Despite the wide-ranging use of performance measures within the public sector and the growing use of performance bonuses to recognize performance achievements, the use of formal performance standards adjustment procedures in public performance measurement systems is relatively rare. Burt S. Barnow of Johns Hopkins University and Carolyn J. Heinrich of the University of Wisconsin–Madison set forth the basic arguments in favor of and against the use of formal or informal procedures for adjusting performance expectations. They describe how performance standards adjustments processes are currently (or have been) in use, review the evidence of their effectiveness or problems encountered in their application, and explore the consequences of the inadequacy of performance standards adjustments. The authors offer recommendations for the improvement of public sector performance measurement systems and conclude why this area remains fruitful for future research experimentation.

The nice thing about standards is that there are so many of them to choose from.

—Andrew S. Tannenbaum

In the last two decades, the use of performance measurement to improve government performance and hold public managers accountable for program outcomes has increased substantially in scope, complexity, and external visibility at the federal, state, and local levels. Some scholars have linked the growing demand for outcomes-based performance measurement to a resurgence of scientific management principles in government reforms and to the corresponding perspective that performance measurement will generate scientifically derived answers to questions about public program effectiveness (Bertelli and Lynn 2006; Frederickson and Frederickson 2006). Other scholars and practitioners view performance measurement as a strategy for eradicating perverse incentives in government and better aligning the interests of government employees with those of the public (Klitgaard and Light 2005). More cynically, still others, such as David Boyle, argue that “counting and measuring are seen as the antidote to distrust” (2001, 38), in the sense that performance measurement gives us (false) confidence about our control over outcomes and the objectivity of government decisions that shape them.

Performance measurement systems take varying forms, including organizational report cards, balanced scorecards, benchmarking, program evaluations, social indicators, annual performance reports, and disclosure requirements, and answer to different internal and external audiences (Gormley 2004). Among the most influential recent developments in public sector performance measurement was the introduction of the U.S. Government Performance and Results Act of 1993, which “brought the full force of the performance measurement movement to the federal government” and invigorated research on the design, implementation, and consequences of performance measurement systems in public programs (Frederickson and Frederickson 2006, 37). With the growing availability of empirical data from performance measurement systems, researchers are going beyond questions of how to “manage for results” and exploring the implications of performance measurement for both individual behavioral responses and organizational outcomes.

It is not an objective of this article to present a thorough review of the performance measurement literature and its findings … but rather to focus attention on a particular feature of performance measurement systems: the use of performance standards adjustments to set expectations for performance results.
findings, which would be beyond the scope of any article, but rather to focus attention on a particular feature of performance measurement systems: the use of performance standards adjustments to set expectations for performance results. In general, the basic goal of any performance standards setting and adjustment process is to establish appropriate benchmark levels of performance to guide the efforts and responses of those managing and conducting the primary work of programs. Adjustments to performance standards may be particularly important in contexts in which standards are applied to a large number of entities (such as states, local units of government, and/or contractors), and performance across entities is ranked (or otherwise compared) to allocate rewards and/or mete out sanctions. As we discuss in this article, failing to take into account factors that influence performance outcomes but are outside program managers’ control can contribute to serious distortions and efficiency losses, particularly if those who face an unlevel “playing field” attempt to adjust for unreasonable expectations in undesirable ways.

Indeed, our research was largely motivated by our observation that despite the wide-ranging use of performance standards and measures in public programs and the growing use of performance bonuses to recognize high performance achievements, the inclusion of formal performance standards adjustment procedures in these performance measurement systems is still relatively rare. Even in high-profile government programs such as Temporary Assistance for Needy Families (TANF)—in which the U.S. Department of Health and Human Services awarded performance bonuses for job entry and retention, Medicaid and SCHIP (State Children’s Health Insurance Program) enrollments, food stamp and child care subsidy receipt, and the number of children residing in married family couple groups—no adjustments were made for the characteristics of adult participants, economic conditions, or other circumstances affecting families’ level of need or capacity for change (Wiseman 2004). In other programs, special requests have to be made for performance standards adjustments, as in the case of the Wisconsin Department of Workforce Development, which only considers adjustments to performance standards in its Wisconsin Works (W-2) welfare to work program for “unusual or nonrecurring events.” At the same time, there is some indication that awareness of how baseline economic and demographic characteristics affect performance is increasing at the federal level, as reflected in the call of the Child Support Performance and Incentive Act of 1998 for research and recommendations on how to adjust to performance standards in the child support enforcement program for these types of factors (Tapogna et al. 2002).

The goal of this article is to examine the design and current use of performance standards adjustments and to marshal the evidence and arguments for increasing or discontinuing their use and for improving their functioning and effectiveness. We begin with a brief description of generic approaches to adjusting performance standards, followed by a discussion of some basic arguments that have been set forth both in favor and against the use of formal or informal procedures for adjusting performance expectations. We describe some of the performance standards adjustments processes that are currently (or have been) in use, giving consideration to their stated objectives, how the approaches were developed, and any evidence of their effectiveness or problems encountered in their use. We also present examples of cases or circumstances in which the absence (or inadequacy) of performance standards adjustments may be problematic, and we explore the consequences of the lack of adjustments for organizational incentives and achievements. We conclude with recommendations for policy makers and program managers aiming to improve public sector performance measurement systems.

The Basics of Performance Measurement and Adjustment

Designing a performance measurement system typically involves the following basic steps. First, one establishes a consensus on specific measurable program goals. Second, one defines empirical measures to use in quantitatively assessing performance toward those goals. Third, and of primary interest in this paper, most programs or organizations also set expectations for progress toward performance goals, that is, targets for performance improvements to be achieved in a given time frame. In many public sector performance measurement systems, these targets are annual, and increasingly, they also incorporate expectations for “continuous performance improvements,” a “total quality management” principle (Deming 1986). Yet as discussed earlier, few public programs opt to undertake the final step of developing formal procedures to regularly adjust performance expectations for unanticipated or uncontrollable factors that might thwart progress toward the goals. In this regard, they neglect a corresponding tenet of total quality management that advocates the use of statistical analysis to adjust for factors outside managers’ control in evaluating and managing performance (Deming 1986).

Although we briefly discuss some alternatives to statistically adjusting performance expectations in this article, in light of the limitations associated with informal procedures and the lack of transparency with their use in practice, we focus primarily on formal statistical methods for performance standards adjustments. One such technique involves adjusting a common performance standard or target ($P_t$) to which an individual or organization’s measured performance ($P_m$) is compared. In public training programs, for example, federal and state officials established target performance goals (e.g., a minimum entered employment rate) for local agencies by using a regression model to adjust for client demographic characteristics ($X$) and other uncontrollable factors ($Z$) that may influence performance (e.g., local area unemployment rates). Typically, baseline data and/or data on past performance ($P_0$) and the vectors $X$ and $Z$ of factors influencing performance are pooled across units and used to estimate a model, such as: $P_m = \alpha + \beta_1 X + \beta_2 Z + \epsilon$. The estimates of $\beta_1$ and $\beta_2$ (vectors) are subsequently used as weights for the influence of these factors in adjusting the common standard ($P_t$) to derive unit-specific performance targets (e.g., for an agency) for a given performance measure. Performance is then judged not by comparing actual performance ($P_m$) across units and/or time, but by comparing the differential between a unit’s target ($P_t$) and its measured performance ($P_m$).

A slight variant of the foregoing approach to adjusting performance is described by Stiefel, Rubenstein, and Schwartz (1999) and illustrated with an application to public elementary school data from Chicago. In the regression equation for this approach, the dependent variable is measured performance ($P_m$), and the independent variables are measures of uncontrollable factors ($Z$) and/or other factors ($X$) that may be controllable under some circumstances (or at some level) and influence $P_m$ in that period. The resulting
adjusted measure of performance for a given unit (e.g., school) is the difference between the predicted value of performance generated by the model for this unit and the unit’s actual value \((\text{Pm}_i\text{)}\). In their analysis, for example, the percentage of students from low-income families had a statistically significant, negative effect on student test scores, so that the predicted values of performance in schools with more low-income students (holding all else constant) were lower. In both of these approaches, if the difference between actual performance and the predicted performance (or the adjusted performance target) is positive, this indicates that the unit or organization is exceeding its performance expectations.

Why Adjust Performance Standards?

We suggest that arguments advanced both for and against the use of performance standards adjustments should simultaneously consider alternative aims of public sector performance measurement systems: to produce knowledge of public program impacts, to shape and manage incentives for individual and/or organizational behavior, and to promote transparency and accountability to the public of government activities and their outcomes. Of course, many government organizations set forth all three of these objectives for their performance measurement systems, yet we contend that the pros and cons of adjusting performance will vary according to the relative weight or importance assigned to them.

The objective of producing accurate knowledge of program impacts or the value added of government activities has recently been strongly advocated by numerous coalitions organized to promote “evidence-based policy making,” that is, government policies and practices based on or guided by scientifically rigorous evidence of their effectiveness. If accepted as the principal objective of a performance measurement system, it is an imperative (in the absence of an experimental evaluation) to model statistically the relationship of government activities (i.e., the technology of public production) to performance outcomes, while adjusting for factors that influence outcomes but are not (or should not be) controlled by public managers. In effect, by adjusting performance expectations for factors that are not controlled in production, the estimates of performance are more likely to accurately (and usefully) reflect the contribution (or value added) of public managers and program activities to any changes in performance.

Bartik, Eberts, and Kline offer exactly this argument in their development of an adjustment model for workforce program performance standards that they suggest should “lead to a better measure of value-added” (2004, 4) of local workforce areas (LWAs), and in turn, “better promote higher-value-added among LWAs by better identifying high-value-added LWAs that should be rewarded and emulated, and low-value-added LWAs that should be reformed.” In their performance adjustment model, Bartik et al. estimate a model of measured performance \((\text{Pm}_i\text{)}\) using cross-sectional data on individual participants (indexed by \(i\)) in programs offered by different LWAs (indexed by \(j\)). In addition to controlling for \(X\) and \(Z\) (vectors of factors that are always or sometimes not controlled by program managers), Bartik et al. add a set of indicator variables, \(W_j\) for each LWA \(j\), to the equation. Importantly, they argue that if it were possible to measure all relevant \(X\) and \(Z\) factors that affect individual performance outcomes, the estimated \(W\) would be an unbiased and consistent estimate of the relative value added of the LWA (i.e., compared to other LWAs) on a given measure. The caveat, of course, is that if there are omitted individual or local area characteristics (i.e., not in \(X\) or \(Z\) ) that differ across LWAs, these estimates of \(W_j\) are likely to be biased. Bartik, Eberts, and Kline contend that even in the presence of some bias, “the relevant issue is whether these estimates are still closer to the true relative value added than the estimates one would obtain by simply comparing LWA means on the common measures” (2004, 8).

In a Monte Carlo data analysis, Brooks (2000) considers whether adjusted performance measures are more likely to approximate true performance, particularly in the case in which omitted or unobservable factors are related to the measured characteristics included in the adjustment model (and to outcomes). Not surprisingly, he finds that the higher the correlation between observed and unobserved factors, the greater the bias in the adjusted performance estimates. In addition, if the adjustment model specified explains little of the variance in measured performance, it will produce unreliable estimates of performance even if the correlation between observed and unobserved factors is low. Estimates of performance may be “extremely misleading” (2000, 3232), Brooks cautions, unless the specification is very accurate and the data for estimating it are complete. Although we concur with Brooks’s technical points, we think the question of interest to policy makers is not whether allowing for adjustments might distort performance rankings, but rather whether any such distortions are more likely to occur than the benefits from statistical adjustments, including the perception of fairness that results from accounting for factors that affect performance but are not controllable by program managers. As we note later, statistical adjustments are not perfect, and provision should be made to overrule the results when they lead to obvious distortions or otherwise defy common sense.

As there are yet no definitive statistical methods for assessing the extent and nature of omitted variable bias and the accuracy of a performance adjustment model, the use of regression models for adjusting performance expectations will always be subject to such criticisms. Indeed, the design of public sector performance measurement systems that precisely measure value added is still primarily an ideal that scholars and policy makers are working toward. There are, however, concrete examples of progress being made, such as the exploratory use of value-added measures of student achievement in public schools (Thorn and Meyer 2006). In the value-added approach to performance measurement developed by education researchers, econometric methods are used to adjust for student characteristics and measurement error in student achievement, and individual growth in student achievement is modeled over time (using a multilevel specification) in order to better assess the contributions of schools to student outcomes (see, e.g., Roderick,
Organizational report cards, which provide ratings of how well organizations perform on one or more criteria, have similar rationales for adjustments. Hospitals, for example, can be rated on the outcomes of particular health care treatments and how well they improve the health of their clients. Because these organizations are likely to face quite different client mixes, a simple comparison of outcomes may not be an appropriate way to assess performance. In describing why the state preferred to compare health outcomes for adult cardiac surgery across hospitals using risk-adjusted mortality rates, the New York State Department of Health (2004) stated, “It is difficult, however, to compare outcomes across hospitals when assessing provider performance, because different hospitals treat different types of patients. Hospitals with sicker patients may have higher rates of complications and death than other hospitals in the state.”

Although promising, some scholars have criticized these more technically rigorous approaches to performance measurement as too positivist or elitist, arguing that they may put performance analysis and results out of reach of practitioners and the public (i.e., limiting their transparency and usefulness for accountability purposes) (Shulock 1999). In the Workforce Investment Act (WIA) program, the U.S. Department of Labor discontinued use of the regression-based performance standards adjustment procedures and replaced them with a system of negotiated standards with the goal of promoting “shared accountability” (U.S. Department of Labor, Employment and Training Administration 2001). It was suggested that factors outside managers’ control that affected program outcomes (some unmeasurable) might be more easily conveyed and weighed in negotiations.

In other cases, the use of performance standards adjustments may be viewed as incompatible with the objective of motivating particular individual or organizational responses to performance requirements. For example, some performance measurement system designers intentionally choose to set a more ambitious standard—also known as a “stretch target”—which is not adjusted in order to motivate laggards to change their ways and aspire to achieve higher standards of performance. In the TANF high performance bonus system, states that in the past had invested little to help clients achieve self-sufficiency had to work harder to meet performance requirements for client work participation, job entry, retention, and earnings gains. A related argument for not developing or using performance standard adjustments is to promote equity in outcomes, that is, to hold up the same standard for all individuals or organizations, despite the greater challenges that may be involved in achieving the minimum level of performance for some. The recent education reforms that require states to set standards for reading and mathematics proficiency and to ensure that all children achieve these minimum levels within a specified time frame, regardless of their backgrounds, special needs, school resources, and so on, are an example of this approach. In fact, the U.S. Department of Education describes its requirements for “challenging state standards” and student testing as one of the “pillars” of the No Child Left Behind (NCLB) Act that is intended to strengthen “accountability for results” (Radin 2006, 62–63). In addition to requiring states to measure and report “adequate yearly progress” under strict timelines and in compliance with federal guidelines, NCLB also established a uniform target for schools to have 100 percent of their students proficient in reading and mathematics within 12 years. Some states have responded accordingly by developing their own performance-based funding incentive systems, in which school districts, schools, and even principals receive incentive payments for schools or students who “demonstrate progress” or exceed performance expectations. Texas and California, for example, are funding incentive awards at approximately half a billion dollars per year, and a number of states (including Arizona, Colorado, Florida, North Carolina, Ohio, and Tennessee) are using value-added statistical models to measure teachers’ contributions to learning and to give teachers credit based on how much better (than expected) their students perform on tests compared to peers (House Research Organization, Texas House of Representatives, 2004).

The responses of schools to these features of NCLB and related education reforms also suggest, however, that all may not be having the intended effects on educational approaches and activities, and in some cases, they may instead be undermining efforts to produce useful information for improving educational outcomes. Faced with limited administrative capacity and technical expertise and state budget shortfalls, some states and schools attempted to “game” what was viewed as an unfair system, or to otherwise misrepresent their progress toward performance targets. For example, Radin (2006) uncovered intentional underreporting of high school dropout numbers by states. In a study of Chicago Public Schools’ test-based accountability system, Jacob (2005) detected sharp gains in students’ reading and math test scores that were not predicted by prior trajectories of achievement but were consistent with the incentives established by the system. In addition, he found some evidence of shifting resources across subjects and increases in student special education and retention rates, particularly in low-achieving schools. Finally, Jacob and Levitt (2003) produced empirical evidence of outright cheating on the part of teachers, who systematically altered students’ test forms to increase classroom test performance.

Would the use of formal performance standards adjustment procedures to “level the playing field” for schools struggling with substantial barriers under NCLB have averted these dysfunctional responses to the...
performance requirements? Although this is a question we are not able to answer empirically in this research, experience with performance standards adjustments in other public programs suggests that they would have likely reduced these problems, even if they were not able to circumvent them altogether. We now discuss some of the performance standards adjustment processes that are currently (or have been) in use and review evidence of their role and potential for producing more useful information on program effects, constructing more effective incentives for public managers, and generating more fair and meaningful performance reports for accountability to the public.

Performance Standard Adjustments in Practice
To identify public programs that use (or have used) adjustments to their performance standards, we conducted an extensive literature search, contacted performance management scholars to ascertain their awareness of the use of performance adjustments in government programs, and carried out an e-mail survey of state workforce agencies (which have the longest experience with performance adjustments). We identified one national program that currently uses formal statistical adjustments (the Job Corps) and one former program (the Job Training Partnership Act) that employed a performance standard adjustment system that evolved substantially over time. A few public programs are exploring options for using performance standard adjustments (such as state programs under WIA), and others, such as the state and local educational agencies discussed earlier, incorporate them into occasional performance evaluations, although their use is not regular or systematic in most cases. This list would be longer if we included organizational report cards that are increasingly used in health and education settings to help consumers select providers, and thus, we discuss the case of cardiac surgery report cards in which logistic regression models are used to produce risk-adjusted mortality rates. We suggest that the lessons from these few programs apply more broadly to other types of programs.

The Job Training Partnership Act Program and Its Performance Management System
Performance management for workforce investment programs began on an exploratory basis in the later years of the Comprehensive Employment and Training Act (CETA), the nation’s major workforce program from 1973 through 1982.5 Economists working for the assistant secretary for policy, evaluation, and research in the late 1970s advanced the idea of holding local CETA programs accountable for their performance as measured by the impact of the programs on earnings and employment, but also recognized the unfairness of setting the same earnings or employment rate standard for local programs that served very different populations in varying local economic conditions. This laid the groundwork for their formalization under the Job Training Partnership Act (JTPA), the nation’s primary workforce development program from 1982 through July 2000.5 The performance measures used for adults, youth, and dislocated workers, which evolved over time, are shown in table 1 for program years 1998 and 1999, the final years in which JTPA operated. The table also shows the adjustment factors and regression-determined adjustments for the adult follow-up employment rate. If, for example, a local program increased the proportion of adult terminees who were women by 10 percentage points, the program’s performance standard for the adult follow-up employ-

In establishing national performance measures and standards, the federal government used data on past experience to set targets that they expected approximately 75 percent of the local service delivery areas (SDAs) would be able to meet or exceed; that is, performance standards were set at the performance level of the SDA at the twenty-fifth percentile in the prior period. Governors were also empowered to select additional measures, set standards for acceptable performance for each measure, determine how performance on the individual standards was to be combined to determine overall performance, adjust standards for the state’s SDAs to reflect differences in participant characteristics and local economic conditions, and determine sanctions and awards, subject to federal requirements. Although the particular variables and adjustment parameters varied from year to year, the basic approach remained the same; the bottom half of table 1 shows the variables in the program year 1998–99 models for the adult, youth, and dislocated worker programs. Local programs that failed to meet half or more of the core JTPA standards were ineligible to receive performance awards, and if an SDA failed to meet half or more of the standards in two consecutive years, the governor was required to implement a reorganization plan for the SDA. Thus, it is important to note that the JTPA performance management system had stronger reward and sanction provisions than were later required by the Government Performance and Results Act—organizations that did well could obtain substantial additional resources, and those that did poorly could lose their right to operate the program.

Although, at the onset, many states simply used the secretary of labor’s standards without adjustments, by the early 1990s, a majority recognized that failure to adjust the standards for local economic conditions and participant characteristics was generating incentives to enroll individuals who would do best in the labor market regardless of the impact of the program (referred to as “cream skimming” or creaming). As a result, more governors opted to use the secretary’s adjustment model, and, following the 1992 JTPA amendments, governors were required to adjust performance standards, either using the optional Department of Labor regression models or some alternative approach that was approved by the Department of Labor (Barnow 1992).7 Yet despite the Labor Department’s efforts to “level the playing field” and reduce creaming, there is still ample evidence of gaming of the performance management system by local programs to improve their measured performance, through actions such as the strategic enrollment and timing of clients’ entry and exit from the program (Courty and Marschke 1996, 2004). A review of research in this area (Barnow and Smith 2004) concluded that the JTPA system was highly susceptible to manipulation by local program operators. At the same time, recent research contrasting the JTPA and WIA systems suggests that the regression model adjustments likely tempered such problems (Heinrich 2004); without them, both incentives and means for local programs to engage in cream skimming behavior are greater.

This research also suggests some aspects of the JTPA performance standards adjustment procedure that one might question.8 For example, in using data from the prior program year to develop regression adjustment models, only statistically significant variables of the
expected sign and a reasonable magnitude were retained in the model. Thus, if the coefficient for the percentage of participants who were black in the entered employment rate model was either positive (counterintuitively) or not statistically significant, that variable would be excluded from the model for that year. Although Barnow (1996) showed inconsistent treatment of people with disabilities because of this feature, changes in the adjustment procedures did not follow. And in other ways, the implementation of the performance standards adjustments may have been too rigid. If a governor decided to use the model, he or she could modify the target level of performance, but governors could not change the regression coefficients themselves (e.g., to encourage enrollment of certain groups beyond simply holding local areas harmless for serving that group), nor could they add adjustments for characteristics not in the models (e.g., to give credit for serving refugees or the disabled in years when such characteristics were not included in the model).

Overall, despite the deficiencies, the use of the regression adjustment models was largely accepted by the various parties and appears to have contributed to leveling the playing field and to producing more valuable information for policy makers to use both for program management and accountability purposes.
have contributed to leveling the playing field and to producing more valuable information for policy makers to use both for program management and accountability purposes. By the time the JTPA program ended, the national regression-based adjustments were the default case and were generally perceived as a fair and appropriate way to set performance standards.

The Job Corps Performance Management System and Adjustments Processes

The Job Corps is an education and vocational training program administered by the Department of Labor that helps young people ages 16 through 24 learn a trade, earn a high school diploma or GED, and get help finding a good or better job. In 2007, there were 94 Job Corps centers operated by nonprofit and for-profit vendors under contract with the Department of Labor, four satellite centers, and 28 Civilian Conservation Centers managed by the U.S. Departments of Agriculture and Interior.

The Job Corps measures center performance using 11 performance measures that are weighted and aggregated to provide an overall report card grade (U.S. Department of Labor, Office of Job Corps 2001). Five of the 11 performance measures, constituting 37.5 percent of the total score for a center, use regression models to adjust the performance standards; the other six measures are not adjusted. The adjustment factors for the program year 2007 graduate six-month weekly earnings model include the percentage with a high school diploma, reading and math competency, field of occupational training, average wage in the county, the percentage of families in the county in poverty, and the percentage placed in a job in a state with a high minimum wage. The rationale offered for adjusting some of the standards is the same as that of JTPA: “By setting individualized goals that adjust for differences in key factors that are beyond the operator’s control, this helps to ‘level the playing field’ in assessing performance” (U.S. Department of Labor, Office of Job Corps 2001, 6, appendix 501). Models are updated annually, and the Job Corps also periodically changes the weights assigned to the adjustment factors in computing the overall score.

In practice, the Job Corps adjustment models often lead to large differences in performance standards across the centers. In 2007, for example, the diploma/GED attainment rate standard ranged from 40.5 percent to 60.8 percent, and the average graduate hourly wage standard ranged from $5.83 per hour to $10.19 per hour. Discussions with Job Corps staff suggest that center operators believe that the performance standards adjustments are fair and help to account for varying circumstances across centers. Given this positive view, it is puzzling that adjustments are not used for all 11 Job Corps performance measures.

Cardiac Surgery in New York State

Several states, including New York and Pennsylvania, have developed highly regarded organizational report cards for health care procedures that are intended to assist consumers and referring physicians in selecting the most appropriate provider for high-risk procedures (Gormley 2004). The New York State report card for coronary artery bypass graft surgery (CABG) is discussed here; similar procedures are used for cardiac valve procedures. CABG is a procedure in which a vein or artery from another part of the body is used to replace a defective cardiac artery and improve the supply of blood to the heart. In 2000, 421 patients of the 18,121 receiving the procedure died, for a statewide mortality rate of 2.32 percent. The mortality rate varied among the 34 hospitals that performed the procedure, from 0 percent to 5.00 percent.

To account for the fact that hospitals providing CABG vary in the characteristics of the patients they serve, a logistic regression model was developed to adjust the expected mortality rate for five categories of patient characteristics, including demographics (age and gender), hemodynamic state (unstable, shock, and cardiopulmonary resuscitation), ventricular function (three variables measuring the functioning of the heart); comorbidities (seven variables measuring the presence or absence of other diseases), and whether the patient had had previous open heart operations. The coefficients estimated from the logit equation were then used to develop risk-adjusted mortality rates for patients and the hospitals in the sample. As would be expected, after adjustments, the risk-adjusted mortality rates differed from the observed mortality rates, although the range did not change substantially. The report notes that only three hospitals have risk-adjusted mortality rates that exceed the 95 percent confidence interval for the state average, and two hospitals have risk-adjusted mortality rates that are lower than the 95 percent confidence interval for the state average. Similar procedures were used to estimate risk-adjusted mortality rates for surgeons performing CABG procedures.

In studying how the New York CABG report card predicted performance and influenced consumer choice, Jha and Epstein (2006) concluded that the report cards did well in predicting the risk-adjusted mortality rates of hospitals and surgeons. In terms of the impact on the market, however, they reported mixed findings. At the hospital level, the report cards appeared to have no impact on market share, although surgeons with high risk-adjusted mortality rates were more likely to retire or leave CABG practice after receiving a high mortality rate grade.

In general, the examples in this section suggest that the use of performance standards adjustments is largely perceived by those engaged in performance management as generating fairer expectations and more appropriate incentives for improving performance, as well as more meaningful information on performance outcomes for program operators and consumers.

Where the Absence of Adjustments Is Problematic

As noted in the introduction, most government programs do not have adjustment mechanisms in place for their performance standards. We argue that the lack of adjustment mechanisms is likely to be particularly problematic for programs that involve multiple levels of government and numerous organizations. For example, if a national student loan program is setting performance measures only at the national level, there may be little need to have a formal adjustment procedure, as any shortfalls in performance can be addressed on an ad hoc basis in the agency’s annual performance report. On the other hand, when the federal government rates and ranks states or other subnational units on performance and recognizes or rewards (or sanctions) performance accordingly, failure to take into account relevant factors that are outside program managers’ control can contribute to serious problems and unintended consequences.
The child support enforcement program established under Title IV-D of the Social Security Act is one example in which the absence of adjustments affects performance rankings. In 1998, Congress enacted the Child Support Performance and Incentive Act (CSPIA) with the goal of altering the incentive structure and rewarding states for performance on establishment and enforcement practices. Specifically, Congress linked incentive payments to states’ performance in five areas: (1) paternity establishment (statewide or specific to the IV-D caseload), (2) establishment of child support orders, (3) collections on current support due, (4) cases with collections on arrears (past support due), and (5) cost-effectiveness (i.e., total collections divided by total administrative costs) (Tapogna et al. 2002). The CSPIA legislation mandated a study of the economic and demographic characteristics of states and their influence on performance and called on the secretary of the Department of Health and Human Services to recommend adjustments for measuring the relative performance of the states. The required study (see Tapogna et al. 2002) identified, through a literature review and by interviewing child support enforcement officials, researchers, and interest groups familiar with the issues, 55 variables likely to influence performance outcomes. The researchers estimated models using two years of available data, and the final models included 13 measures of demographic, economic, and programmatic features of the states.13

For all measures except paternity establishment rates, the researchers concluded that using regression adjustment models was feasible and would increase equity in the system. In most cases, adjustments for a majority of states were relatively small, but in a few instances, the adjustment models led to major changes in the ranking of a state. For example, for the analysis of the percentage of cases with orders, 30 states had adjustments of less than 5 percentage points; for a few states, however, the adjustment was quite large and would have led to a sizable change in the state’s relative and absolute performance.14 The report recommended that prior to adopting the adjustment procedures, further research should be conducted to allow more years of experience and more variation in economic conditions to be captured in the models.

Despite the study findings, states did not support the use of performance adjustments in the Child Support Enforcement (CSE) performance management system. When asked whether the new system favors states with certain profiles, and if so, whether states should be compensated for factors beyond their control, most respondents agreed that CSE programs operate in different socioeconomic environments. However, the majority opposed adjusting incentive payments to account for those differences, because of the complexity and uncertainty they would add to an already complicated payment system, and the difficulty of establishing a consensus on appropriate adjustment factors (Gardiner et al. 2004). The CSE experience confirms that the use of adjustments can be controversial, particularly if public officials perceive them as too complex, lacking transparency, and/or having unclear implications, which may explain in part why so few public programs adopt them.

In public programs in which performance bonuses are used to reward or sanction states or other government entities for performance results, there may be additional reasons, however, to be concerned about the decision to forego performance standards adjustments. Consider, for example, the absence of adjustments in measuring states’ reductions in out-of-wedlock birth rates following the 1996 Personal Responsibility and Work Opportunities Reconciliation Act. To provide additional incentives to states for promoting parental responsibility and encouraging two-parent families, an annual performance bonus system was designed to award $20 million each to as many as five states with the largest reductions in the proportion of out-of-wedlock births to total births. In determining the award winners each year, the Department of Health and Human Services compiled statistics (reported by states) and compared the proportion of out-of-wedlock births to total births for the most recent two-year period to that for the preceding two-year period. For example, in awarding the 1999 bonuses, rankings were based on birth statistics from 1995 and 1996 that were compared to 1997 and 1998. The top five states (as ranked by this measure) also had to show a decrease in their abortion rate between the most recent year and 1995, where the abortion rate is measured as the number of abortions divided by the number of births.

In announcing the 1999 bonus awardees (District of Columbia, 4.13 percent; Arizona, 1.38 percent; Michigan 1.34 percent; Alabama, 0.29 percent; and Illinois, 0.02 percent), the U.S. Department of Health and Human Services (2000) acknowledged that “more evidence is still needed to fully understand the range of factors contributing to the decrease in the proportion of out-of-wedlock births in these particular states.” They also recognized that even before enactment of the 1996 law, teen birth rates had been declining, reaching their lowest rate in the 60 years of recording. An analysis by Paul Offner of the Brookings Institution was more pointed:

When the first year’s results were announced in 1999, the winners were the District of Columbia, California, Michigan, Alabama, and Massachusetts—all jurisdictions with large African American and Hispanic populations (D.C., Michigan, and Alabama also won bonuses in the second round). Between 1994–95 and 1996–97, black and Hispanic non-marital birth rates dropped twice as fast as white non-marital birth rates. This suggests that demographic factors may have been as important as any actions taken by the states in determining the bonus winners. (2001, 4)

In addition, state reports of performance were not audited (Wiseman 2004), and little attention was given to the possibility of measurement or reporting errors. Furthermore, there was no standard or “bar” set for some minimal reduction in out-of-wedlock births that would be worthy of a bonus. In this regard, one might question whether $20 million was an appropriate award for the 0.02 percent reduction achieved by Illinois in the first period. Perhaps most importantly, awarding performance bonuses based on changes in population statistics that may have little to do with any policy actions taken by states—and for which there are no formal procedures for evaluating states’ contributions—would appear to undermine the basic goals of performance measurement systems, that is, promoting accountability for outcomes, improving incentives for performance, and better understanding the outcomes and effectiveness of policy actions. In the most recent welfare reform law reauthorization, the TANF high performance bonus system was discontinued.

Concerns about the distribution of performance bonuses to states also emerged in the Workforce Investment Act program (the
successor to the JTPA), in part because of the discontinuation of the use of statistical models for performance standards adjustments in the performance management system. With the goal of promoting “shared accountability,” any adjustments to the performance standards are now made by the federal government through a negotiation process with each state, and states may adjust local standards as they see appropriate (U.S. Department of Labor, Employment and Training Administration 2001). States are also permitted to request a change in the negotiated standards under specific conditions: “This policy allows adjustments to negotiated performance goals in order to account for changes in economic conditions, changes in the characteristics of the participants served by the program, and changes in service delivery design” (U.S. Department of Labor, Employment and Training Administration 2002). Thus, in this new method for setting performance standards, there is no longer a common adjustment model that is used by the states to account for factors outside program managers’ control, and it is incumbent on the parties involved in these negotiations to adjust for economic and population characteristics in setting performance expectations.

Two national studies of the WIA reported concerns about these new adjustment procedures implemented by the Department of Labor. One study concluded,

Although one State noted that it perceived the process as fair, most felt there was very little real negotiation between themselves and DOL. A common perception was that DOL handed down the performance goals, or at most allowed only minor adjustments based on the state’s proposals or arguments that its goals should be lowered due to prevailing economic conditions. (Social Policy Research Associates 2004, V-3)

In a second study (Barnow and King 2005), most officials interviewed decried the absence of formal procedures to adjust for characteristics of participants served and local economic conditions and expressed concern that the Labor Department’s Employment and Training Administration regional office officials did not enter into negotiations with state officials but rather pressured them to meet federal standards. Both studies concluded that abandonment of the regression model adjustment process and failure to negotiate seriously led to standards often considered arbitrary and to incentives for creaming among potential participants. In a related empirical study of the WIA performance management system, Heinrich (2004) concluded that in the absence of regular adjustments to performance standards for changing local conditions, the WIA system appeared to promote increased risk for program managers rather than shared accountability. Program managers, in response, appeared to make undesirable post hoc accommodations to improve measured performance.

Although the Department of Labor plans no significant changes in how performance standards are negotiated (U.S. Department of Labor, Employment and Training Administration 2007), there have been some efforts by states to take into account varying circumstances among their local workforce areas and to include adjustment procedures. A few states, such as Texas, that had a large enough number of local workforce areas developed their own regression models. In Maryland, local workforce areas were divided into high, medium, and low performers on each performance measure, and the better-performing local areas were expected to achieve higher performance (Governor’s Workforce Investment Board 2001). The Department of Labor also awarded grants to two states, Michigan and Washington, to develop regression-based adjustment models for the WIA performance standards. Michigan is using models to adjust the “common measures” promulgated by the U.S. Office of Management and Budget, while the Washington models are used to develop adjustments for the current WIA performance measures. The Michigan models are intended to be illustrative only, however, and are not used in setting standards, while the Washington models generate a starting point for negotiations of both state and local performance standards.

In 2007, the Department of Labor provided states with data on how economic and participant characteristics affect performance, in effect inviting states to negotiate based on these factors. Thus, although workforce development programs are still not subject to automatic regression-based adjustments, as used in the JTPA, recent changes suggest more support for efforts to take into account factors that are recognized as outside managers’ control in managing performance. Overall, although program managers prefer to have input into the processes that determine the models or approaches that will be used to set performance standards, the more systematic, regression-based approaches appear to produce standards that are more likely to effectively “level the playing field.” Past experience also shows that regression-based approaches can be developed collaboratively, and that there are some advantages to allowing for adjustments over time based on the analysis of multiple years of data, which contributes to stability in the process.

Conclusions and Recommendations

Although many public programs, including all federal agencies, are required to establish performance standards, our research finds few cases in which adjustments to performance standards have been considered, and even fewer in which they have actually been applied. The concepts of fairness and equity have been set forth to argue both for and against the use of performance adjustments. The most often cited reason for adjusting standards is to “level the playing field,” or to make performance management systems as fair as possible by establishing expectations that take into account different demographic, economic, and other conditions or circumstances outside of public managers’ control that influence performance. It has also been argued, however, that it is not acceptable to set lower expectations for some programs than others, even if they serve more disadvantaged populations or operate in more difficult circumstances. For example, do we perpetuate inequities in education if less rigorous standards for reading and math performance are established for schools serving poorer children? Or if a single standard is set for all, could governments instead direct more resources

---

Although many public programs, including all federal agencies, are required to establish performance standards, our research finds few cases in which adjustments to performance standards have been considered, and even fewer in which they have actually been applied.
to those programs that face more difficult conditions or disadvantaged populations to help put them on a more level playing field?

Another argument of those advocating performance adjustments is that they better approximate the value added by programs (rather than gross outcome levels or change). For policy makers or program managers, having a better understanding of the contributions of program activities to performance (net of factors that are not influenced by the production or service processes) may contribute to more effective use of the performance information to improve program operations and management. The use of adjusted performance measures is also more likely to discourage (if not eliminate) “gaming” responses, in which program managers attempt to influence measured performance in ways that do not increase impacts (e.g., by altering who is served and how). A system that adjusts for population characteristics and other such factors will reduce the efficacy of these gaming strategies and the misspent effort and resources associated with them.

Of course, these benefits may be contingent on program managers understanding and having confidence in the adjustment mechanisms. Regression-based performance adjustment models have been criticized for having low explanatory power (as measured by $R^2$) and flawed specifications, suggesting that sometimes adjustments may be biased or unreliable. The argument that a low $R^2$ implies that the statistical model is not useful is in most cases false. A low $R^2$ means that there is a lot of noise in predicting the overall level of the dependent variable, not necessarily that the results are unreliable. Indeed, one may obtain statistically significant coefficients for the adjustment factors even with a low $R^2$, implying that there are important factors that have a strong effect on predicted performance and should be accounted for in measuring performance. Based on similar reasoning, Rubenstein, Stiefel, and Schwartz (2003) argue that performance standard adjustments should also be attempted even when the number of organizations available for comparison is small.

While we recognize the merits in these arguments both for and against the use of performance adjustments, our primary concern is that so few public programs appear to even consider or attempt to develop adjustments for performance standards. Until more experimentation with performance adjustments takes place in public programs, we will continue to be limited in our ability to understand not only whether they have the potential to improve the accuracy of our performance assessments, but also if they contribute to improved performance over time as public managers receive more useful feedback about their programs’ achievements (or failures) and what contributes to them.

We thus conclude with the following recommendations. First, policy makers and program managers should, at a minimum, give more consideration to the concept of adjusting performance standards. Specifically, programs should ask whether they can make a strong case for having the same standard for all jurisdictions or entities regardless of the context or circumstances in which they operate. Second, statistical modeling should be viewed as one tool in the adjustment process (and not the only technique to be applied). There is no single approach to statistical modeling or to combining statistical analysis with other methods such as negotiation or subgroup performance analysis that will work best for all programs. In fact, we suggest that statistical modeling should be viewed as a complement rather than a substitute for negotiating performance standards. In Washington state, for example, statistical models are a starting point for negotiations of local WIA performance standards, and at the national level, the Department of Labor is now providing guidance on how changes in circumstances (like the unemployment rate) can affect outcomes. Likewise, if regression models produce counterintuitive findings or findings that are contrary to other policies of interest, the models, data, and time frame should be investigated and refined accordingly (or discarded). In fact, we suggest that the process of thinking through and estimating adjustment models may be of value itself for learning about how different factors (both within and beyond managers’ control) may affect outcomes of interest and how program management might correspondingly be improved. Finally, the use of statistical modeling for performance adjustments does not negate the use of other incentives for guiding program managers or the incorporation of other performance management system features or requirements such as “continuous performance improvement.” This is clearly a fertile area for additional experimentation with the design and implementation of performance management systems as well as for further academic research (and even better if the two go hand in hand).

Notes

1. Federal government agencies are required under the Government Performance and Results Act to establish performance goals, measures, and plans; to provide evidence of their performance relative to targets; and to report their results annually to the public.


4. For example, there is a Center for Evidence-Based Policy (at Oregon Health and Science University), a national Coalition for Evidence-Based Policy, the Cochrane Collaboration and the Evidence for Policy and Practice Information and Coordinating Centre (both established in the United Kingdom), evidence-based policy networks, evidence-based journals and journal clubs, and evidence-based policy making newsletters and bulletins that review and disseminate current research findings on the effectiveness of interventions.

5. We are grateful to Christopher King for discussions on the history of performance management under the CETA and JTPA.

6. States distributed 78 percent of the federal funds to approximately 600 local units of government and consortia of local units of government that were referred to as service delivery areas (SDAs) and were responsible for service provision. JTPA program activities, including vocational or basic skills classroom training, on-the-job training, job search assistance, and work experience, were sometimes provided by SDAs themselves but were more commonly delivered through contracts with community colleges and other nonprofit and for-profit training organizations.

7. In fact, the criteria for using adjustments other than the Department of Labor regression model were quite strict and discouraged states from developing their own adjustment procedures. The Department of Labor’s guide to performance standards (Social Policy Research Associates 1999) stated that the adjustment procedures had to meet the following criteria: (1) procedures for adjusting performance standards must be responsive to the act, consistently applied among the SDAs and substate areas, objective and equitable throughout the state, and in conformance with widely accepted statistical criteria; (2) source data must be
of public use quality and available upon request; (3) results must be documented and reproducible, and (4) adjustment factors must be limited to economic factors, labor market conditions, geographic factors, characteristics of the population to be served, demonstrated difficulties in serving the population, and type of services to be provided.

8. An issue we do not discuss here is whether it was correct to estimate the adjustment models at the local workforce area level rather than the individual level. There are arguments for and against both approaches, and sometimes grouping data can lead to large changes in the estimated relationships (Blalock 1961).

9. In addition to the two criteria noted in the text, five other criteria were also used to decide which variables were included in the adjustment models (Social Policy Research Associates 1999, III-8).

10. For example, data from a number of years could have been pooled, which would likely have led to fewer changes in the variables included in the models and to smaller changes in the magnitude of the adjustments. Another alternative would have been to limit the magnitude of the change in the adjustments so that the incentives would not vary so much from year to year.


12. These five performance measures are: high school diploma/GED attainment rate, average literacy gain, average numeracy gain, graduate average wage at placement, and graduate six-month average weekly earnings. The six measures for which no adjustments are made include: career technical training completion rate, career technical training placement rate, postenrollment placement rate, graduate placement rate, graduate six-month follow-up placement rate, and graduate 12-month follow-up placement rate.

13. These variables included personal income per capita, poverty rate, percentage of males aged 20–64 not working, rate of job growth, percentage of population living in urban areas, percentage of TANF case heads under age 30, percentage of Title IV-D cases currently participating in TANF, percentage of Title IV-D cases that have never participated in TANF, number of Title IV-D cases per full-time equivalent staff, Title IV-D expenditures per case, population stability (percent of people living in same house 1999 and 2000), judicial or administrative order establishment process, and audit pass/failure indicator. The programmatic variables were included to avoid bias and were not intended for use in adjusting performance standards. All of the variables had the expected sign in the models, and most were statistically significant at the .05 level. The explanatory power of the models was reasonable, with $R^2$ ranging from .3 to .7. Six states that failed all audits in the first year of the analysis were omitted from the study because their data was considered too unreliable.

14. The District of Columbia, which had characteristics associated with a low level of order establishment, had an unadjusted order establishment rate of 26 percent, but its adjusted rate was 48 percent. Other states with large adjustments include Maine, which had its cases with orders score reduced from 89 percent to 79 percent, and New Hampshire, which had its cases with orders rate reduced from 78 percent to 65 percent.

15. For measures based on a percentage, such as the adult entered employment rate, the standard for high performers was set at 102 percent of the state standard; for mid-performer standards, the standard was set at 101 percent of the state standard, and for low performers, the standard was set equal to the state standard. Standards based on dollar amounts used a similar but somewhat more complicated procedure.

16. Eventually, the common measures are expected to apply to all workforce programs, although individual programs may have additional measures.

17. In addition, in Michigan, the performance of each local area is computed directly by including variables for each local area in the adjustment model rather than by subtracting actual performance from predicted performance, and Michigan researchers have also developed an approach to project long-term outcomes based on short-term outcomes (Michigan Department of Labor and Economic Growth 2005). Unlike Michigan or the JTPA procedures, Washington models are based on several years of data rather than a single year of data and reestimated every three to four years rather than annually. These features are likely to make the adjustment procedures more stable not only because of the longer period between model revisions, but also because pooling across years contributes more observations and covers a wider range of economic conditions (Wolffhagen 2006).

References


---

**Keep Learning in 2010 and Beyond…!**

**ASPA Conference Papers**

The 2009 Annual Conference had over 100 applied practice workshops, panel sessions, and discussion circles addressing a wide variety of topics in Public Administration.

Conference Papers are now available at:

http://www.aspanet.org