Making Government Accountable:

Lessons from a Federal Job Training Program

Pascal Courty

London Business School

Gerald Marschke[1]

University at Albany, State University of New York

July 2003

Abstract

We describe the evolution of a performance measurement system in a government job-training program. In this program, a federal agency establishes performance measures and standards for sub-state agencies. We show that the performance measurement system's evolution is at least partly explained as a process of trial-and-error, characterized by a feedback loop: the federal agency establishes performance measures, the local managers learn how to game them, the federal agency learns about gaming and reformulates the performance measures, leading to possibly new gaming, and so on. The dynamics suggest that implementing a performance measurement system in government is not a one-time challenge but benefits from careful monitoring and perhaps frequent revision.

Gerald Marschke is Assistant Professor in the Department of Public Administration and Policy and in the Department of Economics at the University at Albany, State University of New York. His research interests include organizational incentives and performance, research and development policy, and the economics of innovation and technology. **Email: marschke@albany.edu.**

Pascal Courty is Professor of Economics in the Department of Economics at the European University Institute. He is on leave from the London Business School. His research focuses on contract theory with applications to the design of incentives in organizations and firm pricing policies. **Email: Pascal.Courty@IUE.it**

Introduction

Over the past decade and a half, interest in developing and implementing performance measurement systems in the public sector has grown as many policymakers and analysts now believe that such systems can improve accountability and management (Osborne and Gaebler 1992; Gore 1993). Greater use of performance measurement systems in the public sector has also received support in academic circles (National Academy of Public Administration 1991; Wholey and Hatry 1992; Bouckaert 1993; Kravchuk and Schack 1996). By focusing on objectives through performance measures rather than on bureaucratic inputs through monitoring, supervision, and rules, it is argued, such systems encourage local government workers to exploit their familiarity with conditions on the ground and to use their initiative.[2] This argument is at the core of the Government Performance and Results Act (GPRA) of 1993, which arose from the most recent in a long line of high-profile government reinvention campaigns. GPRA requires federal agencies to formulate measures of performance and set performance goals to improve public accountability and permit scrutiny by congressional oversight committees and the public.

This paper chronicles the evolution of a performance measurement system used by a federal oversight authority to evaluate local government agencies that administer a large federal job-training program. We follow the development of this system from the early 1980s through the present, as dictated first by the Job Training Partnership Act (JTPA) of 1982 and later the Workforce Investment Act (WIA) of 1998. In both incarnations, the federal authority has been the U.S. Department of Labor (DOL), which has overseen sub-state job training facilities. Many analysts regard this job training program's performance measurement system as a success.[3] Nonetheless, as in other government programs that have incorporated performance measurement, the performance measures in this program are imperfect proxies of the program's objectives.

Because these proxies fail to capture all dimensions of the local agency's contribution to the program objective, as the evidence reported here and elsewhere shows, these measures sometimes elicit the wrong behavior.

The focus of this paper is a sequence of legislative and regulatory modifications to the JTPA/WIA performance measurement systems. We present evidence that these modifications are at least in part attempts to cope with undesirable responses that the designers of the performance measures did not foresee. The evidence suggests that performance measures elicit unanticipated responses because line workers and their managers gain a superior understanding of how to influence these measures. Managers and workers acquire through their day-to-day operation of their programs an expert's knowledge of the levers available to manipulate performance outcomes. Because the designers of the performance measures are remote from the everyday operations of the agencies they oversee, they lack this knowledge. This information asymmetry means that the designers of performance measures cannot anticipate all behavioral responses ex ante. The designers learn only *after* the performance measure is in place how well aligned the performance measure is with the objective of the program. Consequently, the performance measurement system must evolve according to a dynamic of trial-and-error, in which the designers try out performance measures, observe agencies' responses, and then modify or discard the measures, possibly leading to additional rounds of responses and counter-responses. That the mechanism for developing performance measurement systems must be an evolutionary one is a point we have not seen emphasized in the literature.

Numerous scholars have noted the difficulty of establishing measures of performance for government agencies, however.[4] Government agencies often cite the identification of performance measures as the greatest challenge in implementing GPRA (U.S. GAO 1997).

Many have attributed the difficulty to the complex or ill-defined nature of goals in the public sector (e.g., Thompson 2000). Performance measures that are not well aligned with the objectives of the organization may encourage agencies to pursue wasteful or destructive policies (see Hatry 1999; Joyce 1993; Hayes 2001; Perrin 1998; Ammons 2001; see Blau 1955, for an early discussion of the unintended consequences of performance measurement). For example, an empirical literature documents dysfunctional responses to performance measures by high school teachers (Jacob and Levitt 2002), health care workers (Goddard et al. 2000 and Dranove et al. 2003), and navy recruiters (Asch 1990).

Researchers have evaluated how well the performance measures typically used in job training programs serve as proxies for the programs' objectives (see Barnow 2000 and Heckman, Heinrich, and Smith 2002; Heckman et al surveys this literature).[5] These researchers have found that JTPA's performance measures---especially its early ones---do not seem very well correlated with the impacts of job training on enrollee earnings or employment. While this literature is pessimistic that these measures give an accurate accounting of caseworkers' contribution to their programs, Heinrich (2002) argues that JTPA authorities can nevertheless use these measures to improve managerial strategies and organizational design. Our findings build on Heinrich's work by showing that JTPA's performance measure designers, at least, do appear to learn from their experience with performance measures and to reformulate performance measures in response to what they learn.

The evolutionary dynamic we observe in JTPA/WIA has at least two important implications. First, the evidence suggests the usefulness, and perhaps the necessity, of an extended period of monitoring and evaluation following the launch of a new performance measurement system. During this period, the designers of performance measures may find it

necessary to make major changes to the performance measures. Past literature has emphasized the costs of dysfunctional behavior caused by misaligned performance measures. We offer evidence that designers of government performance measures do learn how they are misaligned and then take steps to improve them. The slow pace with which these designers identify and correct dysfunctional behavior in the program we study, however, suggests that the system of monitoring and evaluation necessary for dynamically managing performance measurement may also be costly.

Second, our findings emphasize the trade-off between obtaining timely measures of performance and accurate ones. Careful evaluations of government programs typically involve several years of follow-up after an intervention. Providing managers and oversight agencies useful and prompt feedback requires shorter measurement periods (see, e.g., Barnow 1992). This trade-off will be especially important for any public-sector program---such as many that provide human services---whose impacts are long-term. JTPA/WIA evidence demonstrates that not only are measures based on short-term performance unlikely to capture long-term effects, but also that short-term measures are especially vulnerable to manipulation.

The next section describes the organization created under JTPA and its performance measurement system. Here we discuss the trade off that exists between making performance measures accurate, on the one hand, and useful for managers and administrators as well as cheap to produce, on the other. This section then describes the set of performance measures and performance measurement rules that existed at the JTPA's outset, the starting point for the evolution that takes place later. Next we detail how program administrators responded to performance measures and chronicle the sequence of modifications that we argue were an attempt to attenuate the undesirable responses to the measures. We consider an alternative

reform of job-training measurement procedures to reduce these dysfunctional responses. The paper concludes with a discussion of the implications of our findings for performance measurement in government.

A Federal Job-Training Program

Federal involvement in job training for the economically disadvantaged began in the Kennedy administration and has since been modified by a series of congressional acts, the most recent passed in 1998. In 1982, the Job Training Partnership Act transformed the bureaucracy that administered the program in two important ways.[6] First, the program under JTPA became highly decentralized: more than 620 semi-autonomous sub-state training centers administered the program with significant discretion over who to admit and how to conduct the training. Second, and most important for this study, the federal government began to use a loose set of performance measures to influence outcomes. In addition, as an incentive, training centers that performed well received modest budgetary increases.

To understand the relevance of our analysis, it may be worth putting our case study into perspective and contrasting it with other government organizations. Compared with the missions of many public-sector agencies, the mission of job training programs for the economically disadvantaged is fairly straightforward. The act directed training centers to establish programs that help the economically disadvantaged and others "facing serious barriers to employment" to develop skills that would enable them to obtain employment, increase their earnings, and reduce their dependency on welfare programs. Because this job-training program has a clear mission, it offers a natural environment to experiment with explicit performance, and our analysis is relevant to other public organizations that share this characteristic and where explicit

performance measurement has been introduced. After all, some of the behavior we identify in this program has also been found in health, education, and defense organizations. Public organizations that lack a clear mission will face additional challenges in implementing performance measurement to the ones identified in this work, but this does not make the implications of our analysis any less relevant. It only suggests that these additional challenges should be taken into account.

The Challenge of Measuring Programmatic Impacts

Section 106(a) of the Job Training Partnership Act directed DOL to develop performance measures to assess each training center's success in maximizing its "return on investment in human capital" [Section 106(a)].[7] At the training center level, this translates into maximizing the amount of new human capital it generates, net of training expenses. While an enrollee's stock of human capital is not directly observable, any increase should be reflected in that person's labor market earnings. Thus, a measure of the impact of job training on a single enrollee's human capital is the sum of her earnings from the beginning of her training into the future, minus the sum of earnings over the same period had she *not* experienced training.[8]

An important difficulty is that while the costs and earnings of program enrollees may be measurable, their earnings in the absence of training are not directly measurable. Producing reliable estimates of the counterfactual earnings is costly and takes time.[9] For example, an experimental evaluation of JTPA involving only 16 of 620 training centers took seven years to complete and, according to one estimate, cost over $21 million (Smith 1996).[10]

As a compromise, DOL attempted to capture earnings impacts using performance measures based on the employment status of enrollees after training. For example, an important

measure early in the program was the *employment rate at termination*, computed as the fraction of enrollees who were employed on the date they officially completed---or were terminated from---the program. Additional measures were based on enrollees' wage rates at training end. In other words, DOL chose measures that captured only outcome *levels,* whereas the objective of job training was to facilitate outcome *changes*. Thus, the employment measure counted the number of employed enrollees, whether or not training was responsible for the employment. DOL partly addressed this problem by basing its evaluation of performance and the budgetary award not on the performance outcomes alone, but on a comparison of performance outcomes with a center-specific performance standard---a prediction of the training center's performance. The prediction was based partly on factors outside of the training center's control, such as the health of the local labor market, but also on the characteristics of the enrollees the training center served (see Courty and Marschke 2003b).[11] DOL's hope was that the difference between the outcome and the predicted outcome would capture the training center's contribution to the enrollee's labor market skills. Such adjusted performance standards are not uncommon in the construction of performance measures (Stiefel et al. 1999).

Measuring Long-term Impacts with Short-term Outcomes

The objective of job training is to produce long-term effects in earnings by imparting skills that enrollees can use for the rest of their lives. Ideally, one would therefore want a measure of an enrollee's earnings and employment over the rest of her working life.

The problem of measuring performance when managerial decisions produce effects far into the future challenges designers of performance measures in many other contexts. Consider, for example, the problem of evaluating a chief executive officer in the private sector, the subject

of much of the contracting literature in economics. The CEO makes decisions that have both near- and long-term consequences for the prospects of the firm. Evaluations of the CEO at the end of the year based on the firm's year-end net earnings do not capture the long-term effects of the CEO's efforts. Moreover, evaluating the CEO based on year-end net earnings may lead the CEO to make decisions that boost current-period net earnings at the expense of the long-term health of the firm. For example, a CEO may avoid lucrative projects with high up-front costs that depress current period net earnings because such projects do not start generating revenue for several years. If the firm is public, however, ownership can evaluate the CEO based on changes in the firm's share price, which capitalizes both the short and long-term effects of CEO decisions.[12]

Because ready-to-use summary measures of a job training program's long-term impacts do not exist, to accurately capture long-term effects, job-training personnel must track enrollees' employment experience long after training ends.[13] Stretching the post-training evaluation periods, however, increases the cost of performance measures and makes them less useful as management, evaluation, and motivational tools. In practice, in government agencies whose activities produce effects that play out over time, the timing of measurement---both the date performance measurement commences and its duration----must strike a balance between producing immediate feedback and accurately measuring performance.[14]

Initially, JTPA program designers chose measures of short-term labor outcomes based on an enrollee's employment status and hourly wage on the day training centers chose to terminate the enrollee.[15] An enrollee's labor market status could swing wildly from week to week as she moved in and out of employment---the job turnover rate among persons served by this program was high---even while the long-term value of her labor market skills produced by job training

remained constant. This meant that even persons with infrequent contact with the labor market could appear as successes if the measurement date happened to coincide with an employment spell. Generally, when the period over which performance is evaluated is short (initially the measurement window in JTPA was a point in time), the choice of when relative to the end of training measurement is taken plays a decisive role in determining whether an enrollee is recorded as a success or failure. The longer the measurement period, however, the more likely this variation smoothes out and the less important the moment that measurement starts.

Thus, when should performance be measured? Applicants enroll throughout the year in job-training services of different lengths, and they may receive several of these services depending on the outcome of training. Often job-training services do not have a well-defined structure with a clear beginning and end. In the course of training, training centers acquire information about participants and the effectiveness of training decisions. Allowing caseworkers the discretion to make training decisions dynamically, including, for example, the decision *when* to terminate an enrollee, allows them to use this information to improve enrollees' outcomes.

Without a definite beginning and end to training, however, there is no natural measurement date for the labor market outcomes.[16] This lack of a natural measurement date is central to understanding the effectiveness of performance measurement in JTPA. The designers of the performance measures ultimately settled on a performance measurement scheme that permitted caseworkers some discretion in the measurement date.

Evolution of Performance Measurement in JTPA

In this section, we consider the changes the performance measures have undergone. Our approach is to selectively focus on the few changes that are regarded as some of the more

important changes in the measurement system and that have therefore received attention among policy analysts, and to carefully investigate why these changes have occurred. We will argue that the changes we have selected can be understood as attempts by the U.S. Department of Labor to address deficiencies discovered only after the performance measures were implemented. We recognize, however, that the rationale we favor is not the only candidate explanation. Changes in the political agenda of DOL, pressure from lobby groups, or even random responses to bureaucratic shocks may have instead triggered the changes in the measurement system we identify. Although these alternative explanations cannot be ruled out, we do not find any evidence in their support in the specific changes we consider.

The source material we use to understand how the performance measurement system in the federally funded job-training program has evolved include congressional, DOL, and individual state JTPA documents as well as results from the literature investigating training center responses. We augment the above with a survey (described below) administered to managers of job training centers that details some of their responses to reforms of performance measures in the late 1980s and early 1990s.

Origins of the 90-day Rule

At JTPA's outset, performance was judged by training center employment and wage rates at termination.[17] Thus, at the beginning of JTPA, the termination date corresponded to the measurement date. JTPA-style performance measures were first implemented without financial incentives in the last years of JTPA's predecessor program, created under the Comprehensive Employment and Training Act (CETA), to prepare the way for their use with budgetary incentives in JTPA. CETA gave training centers discretion over when to terminate enrollees,

allowing them to uncouple the accounting end of training from its actual end, thereby affording them some flexibility in developing and implementing training strategies. However, this uncoupling provided the case manager with an incentive and enough discretion to "hide" bad performance. That is, training centers could postpone terminating enrollees who were unemployed, or if employed, had low wages. This is apparently what job training centers did. DOL observed that "some [training program] participants continued to be carried in an 'active' or 'inactive' status for two or three years after last contact with these programs."[18]

To avoid this problem in JTPA, DOL required training agencies to measure performance on the termination date, but also to terminate any participant who had not received any training services for 90 days. As DOL itself states, this 90-day rule "was designed to provide programs some latitude in securing jobs for their customers, while at the same time, providing for more equitable performance assessments by precluding [training centers] from retaining indefinitely participants who are "unsuccessful." 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."[19] Thus, the 90-day measurement rule was a compromise between allowing caseworkers discretion to craft effective training strategies and generating undistorted measures of performance.

Gaming the 90-day Rule

The 90-day rule did not eliminate training centers' ability to manipulate termination-based performance outcomes, however. The discussion in this subsection, which shows how training centers used their discretion over the timing of termination to distort performance, borrows from Courty and Marschke (forthcoming).

In the first decade of JTPA, the employment rate at termination was the most important measure in determining a training center's award. A training agency's employment rate at termination for the fiscal year was computed as the fraction of enrollees who terminated during that fiscal year who were employed on the date of their termination. At the beginning of the next fiscal year, the slate was wiped clean, and performance measurement began anew. Training centers that exceeded their assigned standards for the year received higher awards. Thus, if in an arbitrary fiscal year a training center's standard for the employment rate at termination was 64 percent and its overall employment rate for that fiscal year was 65 percent, the training center would receive a budgetary award to begin the following fiscal year.

The following discussion describes training centers' responses to the 90-day rule. Under JTPA, a job-training applicant arrived at a training center for an assessment by training center staff. If she was enrolled, the staff assigned her to training services.[20] At the end of her training, the training center made the decision whether to terminate the enrollee and report her labor market outcome, or to postpone termination in hopes that the outcomes would improve.

The evidence shows that enrollees who were employed on the last day of training were (on average) terminated then,[21] and that enrollees who were not employed at the end of training but became employed within the next 90 days were (on average) terminated as soon as employment commenced, and finally that enrollees who did not obtain employment were (on average) terminated on the last day of the 90-day window. This behavior reflects precisely the measurement strategy that maximizes performance scores. Courty and Marschke show that training centers would have produced an employment rate outcome on average 20 percent lower if they had been required to terminate enrollees (and report their performance outcomes) on the day enrollees finished training.

This behavior, of course, may also have been consistent with strategies that promote the development of enrollee human capital. Courty and Marschke, however, reject this hypothesis, for two reasons. First, the termination strategy that they documented implies that a training center usually arrived at the end of the fiscal year with an inventory of idle, unemployed enrollees on its books. The training agency would then decide in which fiscal year to terminate the unemployed enrollees. If the center found itself either comfortably above or hopelessly below its standard, it could enhance its odds of winning an award in the next fiscal year without jeopardizing its award in the current year by terminating most or its entire inventory. If the training center found itself above but close to the standard, it could increase its award in the present year by postponing termination until the following year. Courty and Marschke show that how a training center disposed of this inventory at year-end was sensitive to the training center's year-to-date performance. In particular, a training center typically terminated an unemployed enrollee at year-end more quickly following the end of her training if the training agency was doing either very well or very poorly relative to the employment standard. This behavior is difficult to reconcile with any simple story of maximizing human capital.

Second, the study finds evidence that this kind of gaming behavior consumed program resources. In particular, they show that the training centers that exhibited the greatest amount of this behavior demonstrated the smallest per capita earnings gains. They interpret this finding as evidence that the kind of monitoring of the caseload necessary to optimally time termination diverts resources from training activities.

In sum, after DOL instituted the 90-day rule, employment rates likely better reflected training center's success at producing employment. On the other hand, the 90-day window afforded training centers enough latitude to boost employment rates significantly beyond what

they would have been had employment been measured at training end. In addition to artificially inflating performance outcomes, this kind of strategic behavior may have cost training resources.[22] If DOL were monitoring training centers' performance accounting practices---indeed, training centers were subject to periodic audits---it would likely have discovered the 90-day rule had reduced but not eliminated gaming behavior based on measurement timing. The next round of changes to the measurement system may have further reduced this behavior.

Follow-up Measurement

In 1990 DOL phased out termination-based performance measures at least partly in response to investigations by the General Accounting Office that seemed to show that they induced training agencies to emphasize job placement-oriented services that had no long-term impact on enrollees' skills. To "[promote] effective service to participants and [assist] them to achieve long-term economic independence,"[23] in place of termination-based measures, the DOL introduced a follow-up measure based on the employment state of enrollees three months after the official end of their association with the training centers. In their focus on employment and wages, the follow-up measures are similar to the termination measures. This switch constituted one of the main changes to the JTPA measurement system.

Whether or not this was an intention of DOL, the switch from the termination to the follow-up-based measures meant that training centers had less control over the employment status of the enrollee on the measurement date, which should have reduced the appeal of timing strategies. The new measurement regime included two measures of employment that followed the enrollee over the three months following termination. One measure was based on the number of weeks worked for the entire three-month period following training. Another was based on the

enrollee's earnings throughout this period. As we noted above, the longer the measurement period, the more robust the training center's performance measurement is to when relative to the end of training the measurement period begins. The addition of these measures would have further limited the advantages of attempting to time terminations to coincide with employment.

The switch from termination-based follow-up measures may have provoked several kinds of responses by local decision-makers. First, the follow-up measures may have encouraged training centers to emphasize intensive training services (as opposed to employment-focused services, such as job search) in the hopes of producing larger and longer-lasting impacts on earnings and employment. Second, the follow-up measures may have induced caseworkers to extend their contact with enrollees beyond their termination dates. While JTPA regulations prohibited training centers from providing extensive services such as on-the-job training and vocational education after termination, the regulations did allow them to provide "support" services such as childcare and job search assistance.[24] This provision was designed to encourage caseworkers to keep in touch with enrollees after the end of training to ensure that training strategies succeeded. Moreover, discretion over the choice of post-training services may have improved training centers' ability to produce lasting employment matches, and increased the chances that enrollees would find a new job quickly in the event that their initial placement ended.

To explore this aspect of local decision-makers' response to the advent of the follow-up measures we surveyed the staff of 16 training centers. We chose these training centers because they had participated in an evaluation of JTPA training sponsored by DOL between 1987 and 1989 and thus much was already known about them.[25] Three of the 16 training centers declined to participate in our survey.[26] We distributed written copies of the survey to supervisors,

managers, and training center directors. Two weeks after they received the survey, we contacted them and received their responses by phone.

Of particular interest for this discussion is the part of the telephone survey that addressed training center responses to the follow-up measures (see Table 1). For confidentiality reasons, we do not link the responses to the respondents. Nine of the eleven administrators responding to this part of the survey indicated that after the introduction of the follow-up measures, case managers began monitoring enrollees from termination through the end of the follow-up period. To increase the chances that employment matches lasted until the thirteenth week after termination, some training centers reported that between termination and follow-up they offered additional services such as childcare, transportation, and clothing allowances. Case managers also actively pressed employers to retain the clients until the third month. If the client lost her job, case managers scheduled job-counseling appointments and offered placement services. All training administrators reported that after the third month, they neither offered placement or counseling services nor contacted the client again.

[Table 1 here]

Some of the post-termination activities that the follow-up measures apparently induced may indeed have increased the long-term employment and earnings prospects of enrollees. Other activities, however, such as the provision of childcare and transportation subsidies, appear less directed at raising the skill levels of enrollees and more focused on their measured performance over the follow-up period. To the extent that these activities were performance-focused, a training center's performance reflected not only its success in generating long-lasting

employment relationships but also its success with strategies whose effects were more transitory. If these strategies also diverted resources from other activities that would have increased long-run earnings and employment impacts more, follow-up measures reduced programmatic efficiency. Moreover, if short-term activities had a higher impact on follow-up performance measures than the long-term services these measures were supposed to stimulate, then the follow-up measures failed to redirect training centers' focus toward a long-term vision of training impact. In practice these measures may have simply shifted training centers' attention from the day of termination forward three months.

The Cost Measure

In addition to employment and hourly wage measures, in the early years JTPA training centers faced a cost-based measure that judged the program's managers by how much they spent to produce an employment at termination. The incentives inherent in the cost and employment rate at termination measures were very similar. The cost measure was defined as the training center's program expenditures divided by the number of enrollees who were employed at the end of their training. The JTPA award scheme rewarded managers for achieving low cost measures. Because JTPA training center budgets were fixed (and hence training centers' expenditures were fixed), training centers could only lower their cost outcomes by boosting the number of enrollees employed on their termination dates. That is, holding the number of enrollees constant, the same strategies that increased employment rates at termination reduced cost outcomes. Thus, in time, JTPA officials came to believe that the cost measure, also, was encouraging short run, "quick fix"-type job placement activities in lieu of longer-term activities with more training content.[27] In 1990, when they replaced the termination-based measures with follow-up-based measures,

JTPA officials also phased out the cost measure because "research and experience have shown that the use of cost standards in the awarding of incentives has had the *unintended effect* of constraining the provision of longer-term training programs" (Federal Register, January 5, 1990, italics ours). The cost measure's implementation scenario illustrates the dynamic that is the paper's focus. JTPA's designers instituted the cost measure to give some weight to efficiency in shaping agency behavior and did not entirely foresee local decision makers' responses. Once the designers perceived that this measure's effects were counterproductive, however, they removed it.

The Workforce Investment Act of 1998

Performance measures based on labor market outcomes may have influenced training center behavior in another important and unintended way. Although the performance standard adjustment procedures took into account some demographic characteristics, they did not fully account for the labor market potential of the training center's enrollee population at entry into training.[28] That is, the performance standards' construction meant that training centers were evaluated only partly on their success at increasing the human capital stock of its enrollees. Training centers were at least partly evaluated, therefore, on their ability to select enrollees most able to achieve high levels of employment at high wages, instead of the enrollees most likely to benefit from the program.[29] This behavior, called *cream-skimming*, was anticipated at the outset of JTPA and partly accounts for early opposition to the use of performance measures in JTPA. Evidence of its existence in the academic and non-academic literatures began to appear several years before WIA's enactment (see Heckman, Heinrich, and Smith 2002, for a recent survey of this literature).

Between 1990 and 2000 the JTPA performance measures remained largely unchanged. In 2000, the Workforce Investment Act supplanted JTPA and brought three changes, seemingly improvements to the performance measures.[30] As in JTPA, in WIA most of the performance measures are based on the labor market outcomes of enrollees. However, not all labor market outcomes are measured after services cease, as in the latter-day JTPA. WIA includes a new before-after measure of enrollees' earnings. Measuring employment and earnings at both entry and exit may provide a better estimate of the contribution of job training and address the problem of constructing performance standards that capture this contribution (Heckman et al. 1997). Before-after measures may therefore reduce the incentive managers had under JTPA to cream-skim.[31]

Under WIA, the period of measurement was also extended from three months to six months post-training, further reducing the sensitivity of performance to the measurement start date.[32] Finally, the performance measures under WIA include a measure of "customer satisfaction" produced from post-training surveys of enrollees and their employers. Because enrollees value both the transitory and permanent effects of training, these surveys may capture longer term employment impacts that objective but shorter-term employment measures miss.

Thus, WIA brings further improvements to DOL's performance measures. These improvements appear to address deficiencies uncovered by earlier evaluations and studies. While these changes are not proof that the government designers of job training performance measures are learning from their experience, they are suggestive of it.

Discussion

One can draw at least two implications from this and many previous analyses of JTPA/WIA's performance measurement system. First, workers in government agencies do respond to performance measures. The findings reported here illustrate that training agencies understood the performance measures and responded to them, timing measurement dates to boost their performance outcomes. Even more convincing, we show they changed their behavior when measurement rules changed.[33] Second, as our evidence suggests, the design of effective measurement procedures is especially complicated in a job-training context because there is no natural beginning and ending date to training. Thus, an important lesson is that the challenge in designing effective performance measures for the public sector does not stem solely from the scarcity of good proxies for performance. The effectiveness of measurement rules and accounting procedures depends largely on bureaucrats' freedom to manipulate measurement of performance outcomes. In the context of measuring the contribution of job training to enrollees' skills, the timing of measurement is an important means by which agents manipulate performance outcomes. The timing of measurement thus becomes important in designing the system. Note that the trade-off between allowing caseworkers flexibility in determining the measurement date and accuracy in performance measurement will be present in most case-based human services programs.

Our findings show that the main problem with JTPA measurement rules is that training centers knew when performance outcomes would be measured. Training centers' responses to moving the measurement date beyond the termination date (the follow-up measures) showed that this problem may not be solvable only by changing the timing of measurement, because training centers were still able to anticipate and prepare for the measurement date. Still, extensions of the measurement window, as under the 1990 JTPA reforms and again under WIA, should reduce

some of the strategic behavior that distorts measurement, but at a cost in the timeliness of feedback for JTPA's managers and evaluators.

An alternative measurement rule might divert training centers' attention from the measurement date by concealing from them the date measurement will occur. One such mechanism that might succeed is to measure labor market outcomes on a random date within a fixed window. If the window were large enough, training centers would spread their efforts evenly to maximize employment across the window, which is probably desirable to promote clients' long-term labor market attachment. In fact, randomization is not uncommon in agency settings where one party has superior information that can be used strategically. For example, tax authorities typically randomize the decision to audit tax reports.

Summary and Conclusions

This paper reviewed the evolution of the performance measurement regime in a large federal job training bureaucracy. We identify a process by which system designers appear to learn about and respond to local decision maker responses to the performance measures. This feedback loop suggests the usefulness and perhaps the necessity of having the means to regularly examine performance measurement systems to detect weaknesses and develop improvements. Implementation is not a one-time challenge, but a dynamic one.[34] We suspect this is especially true in areas in the public sector where objectives are less well defined and more multi-dimensional than in job-training programs. We suspect that in most cases in the public sector performance measurement designers cannot expect to anticipate all responses, because local decision makers will gain a superior understanding of how to influence performance measures as they become accustomed to them.

This has important implications for how performance measures should be chosen. Courty and Marschke (2003a) develop a theoretical model of how an organization can manage performance measures when dysfunctional behavior is revealed over time. They show that simply selecting performance measures based on their correlation with the organization's true objective cannot suffice when measures are game-able. Consider the case of a government job-training program where before formulating its first performance measures, the designers observed---perhaps by running an experiment---that training centers that produced high employment rates also tended to produce the highest earnings impacts. Employment outcomes would appear to be a good performance measure. Suppose then that DOL announced to training centers that they would be evaluated based on their employment rates. If gaming, such as by timing measurement strategically, was the low-cost strategy to produce high employment rates, and if such a strategy diverted resources from training efforts, then high employment rates would no longer signal successful producers of earnings impacts. In fact, training centers that generated the highest employment rates might produce the lowest earnings impacts. The model implies that because the game-ability of a performance measure cannot be observed ex ante, the only way to identify a good performance measure is by trying it out. This finding contrasts with arguments in the literature that good performance measures should be highly correlated with the goals of the organization (Heinrich 2002; Heckman and Smith1995; Barnow 2000).

Over the 20 years of performance measurement in JTPA and WIA, the changes in the performance measurement system appear to be improvements that reflect learning from experience. This suggests that performance measurement systems are costly. Clearly, and as noted by others, performance measurement can be costly for the dysfunctional behavior they encourage. But, performance measurement is also costly to the extent that resources must be

committed to their implementation, monitoring, and modification when they seem to fail. These costs may make the use of performance measurement systems uneconomical in many organizations. On the other hand, the JTPA/WIA experience suggests that designers of government performance measures are capable of detecting how performance measures are misaligned with programmatic objectives and then of moving to fix them. The findings suggest that agencies' gaming strategies need not persist, and that performance measures may become more aligned with the objectives of the organization as system designers gain experience.

Table 1: Post-Termination Services Provided after the Introduction of Follow-Up Performance Measures

Training Center	Post-Termination Services (activities performed within the 13-week follow-up period)
1	The training center maintains contact with client until 13th week following termination and helps client find a job if unemployed.
2	No response.
3	The training center maintains contact with client until follow-up period expires.
4	Caseworkers call client at 10th week and if unemployed try to induce client to get job by 13th week.
5	No response.
6	The training center maintains contact with client during 13-week post-termination period.
7	The training center maintains contact with client and firms during follow-up period and engages in problem solving (counsels patience for employers).
8	The training center does not influence retention of client between termination and follow-up.
9	The training center does not influence retention of client between termination and follow-up.

10	The training center calls the client at weeks 2, 6, 10, and 13 after termination to check if the client is "ready for follow-up survey." If client is unemployed, training center schedules counseling appointment and offers additional placement services. After the 13th week, the training center does not contact client again.
11	The training center offers transportation services between termination and follow-up to help client make the transition to employment.
12	The training center offers post-program services to employed terminees between termination and follow-up. Such services include childcare, transportation, and clothing allowances. These services are not offered after the 13th week.
13	The training center maintains contact with employer until 13th week.

Note: This table reports training managers' responses to question 7 in our survey: "Can/Do you influence an employer's retention of a client between termination and follow-up? Can/Do you impose a penalty on the firm for firing/laying off client before follow-up?" The respondent's answers were open-ended. For confidentiality, we do not identify the training centers.

References

Ammons, David N. 2001. *Municipal Benchmarks: Assessing the Local Performance and Establishing Community Standards*. Thousand Oaks, CA: Sage Publications, Inc.

Asch, Beth J. 1990. Do Incentives Matter? The Case of Navy Recruiters. *Industrial and Labor Relations Review* 43(3):89S--106S.

Barnow, Burt. 1992. The Effects of Performance Standards on State and Local Programs. In *Evaluating Welfare and Training Programs*, edited by Charles Manski and Irwin Garfinkel, 277-309. Cambridge, MA: Harvard University Press.

——. 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* 19(1): 118-41.

——. and Jeffrey A. Smith. June, 2002. What Does the Evidence from Employment and Training Programs Reveal About the Likely Effects of Ticket-to-Work on Service Provider Behavior? Working paper, University of Maryland.

Belzil, Christian and Jorgen Hansen. May, 2003. Education and Training over the Lifecycle: The Causal Effect of Accumulated Human Capital on Training Opportunities. Working paper, Concordia University.

Blau, Peter M. 1955. *The Dynamics of Bureaucracy: A Study of Interpersonal Relations in Two Governmental Agencies.* Chicago: University of Chicago Press.

Bouckaert, Geert. 1993. Measurement and Meaningful Management. *Public Performance and Management Review* 17(1): 31-43.

Card, David. 2000. The Causal Effect of Education on Earnings. In *Handbook of Labor Economics*, edited by David Card and Orley Ashenfelter, 1801-63. Amsterdam: North-Holland Publishers.

Coleman, James S. May, 1993. The Design of Schools as Output-Driven Organizations. Manuscript, The University of Chicago.

Courty, Pascal and Gerald Marschke. 2003a. Dynamics of Performance Measurement Systems. *Oxford Review of Economic Policy* 19(2): 268-284.

——. 2003b. Performance Funding in Federal Agencies: A Case Study of a Federal Job Training Program. *Public Budgeting and Finance* 23(3): 22-48.

——. 2004. An Empirical Investigation of Gaming Responses to Explicit Performance Measures. *Journal of Labor Economics* 22(1): forthcoming.

de Lancer Julnes, Patria. 2001. Does Participation Increase Perception of Usefulness? *Public Performance and Management Review* 24(4): 403-418.

Dickinson, Katherine P., Richard W. West, Deborah J. Kogan, David A. Drury, Marlene S. Franks, Laura Schlichtmann, and Mary Vencill. September, 1988. Evaluation of the Effects of JTPA Performance Standards on Clients, Services, and Costs. Research Report No. 88-16, National Commission for Employment Policy.

Doolittle, Fred and Linda Traeger. 1990. *Implementing the National JTPA Study*. New York: Manpower Demonstration Research Corporation.

Dranove, David, Daniel Kessler, Mark McClellan, and Mark Satterthwaite. 2003. Is More Information Better? The Effect of 'Report Cards' on Health Care Providers. *Journal of Political Economy* 111(3): 555-588.

Goddard, Maria, Russell Mannion, and Peter C. Smith. 2000. Enhancing performance in health care: a theoretical perspective on agency and the role of information. *Health Economics* 9: 95-107.

Gore, Al. 1993. *Report of the National Performance Review (The Gore Report).* New York: Random House (Reprint of the U.S. Government Printing Office publication).

Greiner, John M. 1996. Positioning Performance Measurement for the Twenty-First Century. In *Organizational Performance and Measurement in the Public Sector,* edited by Arie Halachmi and Geert Bouckaert, 11-50. Westport, CT: Quorum.

Hatry, Harry. 1980. Performance Measurement Principles and Techniques: An Overview for Local Government. *Public Productivity Review* 4(4): 313-15.

——. 1999. *Performance Measurement: Getting Results*. Washington, D.C.: The Urban Institute Press.

Hayes, Frederick O'R. 1977. *Productivity in Local Government*. Lexington, MA: Lexington Books.

Healy, Paul. 1985. The Effect of Bonus Schemes on Accounting Decisions. *Journal of Accounting and Economics* 7(1):85-107.

Heckman, James J. 2000. Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture. *Journal of Political Economy* 109(4): 674-748.

——, Carolyn Heinrich, and Jeffrey A. Smith. 1997. Assessing the Performance of Performance Standards in Public Bureaucracies. *American Economic Review* 87(2): 389-395.

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

——, Robert J. Lalonde, and Jeffrey A. Smith. 1999. The Economics and Econometrics of Active Labor Market Programs. In *Handbook of Labor Economics, Volume 3A*, edited by Orley Ashenfelter and David Card, 1865-2097. Amsterdam, New York, and Oxford: Elsevier Science, North-Holland.

—— and Jeffrey A. Smith. 1995. The Performance of Performance Standards: The Effects of JTPA Performance Standards on Efficiency, Equity and Participant Outcomes, Working Paper, University of Chicago, Department of Economics.

—— and Jeffrey A. 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): 313-348.

Heinrich, Carolyn. 2002. Outcomes-Based Performance Management in the Public Sector: Implications for Government Accountability and Effectiveness. *Public Administration Review* 62(6): 712-725.

Jacob, Brian A. and Steven D. Levitt. 2002. Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. Manuscript, University of Chicago.

Johnston, Janet W. 1987. *The Job Training Partnership Act: A Report by the National Commission for Employment Policy* (Washington, D.C.: U.S. Government Printing Office).

Joyce, Philip G. 1993. Using Performance Measures for Federal Budgeting: Proposals and Prospects. *Public Budgeting and Finance* 13(4): 3-17.

Kerr, Steven. 1975. On the Folly of Rewarding for A while Hoping for B. *Academy of Management Journal* 18(4): 769-83.

Kravchuk, Robert and Ronald Schack. 1996. Designing Effective Performance-Measurement Systems under the Government Performance and Results Act of 1993. *Public Administration Review* 56(4): 348-358.

Lalonde, Robert J. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review* 76(4): 604-20.

Lynch, Lisa. 1992. Private Sector Training and the Earnings of Young Workers. *American Economic Review* 82: 299-312.

National Academy of Public Administration. 1991. Performance Monitoring and Reporting by Public Organizations. Washington, DC: NAPA.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Winston Lin, George Cave, and Fred Doolittle. 1994. *The National JTPA Study: Impacts, Benefits, and Costs of Title II-A*. Bethesda, MD: Abt Associates Inc.

Osborne, David and Ted Gaebler. 1992. *Reinventing Government.* Lexington MA: Addison-Wesley.

Perrin, Burt. 1998. Effective Use and Misuse of Performance Measurement. *American Journal of Evaluation* 19(3): 367-379.

Smith, Jeffrey A. 1996. A Note on Estimating the Relative Costs of Experimental and Non-Experimental Evaluations Using Cost Data from the National JTPA Study, Unpublished manuscript, The University of Chicago.

Smith, Peter. 1995. On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration* 18(2/3): 277-310.

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

Thompson, James. 2000. The Dual Potentialities of Performance Measurement. *Public Productivity and Management Review* 23(3): 267-281.

Trott, Charles E. and John Baj. 1987. *Development of JTPA Title II-A Performance Standards Models for the States of Region V*. Arlington, VA: James Bell Associates.

U.S. General Accounting Office. 1996. *Executive Guide: Effectively Implementing the GPRA*. Washington, D.C.: GAO/GGD-96-118.

——. 1997. *Measuring for Results: Analytical Challenges in measuring Performance.* Washington, D.C.: GAO/HEHS/GGD-97-138.

Wang, Xiaohu. 2000. Performance measurement in Budgeting: A Study of County Government. *Public Budgeting and Finance* 20(3): 102-118.

Wholey, Joseph. 1999. Performance-Based Management. *Public Performance and Management Review* 22(3): 288-307.

—— and Harry Hatry. 1992. The Case for Performance Monitoring. *Public Administration Review* 52(6): 604-610.

Notes:

[1] Pascal Courty, London Business School, Regent's Park, NW1 4SA, London, UK, pcourty@london.edu. *Contact author:* Gerald Marschke, Department of Public Administration and Policy, Milne 308, University at Albany, State University of New York, Albany, NY 12222, marschke@albany.edu (phone: 518-442-5274; fax: 518-442-5298).

[2] Others have advocated performance measurement for managerial purposes. See Smith (1995), Greiner, (1996), Hatry (1999), and Ammons (2001), for example, for discussions of the uses of performance measures in government.

[3] The JTPA performance measurement system was a prototype for the accountability system presented in the Clinton/Gore National Performance Review (see Gore 1993).

[4] See, for example, Wang (2000) and De Lancer Julnes (2001).

[5] Other studies have looked at dysfunctional responses to JTPA performance measures. We postpone a discussion of this literature until later in the paper.

[6] For more complete descriptions of JTPA and its performance measurement system see, for example, Johnston (1987), Courty and Marschke (2003b), and Dickinson et al. (1988).

[7] An individual's human capital stock is her set of skills and knowledge.

[8] These earning streams would be appropriately discounted.

[9] See Heckman (2000) on the issues involved in estimating program treatment effects.

[10] During the study at the 16 experimental sites, persons who were accepted to the program were randomized into treatment and control groups. Persons in the treatment group were offered the opportunity to enroll in JTPA and receive services. Persons in the control group were embargoed from JTPA for eighteen months. Because the treatment and control groups were nearly identical, a simple comparison of earnings, employment, or welfare recipiency should reveal the impact of job training on these variables. See Doolittle and Traeger (1990) for a description of the implementation of this study and Orr, et al. (1994) for a detailed description of its results. Experimental evaluations are expensive because to obtain meaningful results large numbers of participants must be followed for many months beyond the start of training. Social experiments in which some persons are denied access to a service are also controversial. Non-experimental evaluations are cheaper but many policy analysts believe they produce unreliable estimates of impacts (see Lalonde, 1986, and Heckman, Lalonde, and Smith, 1999).

[11] For example, by taking into account local unemployment measures and other measures of the labor market, the adjustment methodology would lower the employment standard for training centers in depressed job markets compared with training centers in booming job markets.

[12] Decisions by the firm's managers that raise the stream of future profits, whether they raise profits in the near or far term, raise the value of the share and therefore the share price. In this way share prices capitalize both the short- and long-term effects of a CEO's decisions.

[13] There are reasons to suspect that measures of job-training success based on the earnings and employment outcomes of enrollees in the few weeks or months following training do not adequately capture the benefits from training. It is well known that persons who have higher levels of education experience higher wage growth rates (see, e.g., Card 2000). Higher levels of training may also lead to higher rates of growth in wages, because, for example, an individual's past training makes her future training investments more productive (for evidence of this with respect to training provided in the private sector, see Lynch 1992 and Belzil and Hansen 2003). If more training leads to higher wage growth rates then the earnings impact from successful training will increase over the lifetime of an enrollee. That is, the earnings gap between a graduate of training and a similar counterpart who receives no training should widen over time. Performance measures based only on the employment and earnings of enrollees in the few days or weeks following training, therefore, may show no benefits from training even when training is highly successful.

[14] In a survey of government agencies that developed performance measures following the passage of GPRA, many reported that they "found it particularly difficult to translate long-term strategic goals into annual performance goals. This was often because the program had a long-

term mission that made it difficult to predict the level of results that might be achieved on an annual basis." (U.S. GAO May 1997, p. 3).

[15] This paper focuses on performance measures for the adult enrollee population. The smaller youth side of the program was subject to similar but separate performance measures.

[16] In this regard, training differs from schooling, which has well-defined start and end dates. Evaluators can assess students' knowledge at the beginning and the end of the schooling period, or shortly thereafter (Coleman 1993).

[17] At the outset, the performance measures also included a cost measure, discussed below.

[18] See DOL guidance letter TEIN 3-92 available at http://wdr/doleta.gov/directives/corr_doc.asp?DOCN=282.

[19] U.S. DOL, TEIN 5-93, Ch. 2, p. 1. This document is available at http://wdr/doleta.gov/directives/corr_doc.asp?DOCN=770.

[20] Such services would have included vocational classroom training to enable enrollees to become, for example, nursing assistants, office managers, computer programmers, or security guards; on-the-job training; basic or remedial education; and job search assistance, which would have included resume writing and interviewing workshops as well as employment referrals.

[21] Postponing termination of an employed enrollee was risky because job separation rates in the participant population were high.

[22] Note that these findings lend credence to DOL's claim that under CETA training centers were avoiding terminating unsuccessful enrollees to boost performance outcomes. Other analysts have investigated whether local decision makers strategically time the reporting of outcomes to

maximize performance outcomes. See, e.g., Healy (1985) on managers in the private sector and Asch (1990) on military recruiters.

[23] *State of New Jersey Performance Standards Manual*, PY1988-89, Division of Employment and Training, New Jersey Department of Labor, April 1990.

[24] The limits on the services that could be provided after termination typically depended on how states interpreted and enforced the federal guidelines.

[25] The 16 JTPA training centers contacted for the survey were located in Corpus Christi, Texas; Cedar Rapids, Iowa; Fort Wayne, Indiana; Oakland, California; Providence, Rhode Island; Jersey City, New Jersey; Northwest, Minnesota; Butte, Montana; Decatur, Illinois; Larimer, Colorado; Heartland, Florida; Springfield, Missouri; Jackson, Mississippi, Coosa Valley, Georgia; Omaha, Nebraska; and Marion County, Ohio.

[26] We conducted this telephone survey in January and February of 1994. A copy of the survey instrument is available upon request.

[27] Politics may have also played a role in this measure's elimination (Barnow and Smith, 2002).

[28] To keep the models used to construct the standards easy to use and inexpensive to construct, many factors that would have lead to more accurate predictions of performance---such as enrollees' employment histories---were omitted in adjusting the standards. See Barnow (1992) and Trott and Baj (1987) for analyses of the job training standards in JTPA.

[29] Indeed, the performance standards' construction gave training centers an incentive to enroll employed applicants. In a sample of 6500 adult JTPA enrollees (from the experiment described in endnote 10), we found that between a quarter and a third were employed when they started training, and that a large fraction of those were employed in the same job at graduation.

[30] While WIA retains much of the decentralized nature, jurisdictional borders, administrative entities, and performance incentives of JTPA, it represents a departure in several potentially important ways. For example, WIA combines welfare and job-training services geographically, a principle it calls "one-stop shopping." Also, rather than assigning job-training enrollees to providers, WIA training centers issue vouchers to eligible participants, who then shop freely among a list of approved service providers.

[31] While it is more similar to a job-training impact, a before-after earnings measure also suffers from potential problems. For the average enrollee, earnings dip just prior to entering job training, suggesting that her earnings would eventually rise even if the job-training program had no value. This phenomenon, the so-called "Ashenfelter dip," means that before-after earnings differences are biased measures of the true impact of job training (see Heckman and Smith 1999).

[32] For a description of the WIA performance measures, see, for example the U.S. Department of Labor Training and Employment Guidance Letter, No. 9-99.

[33] To demonstrate that training center workers respond to incentives, one must show that the workers change their behavior after introduction of the incentive system. The difficulty here is defining a counterfactual benchmark pattern of behavior in the absence of incentives. The complex JTPA incentive system offers many dimensions along which one could look for performance-driven responses. Our evidence is especially compelling because it is easier to establish a connection between training center behavior and performance reporting procedures with its sharply defined deadlines then between behavior and other dimensions of the system.

[34] The literature includes many discussions of general criteria for choosing performance measures and constructing performance measurement systems (see, for example, Hatry 1980; Wholey 1999; and U.S. General Accounting Office 1996). To our knowledge, however, none

has identified the dynamic process that we note here, nor have they emphasized the importance of monitoring for and adjusting to dysfunctional responses.