to perform. The first step, depending on the data being collected, is to aggregate the data to compute an overall organizational score for each individual performance measure. If the measure is a count variable (such as the number of arrests), it can be summed, or an average or ratio can be computed. If the measure is categorical (e.g., a survey question with five response categories ranging from strongly disagree to strongly agree), the proportion of people choosing each response can be computed. Since a comparative performance measure is intended to measure some aspect of the organization, each measure needs to be aggregated so it represents an organization-wide score.

Standardized Composites or Individual Measures?

When a student takes the SAT, the GRE, or other similar standardized tests, the overall scores represent "composites" of the individual test questions. These composite scores are standardized so that they fall within a certain well-known range, such as 200-800 for the SAT. No one is very interested in performance on individual test questions, only the overall score within each dimension (e.g., math and verbal). Anytime we create new performance measures, we have a series of analytical choices about how we want to use the data. For instance, suppose we generate a list of 7 general dimensions of police performance. Within each one, we collect data on 7 specific performance measures. We will end up with 49 (7 x 7) separate pieces of information from each organization. One possibility is to treat each of the 49 specific items as a performance measure. In some ways, this is analogous to inspecting a student's performance on each individual SAT question. There are commonly used statistical methods, however, that can be used to reduce these 49 separate items into 7 composite scores representing each overall dimension. Furthermore, these composites can also be standardized (e.g., an agency would receive a score falling between 0 and 100 on each dimension). This approach is common in psychology when making standardized instruments to measure a variety of individual traits. Our efforts to create performance measures nationally will focus on creating composite scores. Local police departments implementing performance measures on their own may not have access to sufficient statistical expertise to form composite measures.

Weighting

One of the complaints about some performance measurement systems is that they treat each measure equally. This is acceptable as long as the different domains of performance are equivalent, but if some are much more important than others, it is misleading. Sometimes it is useful to assign greater weight to certain measures when computing composite performance scores. There are a variety of methods for doing this; they require technical expertise, but they can be executed easily. The more pressing question is how to assign the weights in a manner that is not totally arbitrary. For instance, former NYPD Commissioner Bratton explained that his strategies for reducing crime in New York came with some consequences:

We defined brutality as unnecessary behavior that caused broken bones, stitches, and internal injuries. But those were not the figures that had gone up significantly. What had risen were reports of police inappropriately pushing, shoving, sometimes only touching citizens. We were taking back the streets . . . we were being more proactive, we were engaging more people, and often they didn't like it.[92]

Implicit in this explanation is the argument that crime control is a *more important* function of policing than citizen satisfaction or appropriate use of minor levels of force. Policing is certainly not the only industry in which these kinds of questions arise. As Gormley and Weimer note, "A physician with a good bedside manner is not enough when a patient's life is at stake. A teacher with a winning smile is not enough if challenging subjects are being taught."[93]

Some goals may be more important than others. An important decision for those designing comparative performance measures is how to quantify differences in importance between multiple goals. If the differences are minor, they may be worth overlooking. If there are major differences in importance, such as the friendliness of the hospital staff versus its mortality rate, then it will be useful to either consider each performance measure individually, or to use a weighting procedure before forming composite scores.

How can the weights be formed? One method might be to use an expert group and ask them to compile a ranking system. Mark Moore and his colleagues at Harvard's Kennedy School of Government have already used a similar approach for ranking the most important innovations in policing.[94] Focus groups or surveys of citizens could also be used to determine which goals are the most important to them. Once again, a national system of performance measurement should take pains to compute weights for each dimension of performance. Local law enforcement agencies might not have access to the statistical expertise necessary to form an elaborate weighting system, but they should still go through the process of thinking about which dimensions of performance are most important.

After the performance measures have been specified; the data has been collected; and the analysis has proceeded through the possible stages of aggregation, formation of composites, and weighting, it is time to use the data to make comparisons. In the next section, I examine two methods for ensuring that comparisons are as fair as possible.

Making Fair Comparisons

In 1923, Clarence Smith raised a number of concerns about using statistics to compare police departments. His argument, quite simply, was that police in different communities face different circumstances that need to be taken into account when comparing agencies. These differences range from demographic and economic features to topography and culture, including race; population density; the nature of industrial development; the condition and distribution of the streets and highways; the volume of tourist traffic; and "the habits, traditions, and natural law-abiding inclination and disposition of the people of the city."[95] Smith's concern with comparative statistics is apropos today as we consider how to develop the systematic capacity for comparative performance measurement in policing.

This concern with making fair comparisons is not unique to the police. It affects all kinds of organizations. For instance, students graduating from Harvard University are likely to have higher GRE scores than those of students attending state universities. Does that mean Harvard performs at a higher level than state universities? Not necessarily. The typical student admitted to Harvard presumably entered with much greater aptitude and higher SAT scores than the typical state university student. The important question is not whether one organization has better inputs than the other but which one adds more value. The key point here is that organizations often have different inputs, and this variation in inputs should be reflected in performance measures. This notion of "value-added" applies to schools, hospitals, police departments, and many other kinds of organizations.

Like other kinds of organizations, police departments face drastically different workloads, challenges, and environments. One department might work in a poor, ethnically heterogeneous community with high rates of crime and disorder. Another might work in a wealthy, ethnically homogenous, sleepy suburb in which a patrol officer's greatest challenge is to write traffic citations and make an occasional arrest. The key to comparing these two organizations, despite their differences, is known as risk adjustment. Hospitals admitting the most at-risk patients might be expected to have the highest death rates. Prisons admitting the worst offenders might be expected to have the highest recidivism rates; however, inputs of an organization can be controlled when measuring performance. There are two primary methods: (1) stratification (forming peer groups) and (2) calculating "risk-adjusted" performance measures.

Stratification, or forming peer groups containing similar agencies, is one useful way to account for differing inputs. Groups of agencies that are approximately similar in size, type, jurisdiction, and workload will become peers. Each agency within the peer group can compare its performance measurements with the other peer agencies. Forming peer groups is much easier than doing risk adjustment, but it too will be tricky. Some cities are simply unique. Others may belong in certain classes of cities that are difficult to identify in advance. For instance, some "edge cities" have a small population, but due to their proximity to large urban areas, they may face issues that make them unique compared to other similarly sized communities.[96] Nonetheless, the difficulties inherent in peer groups are much less formidable than the difficulties with risk adjustment.

Criminologist Lawrence Sherman acknowledges that cities vary widely with regard to the social and economic correlates of crime. He proposes using statistical methods to purge homicide rate measures of the influence of these other factors. The resulting measure will be a "risk adjusted homicide rate" that is similar to the risk adjusted mortality rates used by hospitals. For instance, one could use relatively straightforward statistical techniques (e.g., regression analysis) to purge homicide rates of the influence of poverty, unemployment, race, divorce, and population density. Once such factors are controlled statistically, the resulting measure can more easily be compared across cities, even those that are very different from one another. One research team has already created a prototype ranking system based on risk-adjusted homicide rates for 21 cities.[97] Sherman suggests that such a measure can be used to rank the performance of police agencies at dealing with crime.[98] This process will require technical expertise and a substantial investment in testing and calibration to ensure that the risk-adjustment procedures are scientifically

defensible. Furthermore, since risk-adjusted crime rates are based on an implicit assumption that demographic and structural characteristics (e.g., poverty, race, and region) influence crime, the risk-adjustment procedures might inspire controversy. Although the research on risk adjustment in policing extends back to at least 1971,[79] much more work remains to be done before police agencies can rely on its scientific foundation. For now, comparative performance measurement initiatives will need to rely on stratification or the formation of peer groups.

Implementing Comparative Performance Measures in Your Agency

Much of this background information might leave you wondering how to establish comparative performance measures in *your* agency. This section walks you through the steps, providing brief pointers to help keep you on track.

1. **Make a commitment to *comparative* performance measures.**
   - This involves comparing your agency's performance over multiple time periods or comparing your agency to other agencies. Other options include comparing district or beat-level performance within a large agency.
   - Conducting *one* data collection exercise (e.g., a citizen survey) in *one* jurisdiction does not provide you with a "comparative" performance measure because it does not offer the opportunity for comparison.

2. **Select the units that you intend to compare.**
   - Will you compare time periods (months, years), beats, districts, or different agencies?
   - Use caution in selecting "peer" agencies. Make sure they are comparable.

3. **Select the dimensions of performance that are valuable for your agency.**
   - This will feel like a philosophical or theoretical exercise.
   - The search for specific performance measures should be a liberating, unconstrained process in which participants are encouraged to think well outside of the traditional boundaries.
   - Do not focus yet on whether you can measure these items. That comes later.
   - What does your community want from its police?
   - Determine the relative importance of your dimensions: Are some more important than others?

4. **Figure out how to measure those dimensions of performance.**
   - Think broadly about potential data sources. Some will be contained in agency data, and some will need to be collected using surveys or other methods.
   - Some alternative methods include general community surveys, contact surveys, employee surveys, direct observation, and independent testing or simulation studies.
   - You may not be able to measure all of the important dimensions you've identified.
   - Do not reverse steps 3 and 4. Step 3 comes before step 4 for an important reason.

5. **Use the measures to improve your organization.**
   - All organizations are capable of self-learning, adaptation, adjustment, experimentation, and innovation. To do so, organizations need information and feedback.
   - Comparative performance measures will provide police organizations with crucial information: how they are doing relative to other police agencies on a variety of performance dimensions and how they are improving relative to their own previous levels of performance.
   - Use them and act on them. Don't just use them as a public relations gimmick for a news article or an annual report.
   - Treat the process as an integral step in organizational learning. Take your organization's temperature. Take its blood pressure. Then, use those measurements to form a diagnosis and implement organizational change.

6. **Repeat the process routinely.**

Conclusion

There are many ways to change organizations, from improved recruitment, hiring and training, to the selection of a new leader. This article presents just one potential method for improving police organizations: comparative performance measurement. All organizations are capable of self-learning, adaptation, adjustment, experimentation, and innovation. To undergo these processes, however, organizations need information and feedback. This article presents a systematic framework for improving policing by creating comparative performance measures. Such measures will provide police organizations with crucial information: how they are doing relative to other police agencies on a variety of performance dimensions or how they are improving relative to their own previous levels of performance. Performance measures are an essential component of an ongoing "organizational learning" strategy.

Endnotes

[78] Weisel, 1999
[79] Hickman & Reaves, 2001
[80] Gormley & Weimer, 1999
[81] Gallagher, Maguire, Mastrofski, & Reisig, 2001
[82] Rape victims, 1998
[83] Tyler, 1990
[84] Arrestee surveys are common in the study of drug use but rare in the study of police performance. In the United States, the Arrestee Drug Abuse Monitoring (ADAM) project collects urine samples from arrestees in multiple cities. Local ADAM sites sometimes conduct arrestee surveys that focus on issues other than drug use, but these are isolated efforts. The ADAM program represents an ideal framework on which to build a national data collection effort on police performance as viewed through the eyes of the arrestee population. Similar programs exist in Australia, South Africa, and other nations, suggesting the possibility of collecting such measures internationally.
[85] Camp, Saylor, & Harer, 1997
[86] Klockars, Ivkovich, Harver, & Haberfeld, 2000
[87] Harms, Clifford, & Cryer, 1998

[88] Fund for the City of New York, 2001
[89] Sampson & Raudenbush, 1999
[90] Mastrofski et al., 1998
[91] ABC News, 2001
[92] Bratton, 1998, p. 291
[93] Gormley & Weimer, 1999, p. 205
[94] Moore, Spelman, & Young, 1992
[95] Smith, 1923, p. 267
[96] Garreau, 1991
[97] www.cjgsu.net/initiatives/HomRates-PR-2003-08-03.htm
[98] Sherman, 1998
[99] Hoffman, 1971

References

ABC News. (2001, May 17). Testing police honesty: A *PrimeTime* investigation with lost wallets. *PrimeTime Thursday* [Television broadcast].

Bratton, W. (with Knobler, P.). (1998). *Turnaround: How America's top cop reversed the crime epidemic*. New York: Random House.

Camp, S. D., Saylor, W. G., & Harer, M. D. (1997). Aggregating individual-level evaluations of the organizational social climate: A multilevel investigation of the work environment at the Federal Bureau of Prisons. *Justice Quarterly*, *14*(4), 739-761.

Fund for the City of New York. (2001). *How smooth are New York City's streets?* New York: Center on Municipal Government Performance.

Gallagher, C., Maguire, E. R., Mastrofski, S. D., & Reisig, M. D. (2001). *The public image of the police: Final report to the International Association of Chiefs of Police*. Manassas, VA: George Mason University, Administration of Justice Program.

Garreau, J. (1991). *Edge city: Life on the new frontier*. New York: Doubleday.

Gormley, W. T., Jr., & Weimer, D. L. (1999). *Organizational report cards*. Cambridge, MA: Harvard University Press.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environment rating scale*. New York: Teachers College Press.

Hickman, M. J., & Reaves, B. A. (2001). *Community policing in local police departments, 1997 and 1999*. Washington, DC: Bureau of Justice Statistics.

Hoffman, R. B. (1971). Performance measurements in crime control. *Journal of Research in Crime & Delinquency*, *8*(2), 165-174.

Klockars, C. B., Ivkovich, S. K., Harver, W. E., & Haberfeld, M. R. (2000). *The measurement of police integrity. Research in Brief, NCJ-181465*. Washington, DC: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Mastrofski, S. D., Parks, R. B., Reiss, A. J., Jr., Worden, R. E., DeJong, C., Snipes, J. B., et al. (1998, December). *Systematic observation of public police: Applying field research methods to policy issues*. Washington, DC: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Moore, M. H., Spelman, W., & Young, R. (1992). *Innovations in policing: A test of three different methodologies for identifying important innovations in a substantive field*. Unpublished manuscript, Harvard University, Kennedy School of Government at Cambridge.

Rape victims rate police performance. (1998, July 23). *Toronto Star*, C3.

Sampson, R. J., & Raudenbush, S. (1999). Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. *American Journal of Sociology*, *105*(3), 603-651.

Sherman, L. W. (1998, July). *Evidence-based policing. Ideas in American Policing*. Washington, DC: U.S. Department of Justice, Police Foundation.

Smith, C. B., Jr. (1923). The adequacy of police forces. *Journal of the American Institute for Criminal Law and Criminology*, *13*, 266-271.

Tyler, T. R. (1990). *Why people obey the law*. New Haven, CT: Yale University Press.

Weisel, D. L. (1999). *Conducting community surveys: A practical guide for law enforcement agencies*. Washington, DC: Office of Community Oriented Policing Services and Bureau of Justice Statistics.

Edward R. Maguire, PhD, is associate professor of Administration of Justice at George Mason University. He received his PhD in criminal justice from the State University of New York at Albany. He has held previous academic and research positions at the University of Nebraska, the U.S. Department of Justice, and the United Nations. His primary professional interest is using organizational theory to understand and improve police agencies. He has led several national studies on police organization and innovation using a variety of research methods. His current research explores the intersection of comparative performance measurement and organizational learning in police agencies.