Robert M.
La Follette School of Public Affairs
at the University of Wisconsin-Madison

Working Paper Series
La Follette School Working Paper No. 2008-023
http://www.lafollette.wisc.edu/publications/workingpapers

One Standard Fits All? The Pros and Cons of Performance Standard Adjustments

Carolyn J. Heinrich
La Follette School of Public Affairs, University of Wisconsin-Madison

cheinrich@lafollette.wisc.edu

Burt S. Barnow
Institute for Policy Studies, Johns Hopkins University

barnow@jhu.edu

Robert M. La Follette School of Public Affairs
1225 Observatory Drive, Madison, Wisconsin 53706
Phone: 608.262.3581 / Fax: 608.265-3233
info@lafollette.wisc.edu / http://www.lafollette.wisc.edu

One Standard Fits All?
The Pros and Cons of Performance Standard Adjustments

Burt S. Barnow
Associate Director for Research
Institute for Policy Studies
Johns Hopkins University
Wyman Building
3400 N. Charles St.
Baltimore, MD 21218-2696
barnow@jhu.edu
Telephone: (410) 516-7174
Fax: (410) 516-8233

Carolyn J. Heinrich
Professor of Public Affairs, Affiliated Professor of Economics
and Regina Loughlin Scholar
University of Wisconsin-Madison
1225 Observatory Dr.
Madison, WI 53706
cheinrich@lafollette.wisc.edu
Ph: 608-262-5443
Fax: 608-265-3233

November 18, 2008

Biosketches

Burt S. Barnow is Associate Director for research and Principal Research Scientist at the Institute for Policy Studies at the Johns Hopkins University. Dr. Barnow's research focuses on program evaluation, performance analysis, labor economics, welfare, and workforce programs. His current research includes an evaluation of the Department of Labor's High Growth Job Training Initiative and an assessment of occupational skill shortages. Dr. Barnow previously directed research and evaluations for the Employment and Training Administration.

Carolyn J. Heinrich is Professor and Regina Loughlin Scholar at the La Follette School of Public Affairs, affiliated Professor of Economics, and the Associate Director of research and training at the Institute for Research on Poverty at the University of Wisconsin-Madison. Her research focuses on social welfare policy, public management, and econometric methods for social-program evaluation. Dr. Heinrich is also the editor of the *Journal of Public Administration Research and Theory*.

Abstract

Despite the wide-ranging use of performance measures in public programs 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 still relatively rare. We explicate the 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 performance standards adjustments processes that are currently (or have been) in use and review the evidence of their effectiveness or problems encountered in their use. We also explore the consequences of the lack (or inadequacy) of performance standards adjustments for organizational incentives and achievements. We conclude with recommendations for the improvement of public sector performance measurement systems, arguing that this is a fruitful area for additional experimentation as well as for further academic research.

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

Introduction

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 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 (2001,38) argue that "counting and measuring are seen as the antidote to distrust," 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 differing internal and external audiences (Gormley, 2004). Among the most influential of recent developments in public sector performance measurement was the introduction of the U.S. Government Performance and Results Act (GPRA) 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).[1] In early field research on GPRA, Beryl Radin (2000), identified important limitations of GPRA's implementation in federal agencies, including a lack of attention to structural and institutional differences across agencies and to complexities of their functions and political environments that complicate efforts to establish viable performance measures and plans. With the growing availability of empirical data from these performance measurement systems, other researchers are also going beyond questions of *how* to "manage for results" and are exploring the implications of performance measurement for both individual behavioral responses and organizational outcomes.

It is not an objective of this paper to present a thorough review of the performance measurement literature and its findings, which would be beyond the scope of any paper, 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 paper, 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.

[1] Federal government agencies are required under GPRA 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.

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 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 (DWD), which only considers adjustments to performance standards in its Wisconsin Works (W-2) welfare to work program for "unusual or nonrecurring events."[2] 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 (CSPIA) 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 paper 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

[2] See the DWD policy memo on W-2 performance standards adjustments at http://www.dhfs.state.wi.us/em/ops-memos/2004/pdf/04-48.pdf (accessed on July 8, 2008).

generic approaches to adjusting performance standards, followed by 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 timeframe. 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" (TQM) principle (Deming, 1986).

Yet as discussed above, few of these 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 TQM that advocates the use of statistical analysis to adjust for factors

outside managers' control in evaluating and managing performance (Deming, 1986). Different approaches to adjusting performance expectations are possible, including statistical techniques (typically implying some type of regression modeling process), as well as less sophisticated procedures, such as negotiating standards or distinguishing or excluding particular cases or groups in performance calculations. For example, the federal No Child Left Behind Act (NCLB) requires annual, statewide testing to measure all students' mastery of academic content (per state standards) and reporting of the test results annually to the public. To account for differences in demographics and the academic preparation of students as they enter school, the results are disaggregated within schools by gender, race, ethnicity, English proficiency, and immigrant status. School districts are also allowed to offer alternative tests to some students (in particular, special education students and those with limited English proficiency), although this has contributed to concerns about how students are classified as exempt from regular testing and whether this provision has been misused to inflate performance.[3] In a Feb. 15, 2007 letter to the Senate Health, Education, Labor, and Pensions Committee (working on the NCLB reauthorization), Senator Russ Feingold and co-signing senators objected to the current form of NCLB testing, stating that: "While we all agree that states and districts should be held accountable for academic outcomes and continue working toward closing the achievement gap

[3] In December, 2003, NCLB regulations were changed for testing special education students, followed by revised policies for limited English proficiency students in February, 2004. States and districts can develop alternate assessments and use them to test special education students; however, only up to 1 percent of students in the grade levels tested can take tests based on alternative achievement standards and have their scores counted for meeting the federal mandate of showing "adequate yearly progress." The rule changes for students with limited English proficiency state that schools are no longer required to give students with limited English proficiency their state's reading test if such students have been enrolled in a U.S. school for less than a year.

among their students, federal education law should not take the form of a one-size-fits-all, cookie-cutter approach."[4]

Although we will briefly discuss some alternatives to statistically adjusting performance expectations in this paper, in light of the limitations associated with less formal 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_s) to which an individual or organization's measured performance (P_m) is compared. In public workforce investment 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: $P_0 = \alpha + \beta_1 X + \beta_2 Z + \varepsilon$. The estimates of β_1 and β_2 (vectors) are subsequently used as weights for the influence of these factors in adjusting the common standard (P_s) to derive unit (e.g., agency)-specific performance targets 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_{si}) and its measured performance (P_{mi}); see Courty et al. (2005) for additional details on this type of model and calculations from the Job Training Partnership Act (JTPA), as well as further discussion of the JTPA program below.

A slight variant of the above approach to adjusting performance is described by Stiefel et al. (1999) and illustrated with an application to public elementary school data from Chicago.[5] In

[4] See http://www.nea.org/esea/lettertohelp0207.html.

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 (P_{mi}). 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. Before elaborating further on how these techniques are applied in practice, we first consider some basic arguments for and against adjusting 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, including: 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.

[5] Rubenstein et al. (2003) illustrate the use of this approach with an example from hospital performance analysis.

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.[6] 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 et al. (2004:4) 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" 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. follow the approach of Stiefel et al. and estimate a model of measured performance (P_m) 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 always or sometimes not controlled by program managers) in this model, Bartik et al. add a set of indicator variables, W_j, for each LWA *j*, to the equation. They

[6] 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.

explain that because the number of LWAs in a given state is limited, estimating a model using data on LWA group means would result in very imprecise estimates. Importantly, they argue that if it was possible to measure all relevant X and Z factors that affect individual performance outcomes, the estimated W_j 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 et al. (p. 8) argue 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."

In a Monte Carlo data analysis, Brooks (2000) addresses the question as to whether adjusted performance measures are more likely to approximate true performance, particularly in the case where 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. In effect, it is unlikely that models for adjusting performance will improve our estimates of performance, and they may, Brooks cautions, be "extremely misleading," unless the specification is very accurate and the data for estimating it are complete (p. 323). Although we concur with Brooks' technical points, we believe that he is probably overstating the danger of making adjustments. 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 below, 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 policymakers 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. For example, a three-level model has been used to estimate, at level one, within-student achievement over time; at level two, student achievement growth as function of student characteristics and student participation in interventions, and at level three, the estimated effects of any policy interventions on student achievement as a function of school characteristics and the environment (see Roderick, Jacob, and Bryk, 2000). Thus, measures of performance (i.e., student achievement) are not only adjusted for factors over which educators have little control, but the models also attempt to explicitly estimate the contributions of policies and interventions

that are directly controlled by educators and administrators to student outcomes. In their analysis that compared "raw" rankings of schools to adjusted school performance measures based on a value-added (conceptual) model, Rubenstein et al. (2003: 612) saw considerable differences in rankings after adjusting for factors such as poverty. They concluded it was "reasonable to describe a school that produces large gains among a primarily low-income student body as a higher-performing school," even if its absolute performance was lower.

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 that "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. DOL-ETA, 2001). It was suggested that factors outside managers' control that affected program outcomes (some unmeasurable) might be more easily conveyed and weighed in negotiations. Along the same lines, Cook and Ludwig (2006: 696) have rebuked the excessive emphasis on strict scientific standards for producing knowledge about program impacts, suggesting that it has led to us to discount much relevant information that is "concerned with basic beliefs about human nature and interactions" and is important to understanding program outcomes.

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 timeframe, regardless of their backgrounds, special needs, school resources, etc., 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 NCLB 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% 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 one-half 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 under-reporting 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. His analysis of particular test question items also showed that the test score gains did not reflect broader skills improvement but rather increases in test-specific skills (related to curriculum alignment and test preparation), primarily in math. 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 adjustments 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 (JTPA) 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 above, 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.[7] Economists working for the Assistant Secretary for Policy, Evaluation, and Research (ASPER) 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. Recognizing that local programs served very different populations in varying local economic conditions, they argued that it would be unfair to set the same earnings or employment rate standard for local areas with high unemployment and disadvantaged populations as for affluent areas and advocated the use of regression analysis to adjust for local differences.

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

The performance management efforts under CETA laid the groundwork for their formalization under JTPA, the nation's primary workforce development program from 1982 through July 2000. JTPA included funding streams for disadvantaged adults, dislocated workers, and youth, although we focus on the subprogram for disadvantaged adults, which had the most sophisticated performance management system.[8] The JTPA statute that mandated the establishment of a performance management system (Section 106) stated that "job training is an investment in human capital… and that it is essential that criteria for measuring the return on this investment be developed." For the adult program, Section 106 specified performance measures including "placement in unsubsidized employment, retention for not less than 6 months in unsubsidized employment, an increase in earnings including hourly wages, a reduction in welfare dependency, and acquisition of skills…"[9] 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 that 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 employment rate would decline by 0.5 percentage points, if all other characteristics remained the same.

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

[8] 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 (OJT), 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.

[9] For youth, the measures included attainment of competencies, dropout prevention, secondary and postsecondary school completion, and enrollment in other training programs, apprenticeships, postsecondary education, and the armed forces.

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 25th 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. 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 GPRA—organizations that did well could obtain substantial additional resources, and those that did poorly could lose their right to operate the program.

Economists at the DOL recognized that local program performance was likely driven by characteristics of the population served and local economic conditions, as well as the management and programmatic strategies of SDA staff. Hence, from the program's inception, the DOL regularly developed and disseminated regression models that could be used by states to put SDAs on a "level playing field," as stated in DOL guides:

> Performance standards are adjusted to "level the playing field" by making the standards neutral with respect to who is served and to local economic conditions. For example, an SDA serving a hard-to-serve population would be given a lower standard than an SDA serving a less hard-to-serve population. Although set at different levels, meeting those two standards would require the same level of SDA effort. Similarly, an SDA facing difficult local economic conditions might be given a lower standard than an SDA in a booming economy. (Social Policy Research Associates, 1999).

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 PY 1998-1999 models for the adult, youth, and dislocated worker programs.

The regression models developed *at the SDA level* each year were based on data from the previous program year and only retained variables with coefficients that were statistically significant and of the expected sign. Thus, for example, if the coefficient for percent of participants who were black in the entered employment rate model was either positive (counter-intuitively) or not statistically significant, that variable would be excluded from the model for that year.[10] 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 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 DOL regression models or some alternative approach that was approved by the DOL (Barnow, 1992).[11]

Despite the DOL's efforts to "level the playing field" and reduce creaming, there is ample evidence of gaming of the performance management system by local programs to improve

[10] 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, p. III-8).

[11] In fact, the criteria for using adjustments other than the DOL 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.

their measured performance, through actions such as the strategic enrollment and timing of clients' entry and exit from the program (Courty and Marschke, 1996, 1997, 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.[12] The practice of only retaining statistically significant variables in the regression adjustment models that had a coefficient of the expected sign and a reasonable magnitude implies that in some cases, a variable would lead to adjustments in some years but not in others, and the magnitude of the adjustment could vary significantly from year to year. Although Barnow (1996) showed inconsistent treatment of people with disabilities because of this feature, changes in the adjustment procedures did not follow.[13] 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).

[12] 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).

[13] 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.

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 policymakers 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 DOL 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.[14] In 2007, there were 94 Job Corps centers operated by nonprofit and for-profit vendors under contract with the DOL, 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, 2001). In addition, the centers that are not operated by the federal government are paid through performance-based contracts. 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.[15] To illustrate how the system operates, Table 2 shows the adjustment factors and the adjustments for the Program Year 2007 Graduate Six-Month Weekly Earnings model. The rationale offered for adjusting some of the standards is the same as that of

[14] The description of Job Corps is from http://jobcorps.dol.gov/about.htm on September 28, 2007 and from http://jobcorps.dol.gov/docs/jc_directory.pdf on September 28, 2007.

[15] 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, post-enrollment placement rate, graduate placement rate, graduate six-month follow-up placement rate, and graduate 12-month follow-up placement rate.

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, 2001, Appendix 501, p. 6). For the average wage at placement model, some of the adjustment factors include: participant age, reading and numeracy scores (for GED attainment), occupational group (e.g., business, construction, and food services), the average wage in all industries in the county, the percent placed in a job in a state with a high minimum wage, and the average percent of families in poverty in the county where the center is located. Models are updated annually, and 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 (Gormley, 2004). These report cards are intended to assist consumers and referring physicians in selecting the most appropriate provider for high-risk procedures. The New York State report card for coronary artery bypass graft surgery (CABG) is discussed here; similar procedures are used by New York for cardiac valve procedures. CABG is a procedure where 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 or not the patient 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 CABGs.

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 which 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 sub-national 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 where the absence of adjustments affects performance rankings. The child support enforcement program in the United States is a federal-state partnership, with funding provided by both the federal government and states. 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 a state's performance in five areas: (1) paternity establishment (in which states choose between 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, calling on the Secretary of the Department of Health and Human Services (DHHS) to recommend adjustments to ensure that the relative performance of the states is measured from a baseline that takes into account such factors. 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. [16]

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 five percentage

[16] These variables included: personal income per capita, poverty rate, percent of males aged 20-64 not working, rate of job growth, percent of population living in urban areas, percent of TANF case heads under age 30, percent of IV-D cases currently participating in TANF, percent of IV-D cases that have never participated in TANF, number of IV-D cases per full-time equivalent (FTE) staff, 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.

points; for a few states, however, the adjustment was quite large and would have led to a sizeable change in the state's relative and absolute performance.[17] 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:

> Child support officials were asked if the new system favors states with certain profiles, such as affluent populations, and, if so, whether states should be compensated for factors beyond their control. Most respondents agreed that CSE programs operate in different socio-economic environments. However, the majority opposed adjusting incentive payments to account for those differences, because such adjustments would add complexity and uncertainty to an already complicated payment system and gaining consensus on an appropriate list of adjustment factors would be extremely challenging (Gardiner et al., 2004, 3).

The CSE experience illustrates key points we have made in this paper. First, including adjustments to performance standards has the potential to significantly influence performance ratings for some programs. At the same time, 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 where 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

[17] 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.

two-parent families, an annual performance bonus system was designed to award $20 million each to as many as five states with the largest reduction in the proportion of out-of-wedlock births to total births. In determining the award winners each year, HHS 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), DHHS 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" (HHS News, September 15, 2000). 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 (2001, 4) 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.

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 which may have little to do with any policy actions taken by states—and for which there are no formal procedures for evaluating states' contributions—appear to undermine basic goals of performance measurement systems described earlier, i.e., 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 (WIA) program (the successor to JTPA), in part due to 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. DOL-ETA, 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, 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 upon the parties involved in these negotiations to adjust for economic and population characteristics in setting performance expectations.

Two national studies of 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, p. V-3.)

A second study, released the following year, reached similar conclusions:

> All states and local areas in the study sample expressed concerns about the WIA performance management system. Most officials interviewed indicated that the WIA system represents a step backward to the approach used under JTPA as follows:
>
> - They decried the absence of formal procedures to adjust for characteristics of participants served and local economic conditions; state and local officials stated that failing to adjust for differences in these factors implies that states and local areas are not placed on a level playing field.
> - State officials expressed concern that the DOL ETA regional office officials did not enter into negotiations with state officials; they indicated that federal officials did not discuss what the state standards should be, citing pressure from the federal government to meet its standards (Barnow and King, 2005).

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.

While the DOL recently stated that "In general, the process for this current round of negotiations will not change significantly from the processes used in past rounds" (U.S. Department of Labor, 2007), new guidance from the DOL includes information on relationships

between 14 of the 17 performance measures and six potential adjustment factors.[18] In addition, 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, like 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).[19] The Maryland approach was abandoned, however, after its first year of use, in part because the adjustments were viewed as arbitrary and involved considerable work for relatively small adjustments.

The DOL also awarded grants to two states, Michigan and Washington, to develop regression-based adjustment models for WIA performance standards. This effort was motivated in part by concerns that a local workforce area "may be credited with greater success than others, not because its services are more effective, but simply because the individuals it enrolls are better qualified to obtain or retain a job or because the local economy is more robust in creating jobs" (Bartik, Eberts, and Kline; 2004). Although both states are using regression analysis to develop adjustments for local workforce area performance standards, the two approaches differ in important ways. The Michigan models are used to adjust the "common measures" promulgated by the Office of Management and Budget, while the Washington models are used to develop

[18] The potential adjustment factors for the adult entered employment rate are the unemployment rate, percent female, percent age 55 or older, percent not high school graduate, percent low income, and percent with disabilities. The guidance indicates that all the relationships are statistically significant, but it does not indicate if the relationships were obtained from univariate or multivariate regressions.

[19] 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-level performers, 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.

adjustments for the current WIA performance measures.[20] Second, the Michigan models are intended to be illustrative only 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 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 are re-estimated 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 (Wolfhagen, 2006).

In 2007, the DOL 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 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

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

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 where adjustments to performance standards have been considered, and even fewer where 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 oft-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 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 of 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 et al. (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 if 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 DOL 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 timeframe 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.”)

Table 1: JTPA Performance Measures, National Standards and Factors in the National Adjustment Model for JTPA Adult Follow-Up Rate Performance Standards in Program Years 1998 and 1999

Performaneice Measures	**National Standards**
Title II-A Adult	
Adult follow-up employment rate	60%
Adult weekly earnings at follow-up	$289
Welfare follow-up employment rate	52%
Welfare weekly earnings at follow-up	$255
Title II-C Youth	
Youth entered employment rate	45%
Youth employability enhancement rate	40%
Title III Dislocated Workers	
Entered employment rate	73%

Factors and Adjustments in National Model for Adult Follow-Up Employment Rate			
Terminee Characteristics (Percent)	*Regression Adjustment*	*Local Economic Conditions*	*Regression Adjustment*
Female	-0.050	Unemployment rate	-0.608
Age 55 or more	-0.130	Three-year growth in earnings in trade	0.245
Not a high school graduate	-0.066	Annual earnings in trade	-0.539
Post high school (including college)	0.008	Families with income below poverty level (percent)	-0.211
Dropout under age 30	-0.015		
Black (not Hispanic)	-0.027		
Minority male	-0.026		
Cash welfare recipient	-0.031		
Long-term TANF recipient	-0.018		
Supplemental Security Income (SSI) recipient	-0.133		
Basic skills deficient	-0.037		
Individual with a disability	-0.096		
Lacks significant work history	-0.055		
Homeless	-0.043		
Viet Nam era veteran	-0.081		
Not in labor force	-0.108		
Unemployed 15 or more weeks	-0.073		
Unemployment insurance claimant or exhaustee	0.022		

Source: Social Policy Research Associates (1999). Guide to JTPA Performance Standards for Program Years 1998 and 1999. Menlo Park, CA: Social Policy Research Associates, p. I-3 and II-10.

Table 2: Job Corps Performance Standards Adjustment Factors and Adjustments for Program Year 2007 Six-Month Weekly Earnings

Performance Measure	Adjustment from Regression Model
Average Age at Termination	7.8121
% High School Diploma or GED at termination	0.1449
% Vocational Completion at Termination	0.2661
% Reading Functional Level 5 at Termination	0.1231
% Reading Functional Level 6 at Termination	0.2233
% Math Functional Level 4 at Termination	0.1578
% Math Functional Level 5 at Termination	0.2660
% Math Functional Level 6 at Termination	0.3543
% Training in Bricklayer or Cement Occupations	0.2920
% Training in Business Occupations	-0.5529
% Training in Carpentry Occupations	0.2524
% Training in Construction Occupations	0.3111
% Training in Food Service Occupations	-0.5301
% Training in Health Occupations	-0.4191
% Training in Mechanical Occupations	0.3132
% Training in Service Occupations	-0.3287
% Training in Welding Occupations	0.6586
% Training in Other Occupations	0.0035
Average Wage in All Industries in County ($1,000's)	1.8077
% Placed in Job in State with High Minimum Wage	0.3270
Average Percent of Families in County in Poverty	-2.1928

Source: U.S. Department of Labor, Office of Job Corps. 2001. *Policy and Requirements Handbook: Job Corps*. Washington, DC: U.S. Department of Labor, Office of Job Corps. Revised July 2007, Appendix 501C, Attachment 1, p. 4.

References

Barnow, Burt S. 1992. "The Effects of Performance Standards on State and Local Programs: Lessons for the Job Opportunities and Basic Skills Programs." In C. Manski & I. Garfinkel (Eds.), *Evaluating Welfare and Training Programs* (pp. 277-309). Cambridge, MA: Harvard University Press.

Barnow, Burt.S. 1996. "Policies for People with Disabilities in U.S. Employment and Training Programs," in *Disabilities, Cash Benefits, and Work*, Jerry L. Mashaw, Virginia Reno, Richard Burkhauser, and Monroe Berkowitz Eds. Kalamazoo, Michigan: Upjohn Institute for Employment Research.

Barnow, Burt S. and Christopher T. King. 2005. *The Workforce Investment Act in Eight States.* Albany, NY: The Nelson A. Rockefeller Institute of Government.

Barnow, Burt S. and Jeffrey A. Smith. 2004. "Performance Management of U.S. Job Training Programs: Lessons from the Job Training Partnership Act." *Public Finance and Management,* 4(3): pp. 247-287.

Bartik, Timpthy J., Randall Eberts, and Ken Kline. 2004. "Estimating a Performance Standards Adjustment Model for Workforce Programs that Provides Timely Feedback and Uses Data from Only One State." Kalamzoo, MI: The W.E. Upjohn Institute for Employment Research. Unpublished manuscript.

Bertelli, Anthony M. and Laurence E. Lynn, Jr., 2006. *Madison's Managers: Public Administration and the Constitution*. Baltimore: The Johns Hopkins University Press.

Blalock, Hubert M. Jr. 1961. *Causal Inferences in Nonexperimental Research.* Chapel Hill, NC: University of North Carolina Press.

Boyle, David. 2001. *The Sum of Our Discontent: Why Numbers Make Us Irrational*. New York: Texere.

Brooks, Arthur C. 2000. "The Use and Misuse of Adjusted Performance Measures." *Journal of Policy Analysis and Management*. 19(2): 323-329.

Cook, Phillip J. and Jens Ludwig, 2006. Aiming for evidence-based gun policy. *Journal of Policy Analysis and Management* 25(3): 691-735.

Courty, Pascal, Carolyn J. Heinrich, and Gerald R. Marschke. 2005. "Setting the Standard in Performance Measurement Systems." *International Public Management Journal,* 8(3): 321-347.

Courty, Pascal and Gerald Marschke. 1996. "Moral Hazard under Incentive Systems: The Case of a Federal Bureaucracy." In *Advances in the Study of Entrepreneurship, Innovation and Economic Growth*, Volume 7. Ed. Gary Libecap. Greenwich, CT: JAI Press, pp. 157-190.
.
Courty, Pascal, and Gerald Marschke. 1997. "Measuring Government Performance: Lessons from a Federal Job-Training Program." *American Economic Review*. 87(2): pp. 383-388.

Courty, Pascal and Gerald Marschke. 2004. "An Empirical Investigation of Gaming Responses to Explicit Performance Incentives." *Journal of Labor Economics*. 22(1): pp. 22-56.

Deming, W. Edwards 1986. *Out of the Crisis*. Cambridge, MA: MIT Institute for Advanced Engineering Study.

Frederickson, David G. and George H. Frederickson. 2006. *Measuring the Performance of the Hollow State*. Washington, DC: Georgetown University Press.

Gardiner, Karen N., Michael E. Fishman, Asaph Glosser, and John Tapogna. 2004. *Study of the Implementation of the Performance-Based Incentive System.* Falls Church, VA: The Lewin Group.

Gormley, Willaim T., Jr. 2004. Using Organizational Report Cards. In *Handbook of Practical Program Evaluation* Second Edition, edited by Joseph S. Wholey, Harry P. Hatry, and Kathryn P. Newcomer. San Francisco, CA: Jossey-Bass. 628-648.

Governor's Workforce Investment Board. 2001. "Performance Incentive Plan for Maryland Workforce Investment Areas." Baltimore, MD

Heinrich, Carolyn J. 2004. Improving Public-Sector Performance Management: One Step Forward, Two Steps Back? *Public Finance and Management,* 4(3), 317-351.

HHS News. "HHS awards $100 million bonuses to states achieving largest out-of-wedlock births." September 15, 2000. (See: www.hhs.gov/search/press.html*).*

House Research Organization, Texas House of Representatives. 2004. "Examining Teacher Performance Incentives." Focus Report Number 78-17.

Jacob, Brian A. 2005. Accountability, incentives and behavior: Evidence from school reform in Chicago." *Journal of Public Economics* 89(5-6): 761-796.

Jacob, Brian A. and Levitt, Stephen D. 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics* CXVIII(3): 843-877.

Jha, Ashish and Arnold M. Epstein 2006. The Predictive Accuracy of the new York State Coronary Artry Bypass Surgery Report-Card System. Health Affairs 25(6): 844-855.

Klitgaard, Robert and Paul C. Light. 2005. *High-Performance Government*. Santa Monica: RAND Corporation.

Michigan Department of Labor and Economic Growth. 2005. "The Value Added Performance Improvement Model. Measuring Workforce Development Program Success Technical Assistance Guide." Lansing, MI: Michigan Department of Labor and Economic Growth.

New York State Department of Health. 2004. *Adult Cardiac Surgery in New York State: 1998-2000*. Albany, NY: New York State Department of Health.

Offner, Paul. 2001. Reducing Non-Marital Births. The Brookings Institution Policy Brief No. 5, August.

Radin, Beryl 2000. The Government Performance and Results Act and the Tradition of Federal Management Reform: Square Pegs in Round Holes? *Journal of Public Administration Research and Theory,* 10(1), 11-35.

Radin, Beryl 2006. *Challenging the Performance Movement*. Washington, D.C.: Georgetown University Press.

Roderick, Melissa, Brian Jacob, and Anthony S. Bryk. 2000. "Evaluating Chicago's Efforts to End Social Promotion." In *Governance and Performance: New Perspectives*, eds. Lynn, L. and Heinrich, C. Washington, D.C.: Georgetown University Press.

Rubenstein, Ross, Leanna Stiefel and Amy Ellen Schwartz. 2003. "Better than Raw: A Guide to Measuring Organizational Performance with Adjusted Performance Measures." *Public Administration Review* 63(5): 607-615.

Shulock, Nancy B. 1999. The paradox of policy analysis: if it is not used, why do we produce so much of it? *Journal of Policy Analysis and Management* 18(2): 226-244.

Social Policy Research Associates. 2004. *The Workforce Investment Act after Five Years: Results from the National Evaluation of the Implementation of WIA*. Oakland, CA: Social Policy Research Associates.

Social Policy Research Associates. 1999. *Guide to JTPA Performance Standards for Program Years 1998 and 1999*. Menlo Park, CA: Social Policy Research Associates.

Stiefel, Leanna, Ross Rubenstein and Amy Ellen Schwartz. 1999. "Using Adjusted Performance Measures for Evaluating Resource Use." *Public Budgeting and Finance* 19(3): 67-87.

Tapogna, John, K, Karen Gardiner, Burt Barnow, Michael Fishman, and Plamen Nikolov. 2002. *Study of State Demographic and Economic Variables and their Impact on the Performance-Based Child Support Incentive System.* Falls Church, VA: The Lewin Group.

Thorn, Christopher A. and Robert H. Meyer 2006. "Longitudinal data systems to support data-informed decision making: A tri-state partnership between Michigan, Minnesota and Wisconsin." Report of the Wisconsin Center for Education Research.

U.S. Department of Labor, Office of Job Corps. 2001. *Policy and Requirements Handbook: Job Corps*. Washington, DC: U.S. Department of Labor, Office of Job Corps. Revised July 2007.

U.S. Department of Labor, Employment and Training Administration. 2001. 2002 Annual Performance Plan for Committee on Appropriations.

U.S. Department of Labor. 2002. Employment and Training Administration. Training and Employment Guidance Letter No. 11-01. "Guidance on Revising Workforce Investment Act (WIA) State Negotiated Levels of Performance." Washington, DC: U.S. Department of Labor, Employment and Training Administration.

U.S. Department of Labor. 2007. Employment and Training Administration. Training and Employment Guidance Letter No. 19-06. "Negotiating Performance Goals for the Workforce Investment Act Title 1B Programs and Wagner-Peyser Act Program for Program Years 2007 and 2008." Washington, DC: U.S. Department of Labor, Employment and Training Administration.

Wiseman, Michael 2004. "The High Performance Bonus." Washington, DC: The George Washington University. Unpublished manuscript.

Wolfhagen, Carl 2006. "WIA 1-B Performance Regression Models of Federal and State Performance Measures." Olympia, WA: Washington State Workforce Training and education Coordinating Board. Unpublished manuscript.