The Role of Performance Management in Workforce Investment Programs

Burt S. Barnow
Johns Hopkins University

Prepared for the Conference
What the European Social Fund Can Learn from the WIA Experience

Sponsored by

European Commission Directorate-General for Employment,

Social Affairs and Equal Opportunity

and

University of Maryland School of Public Policy

Washington, DC   November 7, 2009

1. Introduction

Since the late 1970s, major federal workforce development programs in the United States have included performance management systems that are used to assess how well the programs are performing at the national, state, and local levels. The use of performance management in workforce programs predates the more general congressionally mandated performance requirements of the Government Performance and Results Act (GPRA). This paper draws on the previous work of the author and others in assessing the lessons of the past 30 years of experience with performance management in workforce programs. Although the paper focuses on the U.S. system, the lessons should apply to programs in other countries as well.

The paper first discusses what performance management is in the context of workforce programs. Next, performance management is contrasted and compared with program evaluations. Policy officials often would like to implement performance measures that are based on program impact; the next section of the paper describes why that is generally not possible to do and presents empirical findings on the success of such efforts. The following section describes how the performance management system used for U.S. workforce programs can lead to unintended results and summarizes some of the research on this topic. This is followed by a discussion of whether standards should be absolute or adjusted for factors such as participant characteristics and economic conditions. The final section presents lessons for countries that are considering establishing a performance management system.

Although related, the concepts of performance measurement and evaluation are distinct and serve different purposes, they are sometimes confused.[1] Performance measurement is a management tool that is used to monitor implementation on a real time basis. Performance measures may track data that indicates fidelity in program implementation, inputs (such as participant characteristics) that are considered important to the program's purpose, process measures (e.g., use of "best practices"), outputs expected from the program, and sometimes short-term gross outcome measures. Program evaluation, on the other hand, is intended to answer specific questions about programs. Process studies document what happened while the program is implemented, impact evaluations assess what difference in outcome measures was due to the intervention, and cost-benefit analyses assess whether the benefits of a program exceed the costs.

If the program has limited capacity, participant characteristics may be a useful performance measure, and one or measures could be established to track the characteristics of customers served.[2] Process measures rather than output or outcome measures are sometimes used. For example, if particular practices are known to be more effective or less expensive than the alternatives, a case can be made for including process measures of performance. In the current health care reform debate in the United States, some advocates argue that costs can be driven down by requiring providers to use best practices or by providing financial incentives to do so; similar arguments can be raised in setting standards for education. In the past, however, some in the workforce field have argued that so long as the grant recipients are held accountable

[1] See Blalock and Barnow (2001).

[2] The performance management system for JTPA required local programs to have at least 65 percent of participants served classified as "hard to serve" to qualify for performance bonuses.

for the desired results, they should be free to adopt the approach they believe is best rather than relying on processes prescribed by the federal government.[3] A reasonable approach might be to monitor use of best practices and provide technical assistance, rewards, and sanctions only when an organization fails to achieve satisfactory outcomes.

In a system characterized by delegation of authority from the central government to lower levels of government (state and local government for many U.S. programs, but the concepts apply to a system of grantees or for-profit contractors as well), the goals of the level of government providing the funds may not be aligned with the goals of the level of government providing the services. By instituting a performance management system that provides rewards and sanctions based on how well the lower level of government meets the goals of the funding agency, the so-called principal-agent problem can be (in theory) resolved.

The differences between performance measurement and evaluation often are matters of depth of analysis and causality. Because of the need for rapid feedback, performance measurement activities generally track easy to collect data on inputs, activities, and outputs. Data for performance measurement generally comes from management information systems maintained by the programs and from administrative data collected for other purposes.[4] Evaluations are usually conducted less often and with greater resources, so a process study can

[3] The JTPA Advisory Committee explicitly rejected the idea of dictating process to state and local governments: "In the business world, it is now widely accepted that the excellent companies define their expected results explicitly, and tightly measure performance against them, while allowing their producers to have discretion in how they attain those results. We suggest that JTPA emulate this model" (Job Training Partnership Advisory Committee, 1989, p. 27.)

[4] In the United States, data on employment and earnings for workforce programs comes from administrative data maintained by states to determine eligibility and benefit levels for unemployment insurance. Although such data is not perfect, as it misses self-employment earnings, off-the-books employment, and sometimes employment in other states or for government, it is inexpensive to use relative to conducting a survey, and it avoids recall issues.

make use of extensive interviews to document program implementation. Performance management activities cannot usually afford the time and resources required for tracking long-term outcomes and establishing and tracking a control group or comparison group, so performance measures are usually based on gross post-program measures (such as earnings during a post-program period), while evaluations can estimate program impact (by, for example, comparing earnings of participants after participation with earnings of a control group of applicants that was excluded from the program through random assignment). For example, performance measures for a vocational training program can include placement rates, wages at placement, and perhaps short-term follow-up measures of employment and earnings for participants, but an impact evaluation will focus on the change in employment and earnings due to the program, usually for a significantly longer period. Both types of activities are important for management and policy development, but, as discussed below, one should avoid reading more into performance results than is actually there.

2. Performance Management for U.S. Workforce Development Programs

In the United States, implementation of GPRA has led various programs to embrace alternative concepts of why performance management is useful. According to the statute, GPRA was designed to hold "federal agencies accountable for achieving program results." In particular, GPRA requires that agencies develop performance measures and standards for the programs they administer, as well as strategic plans to achieve their goals.

A performance management system must include three components: (1) measures of performance, (2) standards for acceptable performance, and (3) feedback on performance. As discussed below, measures for U.S. workforce programs have attempted to focus on program impacts, but that need not be the case.

The U.S. Department of Labor established performance measures long before such measures were mandated under GPRA. Performance measures were first established in the late 1970s for the Comprehensive Employment and Training Act (CETA). The Job Training Partnership Act (JTPA), which was the major national workforce program in the 1980s and 1990s, had statutory provisions calling for measuring performance as the impact of the program on employment and earnings relative to program cost.[5] Specifically, Section 106 of JTPA, which provided the requirements for performance standards, stated

> The Congress recognizes that job training is an investment in human capital and not an expense. In order to determine whether that investment has been productive, the Congress finds that-- (1) it is essential that criteria for measuring the return on this investment be developed; and (2) the basic return on the investment is to be measured by long-term economic self-sufficiency, increased employment and earnings, reductions in welfare dependency, and increased educational attainment and occupational skills.

The JTPA statute suggested, but did not require, that measures for adults include the employment rate in unsubsidized employment, employment retention for six months, an increase in earnings and/or the wage rate, a reduction in welfare dependency, and acquisition of skills. In practice, the performance measures used for JTPA were primarily program outcomes that, at best, served as proxies for program impact. Initially, the measures focused on the status of participants at the time of exit from the program or shortly thereafter, but by the time the program was replaced by the Workforce Investment Act (WIA), a follow-up period of 13 weeks

[5] During this period the Work Incentive Program (WIN), which provided employment and training services to welfare recipients, also had a performance management system that distributed some of the funding to states based on their success on measures such as welfare grant reductions, the entered employment rate, wage rates of WIN participants who obtained jobs, and job retention. See U.S. General Accounting Office (1982).

was used for most measures. The statute originally also called for cost measures, but as described below, this requirement was repealed in 1992.[6]

Under WIA, the Office of Management and Budget (OMB) has advocated that all programs with a workforce goal have "common measures," but agencies other than the U.S. Department of Labor have resisted adopting the common measures. Currently, the common measures for adults and dislocated workers are:

- **Entered employment rate:** Of those not employed at the date of participation, the number of participants who are employed in the first quarter after the exit quarter divided by the number of participants who exited during the quarter.
- **Employment retention rate:** Of those who are employed in the first quarter after the exit quarter, the number of participants who are employed in both the second and third quarters after the exit quarter divided by the number of participants who exited during the quarter.
- **Average Earnings:** Of those participants who are employed in the first, second, and third quarters after the exit quarter, total earnings in the second quarter plus total earnings in the third quarter divided by the number of participants who exited during the quarter.[7]

For youth programs, the common measures use a broader concept of a successful outcome by including training and education, and the measures include attainment of a certificate or degree and literacy and numeracy gains.

3. Adjustments to Performance Standards[8]

U.S. workforce investment programs have varied in their approach to adjusting

[6] For a discussion of the JTPA performance management system in its latter years, see Social Policy Research Associates (1999).

[7] See Training and Employment Guidance Letter 17-05, Attachment A. Available at http://wdr.doleta.gov/directives/attach/TEGL17-05_AttachA.pdf on October 24, 2009.

[8] This section is based on Barnow and Heinrich (forthcoming).

performance standards, both among programs and over time. This section first describes how performance standards have been adjusted for major U.S. workforce investment programs and then summarizes the pros and cons of adjusting standards.

For the primary workforce investment programs administered by the U.S. Department of Labor, JTPA in the 1980s and 1990s and WIA beginning in 2000, state and local area standards were subject to adjustment, but the approach has varied greatly. Initially under JTPA, governors had three options for adjusting standards for the SDAs within their jurisdiction: using the national standards established by the Secretary of Labor; using regression models developed by the Department of Labor to adjust standards for variation in participant characteristics and economic conditions; or developing their own adjustment system. The legislation was later amended to make it relatively difficult for governors to use any adjustment mechanism other than the national model. Under WIA, standards were established at the state level through negotiations, although some have commented that the Department of Labor officials imposed standards on the states with no opportunity to truly negotiate; states can determine local workforce area standards in any manner they choose.[9]

Although many public programs, including all federal agencies, are required to establish performance standards, there are 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

[9] See Social Policy Research Associates (2004) and Barnow and King (2005) for a discussion of state and local perceptions of performance management under WIA.

take account of 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 adjustments to performance standards 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. As described below, there is ample evidence that workforce investment programs have responded to performance management systems by cream-skimming from the pool of eligible individuals.

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 estimates of the effects of specific explanatory variables 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.

While there are merits in the arguments both for and against the use of performance adjustments, it appears that 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.

In their assessment of adjusting performance standards, Barnow and Heinrich (forthcoming) 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, Barnow and Heinrich (forthcoming) 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, DOL is now providing guidance on how changes in circumstances (such as 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. 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."

4. Evidence on the Effects of Performance Management in Workforce Investment Programs

The Employment and Training Administration has had substantial experience with performance standards, and a number of studies have been conducted on the impacts of performance management on participants served, activities, costs, and program impacts. While most analysts note the strong rationale for developing performance measures for government programs, there has been considerable controversy in the literature regarding the benefits of performance management systems, particularly as they have been applied since enactment of the GPRA in 1993. This section of the report reviews the literature on performance standards for workforce programs; most of the research was conducted on the performance standards system used under the Job Training Partnership Act (JTPA), WIA's predecessor.[10] Although much of the literature on performance management points to its salutary effects, there is little doubt from the literature that instituting performance standards can have a strong impact on program behavior, and not always in the desired direction. This section of the report summarizes the literature on performance standards in employment and training programs in several key areas: (1) the impact of performance standards on who is served, (2) the impact of performance standards on the services provided, (3) the relationship between performance measures and program impacts, (4) strategic responses by state and local programs to performance standards, and (5) lessons learned by ETA and states/localities on the use and effects of efficiency measures/standards.

The Impact of Performance Standards on Who Is Served. The majority of the employment and training literature on performance incentives addresses the question of their effect on who gets served. Under JTPA, local service delivery areas (SDAs) had strong incentives to serve persons likely to have good labor market outcomes, regardless of whether those outcomes were due to JTPA because the performance measures used focused on post-program levels of employment and earnings. Similar incentives guide the WIA program. In fact, the absence of a regression model to adjust standards for serving individuals with labor market barriers should make these incentives stronger under WIA than they were under JTPA.

The literature divides this issue into two parts. First, do SDAs (WIBs under WIA) respond to these incentives by differentially serving persons likely to have good outcomes,

[10]For a more in-depth review of the literature on performance standards in workforce programs, see Barnow and Smith (2004); most of this section is based on Barnow and Smith (2004). For a critical review of the performance management movement, see Radin (2006).

whether or not those good outcomes result from the effects of the program? This is the literature on "cream skimming." Second, if there is cream skimming, what are its impact effects? Taking the best among the eligible could be economically efficient if the types of services offered by these programs have their largest net impacts for this group. In what follows, we review the literature on each of these two questions in turn.

Do employment and training programs cream skim? Several papers examine whether or not JTPA program staff cream skimmed in response to the incentives provided by the JTPA performance system. The key issue in this literature is the counterfactual: to what group of non-participants should the participants be compared in order to determine whether or not cream skimming has occurred? In all cases, the studies proceed by comparing observable characteristics correlated with outcomes, such as education levels or participation in transfer programs such as Aid to Families with Dependent Children (AFDC) or Temporary Assistance for Needy Families (TANF). A finding that participants have "better" characteristics relative to non-participants in the form of higher mean years of schooling or lower average pre-program transfer receipt, is interpreted as evidence of cream skimming.

Anderson et al. (1992, 1993) compare the characteristics of JTPA enrollees in Tennessee in 1987 with the characteristics of a sample of individuals eligible for JTPA in the same state with data constructed from the Current Population Survey. The literature suggests that less than five percent of the eligible population participated in JTPA in each year (see the discussion in Heckman and Smith, 1999), which allows wide scope for cream skimming. Both papers find modest evidence of cream skimming. In particular, the Anderson et al. (1993) analysis of program participation and post-program job placement suggests that if eligible persons participated at random, the placement rate would have been 61.6 percent rather than 70.7 percent, a fall of 9.1 percentage points.

Heckman and Smith (2004) address the issue of self-selection versus selection by program staff using data from the Survey of Income and Program Participation (SIPP) on JTPA eligibles combined with data from the National JTPA Study. They break the participation process for JTPA into a series of stages – eligibility, awareness, application and acceptance, and participation – and look at the observed determinants of going from each stage to the next. They find that some differences between program eligibles and participants result primarily from self-selection at stages of the participation process, such as awareness, over which program staff have little or no control. The evidence in Heckman and Smith (2004) suggests that while cream skimming may be empirically relevant, comparing the eligible population as a whole to participants likely overstates its extent, and misses a lot of substantive and policy-relevant detail.

The paper by Heckman, Smith, and Taber, (1996) presents a contrasting view. They use data from the Corpus Christi, Texas SDA, the only SDA in the National JTPA Study for which reliable data on all program applicants are available for the period during the experiment. In their empirical work, they examine whether those applicants who reach random assignment (i.e., were selected to participate in the program) differ from those who do not in terms of both predicted outcome levels (earnings in the 18 months after random assignment) and predicted program impacts (projected into the future and discounted). The authors find strong evidence of negative selection on levels combined with weak evidence for positive selection on impacts. They attribute the former to a strong "social worker mentality" toward helping the hard-to-serve among the eligible that was evident in interactions with program staff at the Corpus Christi site.

The Workforce Investment Act (WIA) program offers an interesting contrast to JTPA because the WIA performance standards are not adjusted by a regression model, and they therefore do not hold programs harmless for the characteristics of their participants. Because programs now have stronger incentives to enroll individuals with few barriers to employment, we would expect to observe enrollment shift toward this group. An internal (U.S. Department of Labor, 2002) study finds that this is precisely what appears to be occurring, at least in the area scrutinized:

> *A brief survey of States by our Chicago Regional Office indicated that WIA registrations were occurring at only half the level of enrollment achieved by JTPA. While some of this may be due to start up issues, there are indications that the reduced registration levels are due to a reluctance in local areas to officially register people in WIA because of concerns about their ability to meet performance goals, especially the "earnings gain" measure. It appears that local areas in these States are selective in whom they will be accountable for. Some local areas are basing their decisions to register a person on the likelihood of success, rather than on an individual's need for services.*

A study by the U.S. General Accounting Office (GAO, 2002) confirms these problems. The GAO report, based on a survey of 50 states, indicated "many states reported that the need to meet performance levels may be the driving factor in deciding who receives WIA-funded services at the local level."

Overall, the literature provides modest evidence that program staff responded to the incentives provided by the JTPA performance standards system to choose participants likely to improve their measured performance whether or not they benefited from program services, and studies of the implementation of WIA indicate that, if anything, the situation has been exacerbated by the performance management system used for WIA. At the same time, the evidence from the Corpus Christi SDA indicates that staff concerns about serving the hard-to-serve could trump the performance incentives in some contexts.

What are the impact implications of cream skimming? A number of studies have examined the efficiency implications of cream skimming by estimating the correlation between performance measures and program impacts. Barnow and Smith (2004) summarize the evidence from the seven studies that comprise this literature. The seven papers examine a variety of different programs, ranging from the Manpower Development and Training Act (MDTA) program of the 1960s to the Job Corps program of today. Most rely on experimental data for their impact estimates. With one exception (Zornitsky et al., 1988), the findings are negative or mixed regarding the relationship between outcome-based performance measures of the type typically used in employment and training programs and program impacts. The (Zornitsky et al., 1988) findings refer to a program, the AFDC Homemaker-Home Health Aide Demonstration, which differs from programs such as JTPA and WIA in that it provided a homogeneous treatment to a relatively homogeneous population. Taken together, the literature clearly indicates that, in the context of employment and training programs, commonly used performance measures do *not* improve program impact by inducing service to those who will benefit most. At the same time, the literature indicates that cream skimming likely has a very small effect, if any, on program earnings impact.

Effects of Performance Incentives on Services Provided.[11] At least two papers examine the effect of performance incentives on the types and duration of services offered in an employment and training program, holding constant the characteristics of persons served.

[11] The effects of cost standards on services are covered in a later section.

Marschke's (2002) analysis uses the variation in performance incentives facing the training centers in the National JTPA Study to identify the effects of performance incentives on the types of services received by JTPA participants. Marschke (2002) finds evidence that changes in the performance measures employed in JTPA led SDAs to alter the mix of services provided in ways that would improve their performance relative to the altered incentives they faced. In some cases, these changes led to increases in efficiency, but in others they did not. Marschke (2002) interprets his evidence as indicating that SDAs' service choices are responsive at the margin, but that existing performance measures do a poor job of capturing program goals such as maximizing the (net) impacts of the services provided.

More recently, Courty and Marschke (2004) demonstrated that the JTPA performance management system affects the duration of training for some participants because program managers manipulate the duration of services for some participants to be able to count them on their performance measures for a specific program year. Courty and Marschke (2004) find that these manipulations reduced the overall mean impact of the employment and training services provided by JTPA.

Relationship between Performance Measures and Program Impact. Performance measures for a program may be of intrinsic interest, or they may be a proxy for some underlying factor of interest that is not easy to measure on a relatively quick and inexpensive manner. For example, Blalock and Barnow (2001) note that programs may wish to use program impact as a performance measure, but accurately measuring impact requires many years and the presence of a randomly assigned control group or a carefully selected comparison group. Because this is not generally compatible with obtaining quick, inexpensive measures, programs often rely on proxy measures such as post-program earnings or the pre-post change in earnings. If the goal is to have performance measures serve as a proxy for impact, then it is necessary to assess how well the types of measures that are practical and have been used for the JTPA and WIA programs correspond with program impact.

Two studies have explored this issue for JTPA in recent years, and another study looked at the Job Corps. The studies by Barnow (2000) and by Heckman, Heinrich, and Smith (2002) both made use of the fact that the National JTPA Study provided experimental impact findings in 16 local areas and included the data needed to construct performance measures similar to those used by ETA. However, the approach used to measure performance does not include a control group or even a comparison group, so it is not surprising that the performance measures used are at best weakly correlated with program impact.[12]

The recent evaluation of the Job Corps that was based on a classical experimental design provided Schochet and Burghardt (2008) with an opportunity to analyze how closely the Job Corps' performance standards track the program's impacts. Job Corps is a primarily residential program for highly disadvantaged out-of-school youth. Schochet and Burghardt indicate that during the evaluation period, program years 1994 through 1996, the performance measures included eight measures in three broad areas: (1) program achievement (reading and math gains, GED attainment rate, and vocational completion rate), (2) placement measures (placement rate, average wage at placement, and the percentage of placements related to training), and (3) quality/compliance measures (ratings of federal monitors). Because of the random assignment

[12]A related problem is that performance measures must use short-term post-program earnings to measure performance, but the impact of a program is best measured over a longer period. Barnow and Smith (2004) review the literature on the relationship between short-term earnings impacts and long-term impacts, and they find that most studies find a very weak relationship between the two.

used to assign treatment status, impact can be estimated as the difference between treatment and control group values on the outcome measures. Schochet and Burghardt (2008) compared program impacts for Job Corps centers ranked in each third of the performance distribution. They concluded "Our results indicate that at the time of the National Job Corps Study, measured center performance was not associated with impacts on key education, crime, and earnings outcomes."

Strategic Responses to Performance Incentives. In addition to the substantive responses to performance incentives considered above, in which local programs changed what they actually did, local programs can also attempt to change their measured performance without changing their actual performance. This behavior is referred to as a strategic response, or as "gaming" the performance system. Regardless of their differing goals, all types of organizations have an incentive to respond strategically to performance incentives, provided the cost is low, as doing so yields additional resources to further their own goals. The literature provides clear evidence of such gaming behavior under JTPA.

One important form of strategic behavior under JTPA was the manipulation of whether or not participants were formally enrolled. Under the JTPA incentive system (and WIA as well), only persons formally enrolled counted toward site performance. In addition, for the first decade of JTPA's existence, local programs had substantial flexibility in regard to when someone became formally enrolled. Clever SDAs improved their performance by basing enrollments on job placements rather than the initiation of services. For example, some SDAs boosted performance by providing job search assistance without formally enrolling those receiving it in the program. Then, if an individual found a job, the person would be enrolled, counted as a placement, and terminated, all in quick succession. Similarly, SDAs would send potential trainees to employers to see if the employer would approve them for an on-the-job training slot; enrollment would not take place until a willing employer was found.

There are several pieces of evidence regarding the empirical importance of this phenomenon. The first is indirect, and consists of the fact that DOL found it enough of a problem to change the regulations. Specifically, in 1992, the Department of Labor required that individuals become enrolled once they received objective assessment and that they count as a participant for performance standards purposes once they received any substantive service, including job search assistance.

Other evidence comes from the National JTPA Study. As part of their process analysis of the treatments provided at the 16 SDAs in the study, Kemple, Doolittle, and Wallace (1993) conducted interviews of non-enrolled members of the experimental treatment group at 12 of the 16 sites. These results (available on Table 3.2 of their report) show that 53 percent of non-enrolled treatment group members received services, most often referrals to employers for possible on-the-job training (36 percent of all non-enrollees) and job search assistance (20 percent of all non-enrollees). They report that: "… most of the study sites enrolled individuals in classroom training when they attended their first class or in OJT when they worked their first day." There is also evidence that this type of behavior has continued under WIA. The U.S. General Accounting Office (2002, p. 14) notes that: "All the states we visited told us that local areas are not registering many WIA participants, largely attributing the low number of WIA participants to concerns by local staff about meeting performance levels."

The flexibility of JTPA also allowed strategic manipulation of the termination decision. Because performance standards in JTPA were based on exiters, SDAs had no incentive to terminate individuals from the program who were not successfully placed in a job. By keeping them on the rolls, the person's lack of success would never be recognized and used against the SDA in measuring its performance. As the Department of Labor explains in one of its guidance letters, "Without some policy on termination, performance standards create strong incentives for local programs to avoid terminating failures even when individuals no longer have any contact with the program."[13]

Problems with local programs retaining participants on the rolls long after they stopped receiving services go back to the days of JTPA's predecessor, the Comprehensive Employment and Training Act (CETA). In one of their guidance letters, the Department of Labor observed that "monitors and auditors found that some participants continued to be carried in an 'active' or 'inactive' status for two or three years after last contact with these programs." For Title II-A of JTPA, DOL limited the period of inactivity to 90 days, although some commentators suggested periods of 180 days or more.[14]

Employment and Training Administration Experience with Efficiency Measures. The Employment and Training Administration also has previous experience with efficiency standards under WIA's predecessor, the Job Training Partnership Act (JTPA). Under the original JTPA statute, Section 106(b)(4) required that efficiency measures be prescribed for the JTPA Adult Program and that the efficiency measures be related to the outcome measures used.

[13]See TEIN 5-93 available at: http://wdr.doleta.gov/directives/corr_doc.cfm?DOCN=770

[14] Pascal Courty and Gerald Marschke conducted several studies that verify gaming behavior by local programs participating in the National JTPA Study. See Barnow and Smith (2004) for a review of these studies.

The National Commission for Employment Policy (NCEP) funded an evaluation of the effects of JTPA performance standards on participants, services, and costs.[15] The study included quantitative statistical analysis of JTPA Annual Status Report (JASR) data linked to data on the characteristics of local program areas, as well as qualitative analysis based on interviews with 30 local programs and 87 service providers in eight states.

For the most part, the study found that the JTPA performance standards had the desired effects of holding programs harmless for differences in participant characteristics and local economic conditions. However, the study found that the cost standards had intrinsic problems and created some undesirable effects on participants served:

> *This evaluation found that the federal standards for the entered employment rate and wage rate for adults generally did not have unintended effects on clients or services. ...The federal cost standards, however, had the most unintended effects and were the least comparably measured of all the federal performance measures. The evaluation found that SDAs in states that placed more weight on the federal cost standard tended to serve fewer hard-to-serve clients and that [local areas] concerned about exceeding the cost standards tended to design less intensive services. At the same time, this evaluation found serious measurement problems with the cost standards. We found large differences in the extent to which [local programs] were leveraging JTPA funds, either by using funds from other programs to help fund JTPA Title II-A programs or by using service providers that had alternative funding sources. As a result, it is difficult to compare the cost of services received by JTPA participants across [local programs]. [Emphasis in original document] (p. 5).*

Based on their findings from both the quantitative and qualitative components of the study, the authors recommended that alternatives to the cost measures be explored. The authors note that as a result of concern about the unintended impacts of the cost standards, ETA set more lenient cost standards in PY 1988, but they conclude that this policy change would not eliminate

[15]See Dickinson et al. (1988).

the disincentive problems in states that emphasize exceeding rather than meeting standards. In response to the research findings, the NCEP made a number of recommendations for changing the statutory provisions of JTPA dealing with performance standards. Taking note of the study's findings regarding the undesirable incentives and comparability of cost issues, the Commission's first recommendation was:

> *The Commission recommends that Section 106(b)(4), which requires the Secretary [of Labor] to prescribe performance standards relating gross program expenditures to various performance measures, be amended to direct that cost-efficiency be monitored by states.*

In August 1992, the JTPA statute was amended, and the amendments repealed the federal requirement for efficiency standards and prohibited Governors from using efficiency standards in making awards to local areas. The Workforce Investment Act has no prohibitions against the use of cost standards, and in response to requests by OMB, the U.S. Department of Labor currently has a contractor exploring the use of cost measures for 11 workforce programs administered by the Department's Employment and Training Administration.

5. Conclusions and Lessons for the European Commission

This final section of the paper presents conclusions based on the research and lessons I suggest for countries about to introduce a performance measurement system for its workforce investment programs. The primary conclusions and lessons are:

- **Do not confuse performance measurement with program evaluation.** Performance measurement is used as a management tool, and it cannot and should not be expected to serve as a substitute for program evaluation. Performance must be monitored on a continuous basis to assess whether key elements of the program are being implemented as planned and if immediate program outcomes are consistent with the long-term results expected. As a management tool, performance measurement should provide quick feedback on the operation of a program, but in most cases, performance measures cannot and do not measure program impacts.

- **There are often good reasons to adjust performance standards to take account of program goals, participant characteristics, and environmental conditions.** Performance is generally a function of many factors, so it is likely that programs in different locations will vary in important ways that can affect their performance. U.S. programs that use adjustment mechanisms refer to the adjustments as “leveling the playing field,” an effort to judge programs in different circumstances appropriately. When the Workforce Investment Act (WIA) abandoned the statistically-based adjustment procedures used for the predecessor JTPA program, the states and local governments indicated strong dissatisfaction with the new approach.

- **Programs Need not Have the Same Performance Measures or Standards.** In the United States, the Office of Management and Budget (OMB) has attempted to impose common measures on all programs with a workforce orientation. The programs often differ, however, in significant ways, and there is no reason why programs with different participants, activities, and/or economic conditions should necessarily have identical measures. For example, the Senior Community Service Employment Program, sometimes referred to as the Older Worker program, provides community service opportunities to poor older individuals who would like to work. It can be debated whether placement in an unsubsidized job is a good measure of performance for the program, but it is very unlikely that if the entered employment rate is used a measure that the standard should be the same for programs serving customers with fewer barriers to employment.

- **Be Cautious in Establishing Performance Measures with Large Rewards and/or Sanctions.** This is an extremely important lesson from the literature on U.S. workforce investment program performance management research. The literature on performance management for U.S. workforce programs clearly indicates that the measures are sometimes only weakly related to outcomes of interest such as program impact and that state and local programs can manipulate their data to raise their measured performance without actually increasing the value of what they do—in short they spend resources trying to look good instead of doing good. Worse, there is strong evidence that programs sometimes engage in cream-skimming and reduce their services to those most in need. Thus, for workforce programs it is wise to avoid “high stakes” performance measures if the programs can behave strategically to affect their measured performance.

- **Including Measures to Encourage Efficiency Can Be Useful, but the U.S. Experience Indicates that too Much Emphasis on Efficiency Can Lead to Programs Avoiding Customers Who Require Expensive Service Strategies and to too Much Emphasis on Less Expensive Service Strategies.** Research on the use of cost measures for U.S. workforce investment programs indicates that in the past there was widespread agreement that cost measures led to deleterious consequences. Efficiency is, however, a very important goal, particularly when past studies have indicated that only a small fraction of those eligible for workforce programs can be served at current budget levels. Thus, the current concern about taking cost into consideration when measuring the performance of workforce programs is appropriate. The question is how best to balance the desire the need to use resources efficiently with the knowledge that placing too much emphasis on cost issues can lead to providing the wrong mix of services. I would recommend monitoring efficiency and discussing efficiency with those who perform poorly on such measures, but given the negative experiences with the use of cost measures for workforce programs in the United States, I would recommend against establishing formal efficiency measures.

- **Performance Management Is Still in a Formative Stage; Legislation Should Not Be Overly Prescriptive on the Measures, Standards, and Incentives. Deliberations on the Structure of the Performance Management System Should Include Input by All the Relevant Stakeholders.** When performance management was first introduced in the U.S. workforce investment system, meetings involving representatives of states, localities, training providers, academics, and others were held over the course of several years before a formal system with rewards and sanctions was implemented. This process paved the way for statistical models to be used to adjust for variations across states and local areas. Over time, these work groups explored changes, such as using longer-term measures and using administrative data rather than survey data. The inclusionary process helped lead those involved to have a voice in the system and led them to support the resulting system. In contrast, when the performance management system was changed significantly for the implementation of the WIA program with little if any discussion with stakeholders, there was a strong rejection of the new system. Performance measures are much less likely to drive performance in the manner intended if those being graded consider the system grossly unfair.

.

References

Kathryn Anderson, Richard Burkhauser, Jennie Raymond, and Clifford Russell (1992). "Mixed Signals in the Job Training Partnership Act." *Growth and Change*. 22(3): pp. 32-48.

Kathryn Anderson, Richard Burkhauser, and Jennie Raymond (1993). "The Effect of Creaming on Placement Rates under the Job Training Partnership Act." *Industrial and Labor Relations Review*. 46(4): pp. 613-624.

Burt S. Barnow (2000). "Exploring the Relationship Between Performance Management and Program Impact: A Case Study of the Job Training Partnership Act." *Journal of Policy Analysis and Management*, Winter.

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

Burt S. Barnow and Jeffrey Smith (2004). "Performance Management of U.S. Job Training Programs: Lessons from the Job Training Partnership Act." *Public Finance and Management*, Vol. 4, No.3.

Ann B. Blalock and Burt S. Barnow (2001). "Is the New Obsession with 'Performance Management' Masking the Truth About Social Programs?" with Ann B. Blalock in *Quicker, Better Cheaper? Managing Performance in American Government*, Dall Forsythe Editor. Albany, NY: Rockefeller Institute Press.

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

Katherine Dickinson, Richard W. West, Deborah J. Kogan, David A. Drury, Marlene S. Franks, Laura Schlictmann, and Mary Vencill (1988). *Evaluation of the Effects of JTPA Performance Standards on Clients, Services, and Costs.* Washington, DC: National Commission for Employment Policy Research Report No. 88-15.

James J. Heckman, Carolyn Heinrich, and Jeffrey Smith. (2002). "The Performance of Performance Standards." *Journal of Human Resources* 37(4):778-811.

James Heckman and Jeffrey Smith (1999). "The Pre-Programme Dip and the Determinants of Participation in a Social Programme: Implications for Simple Programme Evaluation Strategies." *Economic Journal*. 109(457): pp. 313-348.

James Heckman and Jeffrey Smith (2004). "The Determinants of Participation in a Social Program: Evidence from JTPA." *Journal of Labor Economics*. 22(2): pp. 243-298.

James Heckman, Jeffrey Smith and Christopher Taber (1996). "What Do Bureaucrats Do? The Effects of Performance Standards and Bureaucratic Preferences on Acceptance into the JTPA Program." In *Advances in the Study of Entrepreneurship, Innovation and Economic Growth*, Volume 7. Ed. Gary Libecap. Greenwich, CT: JAI Press, pp. 191-218.

Job Training Partnership Act Advisory Committee (1989).. *Human Capital: JTPA Investments for the 90's*. Washington, DC: U.S. Department of Labor.
James Kemple, Fred Doolittle, and John Wallace (1993). "The National JTPA Study: Site Characteristics and Participation Patterns." New York, NY: Manpower Demonstration Research Corporation.

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

Gerald Marschke (2002). "Performance Incentives and Bureaucratic Behavior: Evidence from a Federal Bureaucracy." Albany, NY: Working paper, Department of Economics, SUNY Albany.

Peter Z. Schochet and John A. Burghardt (2008). "Do Job Corps Performance Measures Track Program Impacts?" Journal of Policy Analysis and Management. 27(3): pp. 556-576.

Social Policy Research Associates. (1999). *Guide to JTPA Performance Standards for Program Years 1998 and 1999*. Menlo Park, CA: Social Policy Research Associates.
U.S. Department of Labor (2002). "Summary Report on WIA Implementation." Washington, DC: Employment and Training Administration.

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.

U.S. General Accounting Office, (1982). *An Overview of the WIN Program: Its Objectives, Accomplishments, and Problems.* Washington, D.C.: GAO/HRD-82-55.

U.S. General Accounting Office (2002). Workforce Investment Act: Improvements Needed in Performance Measures to Provide a More Accurate Picture of WIA's Effectiveness. Washington, DC: GAO-02-275.

Jeffrey Zornitsky, Mary Rubin, Stephen Bell, and William Martin (1988). "Establishing a Performance Management System for Targeted Welfare Programs." National Commission for Employment Policy Research Report 88-14. Washington, DC: National Commission for Employment Policy.