DISCUSSION PAPER SERIES

IZA DP No. 3909

Curbing Cream-Skimming: Evidence on Enrolment Incentives

Pascal Courty
Do Han Kim
Gerald Marschke

December 2008

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Curbing Cream-Skimming: Evidence on Enrolment Incentives

Pascal Courty
European University Institute

Do Han Kim
University at Albany, SUNY

Gerald Marschke
Harvard University,
University at Albany, SUNY,
NBER and IZA

Discussion Paper No. 3909
December 2008

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 3909
December 2008

ABSTRACT

Curbing Cream-Skimming: Evidence on Enrolment Incentives*

Can enrolment incentives reduce the incidence of cream-skimming in the delivery of public sector services (e.g. education, health, job training)? In the context of a large government job training program, we investigate whether the use of enrolment incentives that set different 'shadow prices' for serving different demographic subgroups of clients, influence case workers' choice of intake population. Exploiting exogenous variation in these shadow prices, we show that training agencies change the composition of their enrollee populations in response to changes in the incentives, increasing the relative fraction of subgroups whose shadow prices increase. We also show that the increase is due to training agencies enrolling at the margin weaker members, in terms of performance, of that subgroup.

JEL Classification: H72, J33, L14

Keywords: performance measurement, cream-skimming, enrolment incentives, bureaucrat behavior, public organizations

Corresponding author:

Gerald Marschke
John F. Kennedy School of Government
Harvard University
79 John F. Kennedy Street
Cambridge, MA 02138
USA
E-mail: jerry_marschke@harvard.edu

* The bulk of this paper was completed while Marschke was at the University at Albany. We thank participants at the 2008 "Vouchers, Contracting-out and Performance Standards" Conference in Nuremberg.

1-Introduction

The recent introduction of performance incentives in several branches of the public service sector, such as in job training, education, and health, has raised concerns as to their impact on enrolment decisions, and on equity and efficiency outcomes. At the center of this debate is the issue that incentives induce case workers to cream-skim, that is, to select applicants on the basis of performance on measured outcomes instead of value added according to the program's stated objectives (Anderson, Burkhauser, and Raymond 1993, Cragg 1997, Heckman, Heinrich, and Smith 2002). A solution that has been proposed to retain control over the recipient population is to adjust the measures that are used to assess performance, effectively setting different 'shadow prices' for different socio-economic subgroups of enrollees. Although such methods are used in practice, and their theoretical underpinning is uncontroversial, there is no evidence that adjustment models actually have an impact on enrolment decisions.

Our case study is a large government job training program that changed the adjustment method used to assess performance three times during our sample period, using four different sets of shadow prices. We use this variation to produce the first econometric evidence on whether it is possible to influence case worker intake choices. We study the federal program created under the Job Training Partnership Act (JTPA), which, between 1982 and 2000, provided job training to the economically disadvantaged. Under JTPA, job training services were administered by over 620 semi-autonomous sub-state training agencies each evaluated according to a set of performance measures defined at the federal level. Specifically, a training agency's yearly performance was adjusted upwards or downwards to account for the particular mix of persons the agency enrolled.

To illustrate, consider the adjustment made to the employment at termination measure for enrolling adults who never received a high school degree.[2] By enrolling more high school dropouts a training agency lowered the minimum performance (the minimum fraction of participants employed at termination) necessary to avoid sanctions and qualify for a performance award. We refer to this minimum performance threshold as the performance standard. The adjustment to the standard for enrolling high school dropouts varied over time. We test whether case workers respond to the changes in these adjustments. We quantify the impact of the adjustment method both on intake populations and on performance outcomes.

There are good reasons to think that JTPA's adjustment methods may not change enrolment patterns in practice. First, case workers' preferences may vary over socio-economic subgroups, or case workers may be subject to pressures by local influence groups that override the typically weak incentives backing the adjustments (Heckman, Smith, and Taber 1996). Second, Heckman and Smith (2004) have shown that most of the selection occurs at the early stages of the participation process, such as between eligibility and awareness, over which the program staff has little or no control. Thus, even if case workers respond to changes in the shadow prices, their response may be negligible. Third, adjustment methods may have little impact in practice because they are complex. In our case study, for example, the adjustment model can potentially distinguish over 16 million different demographic subgroups for each of four different performance

[2] The employment at termination measure, the most important measure in the early years of JTPA, was defined as the fraction of program terminees who terminated with a job.

measures.[3] It may be impossible, or not worthwhile, for a training agency to attempt to factor into its enrolment strategy so many 'shadow prices'.

The paper proceeds as follows. Section 2 describes how performance adjustment was implemented in JTPA, defines the concept of performance adjustment weight (PAW), and reviews the literature on performance adjustment in non-job training areas of the public and non-profit sectors. Most importantly for our empirical study, we argue that the changes in PAW over time in JTPA are exogenous to the training agency's enrolment decision. Section 3 discusses how award maximizing case workers should respond to changes in PAW and derives predictions on changes in enrollee population and performance outcomes. We test these predictions using micro-level data on case workers' enrolment choices and performance outcomes in JTPA, leveraging three exogenous changes in the PAW. We estimate the impact of these changes in PAW following a difference-in-difference approach (time and demographic subgroups) at the agency level.

Our empirical analysis establishes two sets of results. First, we find that changes in the incentive for enrolling members of a subgroup significantly change the fraction of enrollees from the subgroup. Second, we demonstrate the existence of within-subgroup heterogeneity. Case workers increase the number of enrollees from a specific subgroup by enrolling at the margin applicants that perform worse on the measure. This finding is consistent with the cream-skimming hypothesis that case workers use their private information about the eligible population which they use to select enrollees that perform well on the performance measures. In contrast with the literature, which focuses on the

[3] 24 different adjustment factors have been used during our sample period (see Table 1, and the later discussion). Each adjustment factor takes binary values implying 2^{24} different subgroups.

impact of incentives on overall enrolment at the training agency level (differences across subgroups), we demonstrate that private information carries through even within the demographic subgroups defined by PAW.

Literature

Our results are of interest to policy makers and academics for three main reasons. First, our evidence sheds new light on the literature on cream-skimming. Interestingly, the evidence on cream-skimming is mixed (Heckman et al. 1996, Cragg 1997) and one might be tempted to conclude that enrolment decisions are not influenced by incentives. But most previous JTPA studies have focused on the enrolment incentives due to performance measurement without factoring in the role of the adjustment weights in the performance standard. They test variants of the following hypothesis: Does rewarding (or sanctioning) a training agency based on the fraction of its clients who obtain employment, dissuade it from serving high school dropouts and other persons with poor labor market prospects? But the JTPA adjustment model forces the training agency to consider how a person's attributes not only affects the performance outcome but also the standard: the agency knows that enrolling a high school dropout lowers its employment outcome but it also lowers its standard. Thus, the absence of strong evidence of cream-skimming may be because the performance standard adjustment procedure was doing its job—that is, reducing the incentive to cream-skim—and not because JTPA case workers did not respond to incentives. Our study tells a more complex picture where both effects are at play: local agencies respond to enrolment incentives but cream-skimming still takes place due to unobservable characteristics within demographic subgroups.

Second, the issue of allocation of public services receives much attention in the policy literature (Heckman and Smith 2004). This debate is fueled by a general concern over equity and also because policy makers often have specific target populations that they would like served. For example, JTPA itself introduced several constraints on the allocation of JTPA entitlements across socio-economic subgroups (Dickenson et al. 1988). The U.S. Department of Labor (DOL), which administered JTPA, defined the concept of eligible population to restrict the pool of people who could be served. In addition, budget compartmentalization capped the resources that could be used on adults and established a minimum expenditure for youth enrollees. The DOL was also desirous that the 'hard-to-serve' and 'most-in-need' not be neglected and supported the states that introduced incentives to target resources toward these sub-populations (Barnow 1992, Courty and Marschke 2003). Like quotas and budget compartmentalization, PAW are objective and transparent, but in contrast to these schemes they leave some discretion to local decision makers to exploit potential trade-offs between sub-populations. A drawback is that they may convey very complex incentives, and also, perhaps, grant too much discretion over unobserved heterogeneity within subgroups. Our work can help policy makers understand whether PAW can help achieve equity objectives and/or correct distortions due to performance incentives, or whether other methods are needed.

Third, this work contributes to the literature on the effectiveness of incentives in the public sector. Many policy analysts now believe that such systems can improve accountability and management (Osborne and Gaebler 1992, Gore 1993)[4] and such

[4] 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).

systems have become policy through the Government Performance and Results Act (GPRA) of 1993.[5] Our findings can help in the design of future adjustment systems. For example, the Workforce Investment Act (WIA), which supplanted JTPA in 2000, replaced the JTPA's regression based adjustment model with negotiated performance standards. For this and other reasons, policy-analysts have expressed concerns about cream-skimming under WIA (Barnow and Smith 2004, U.S. Government Accounting Office 2004, Heinrich 2004, Barnow and Heinrich 2008). Some analysts have called for the reintroduction of adjustment models in the upcoming reauthorization of WIA. One contribution of this paper is to provide some evidence—the first that we know of—that job training staff respond to JTPA-style adjustment models and to quantify these responses. In fact the literature has repeatedly pointed out the difficulties in separating bureaucrat and applicant motives in explaining participation (Heckman and Smith 2004). Our evidence circumvents this challenge by using a natural experiment that permits one to identify the relation between PAW and enrolment choice. We show that bureaucrats respond to sophisticated contracts that involve a large number of implicit prices and require the ability to compute complex trade-offs between alternative enrolment strategies. Although there may exist bureaucratic preferences over the choice of allocations of public resources as suggested by Heckman et al. (1996), our findings show that it is possible to influence bureaucratic preferences over intake choice.[6]

[5] 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.

[6] Heckman et al. find that JTPA case workers were more likely to enroll the applicants with the lowest prospects for employment after training. Heckman et al. call this behavior "cream avoidance" which they attribute to a "social worker mentality" in training agency staff.

2-Performance Adjustment Weights: Background and Case Study

Much of the literature on PAW has focused on their use as a means to complement and fine tune performance incentive systems.[7] The idea is that PAW can help to correct enrolment distortions due to the introduction of outcome based performance incentives. Performance incentives stimulate agency efforts to produce value added, but they may also distort the characteristics of the population the agency selects. This problem has emerged with incentive schemes in education that measure school performance using standardized test scores (Jacob and Levitt 2003), in job training that evaluate performance using labor market outcomes of trainees (Heckman et al. 2002), and in health care where doctors and hospitals are evaluated using "report cards" (Dranove, Kessler, McClellan, and Satterthwaite 2003).

The literature on PAW has been mostly conceptual or prescriptive in nature. Rubenstein, Schwartz, and Stiefel (2003) and Brooks (2002) lay out rationales for adjusting performance standards, compare and contrast different adjustment strategies that one might employ, and offer recommendations to policy-makers on how to adjust standards. Courty, Heinrich, and Marschke (2005) situate the problem in the principal-agent framework, and discuss how performance outcome measures should be adjusted. Another strand of the literature documents how PAW have been used in the context of specific applications. Trott and Baj (1987), Barnow (1992), Heinrich (2004), and Barnow and Smith (2004) discuss applications to job training programs, Siedlecki and King

[7] PAW can also be used in the absence of outcome based incentives, and the point would then be to correct possible bias due to bureaucratic preferences. In fact, policy makers may reward bureaucrats for enrolling certain groups if they feel that these groups would be otherwise underserved.

(2005) to workforce development programs, and Berne (1989), Stiefel, Rubenstein, and Schwartz (1999) and Stiefel, Schwartz, Rubenstein, and Zabel (2005) to education.

Case study: JTPA

A large literature discusses various aspects of the JTPA program (e.g. Johnston 1987), and offers descriptions of its incentive system (e.g. Courty and Marschke 2003) and the bureaucratic responses they induce (e.g. Heckman and Heinrich, forthcoming). To reduce unnecessary repetition, we present here only those features of the organization that are essential to our analysis of PAW, and direct the reader to more comprehensive sources when required.

The JTPA program was highly decentralized: the 620 plus training agencies administered the program with significant discretion over whom to enroll. While applicants had to meet an income test to be eligible, JTPA was not an entitlement. Given the JTPA annual budget (approximately $4.1 billion in 1993), and the large population that was eligible for training, agencies could serve only one to three percent of the eligibles (550,000 new participants were enrolled in 1993).[8] The decision of which eligibles to enroll constitutes the focus of this paper.

The Act called for financially-backed performance incentives that would measure and reward training agency's success in developing participants' human capital, the primary goal of the program according to the Act (JTPA, section 106(a)). Congress gave the DOL the responsibility of developing a workable set of performance measures that would reflect the Act's mission. The JTPA fiscal year, or *program year*, ran from July 1 to

[8] See Dickenson et al. (1988) for a complete description of the JTPA eligibility rules.

June 30 of the next calendar year. At the end of each program year, training agencies were rewarded (or sanctioned) on the basis of their performance relative to these DOL measures. For the average training agency, the award amounted to about seven percent of the operating budget.

Our empirical analysis focuses on program years 1993-1998 and on the adult JTPA population. For the 1993-1998 period, the DOL used four performance measures constructed from two labor market outcomes, employment and earnings, to evaluate training agencies. A training agency's employment rate at follow-up (ER) for a particular program year was calculated as the fraction of enrollees terminated during that year who were employed 13 weeks after termination. The average weekly earnings (WE) was calculated as the average weekly earnings during the ninety days following termination for those enrollees who were employed 13 weeks after termination. From the ER outcome, two performance measures were constructed: one ER measure was based on the performance of all adult enrollees and another was based on the performance of only the welfare-receiving subset of adult enrollees. Similarly, separate adult and adult welfare performance measures were constructed based on WE. Each measure had associated to it a separate standard. The DOL set lower standards for the welfare versions of the measures. Meeting these standards was a condition for receiving an award and in many states most of the award a training agency was eligible for was paid out for simply meeting the standard. Thus, the structure of the incentives under JTPA meant that a training agency interested in avoiding sanctions and maximizing its award, should focus on meeting its standards.

For each of the four standards, the DOL developed an adjustment model to establish a training agency-specific standard that accounted for the particular agency's enrollee choices (demographic characteristics of the enrollee pool) and local labor market circumstances (socio-economic conditions outside the control of the agency). For example, it was determined that training agencies that enrolled few high school dropouts should be handicapped relative to those that enrolled more, and that training agencies should not be penalized for operating in particularly adverse labor markets. In this study, we focus exclusively on the set of factors in the DOL adjustment models that are based on the demographic characteristics of the enrollee population, as only these factors can influence enrolment decisions.

PAW in JTPA

To illustrate how the adjustment methodology works, assume two demographic factors, gender (female, male) and race (black, non-black). The training agency is rewarded on the basis of excess performance, that is, performance above the performance standard. The DOL model adjusts the performance standard around an exogenously given baseline level, that we denote m_0, depending on the characteristics of the enrollee population. Suppose an agency enrolled x_f percent of females, x_b percent of black and denote by β_j the adjustment weight for demographic characteristic j=f,b. A stylized performance adjustment model can be written as

$$M_0(x_f,x_b)= m_0-(\beta_f x_f+\beta_b x_b) \qquad (1)$$

where M_0 is the adjusted performance standard. The higher the standard, the greater is the difficulty obtaining an award. We define an adjustment factor as a socio-economic variable

(e.g. x_f) that is used to correct the standard and an adjustment weight as the numerical value that is imputed to correct the standard (e.g. β_f). For example, if β_f is positive, then the agency is more likely to receive an award, ceteris paribus, if it enrolls more females.

The DOL chose different sets of factors for each performance measure based upon their availability, their statistical relation with the performance measure, and political considerations. The first line in Table 1 presents the baseline level (m_0), the first column in the bottom panel presents the set of adjustment factors (x) for the adult ER and WE standards, and the core of the table reports the value of the adjustment weights (β) corresponding to these factors.[9] The columns report the weights for different program year cycles. The adjustment weights remain in force for two consecutive program years before they are updated. Thus for example in program years 1992 and 1993, the adjustment weights for the ER standard for females was .072; in program years 1994 and 1995 it was .056; and so on. They are constructed before the beginning of a new two year cycle, using information on demographic characteristics and outcomes observed in the previous cycle, as the coefficient estimates from a regression of performance outcomes on demographic and labor market characteristics.[10]

Table 1 shows that the enrolment incentives embedded in the DOL adjustment model can significantly impact the performance standard and therefore the agency's likelihood to receive an award. For example, an agency in either 1992 or 1993 enrolling only applicants that embodied all of the characteristics associated with positive weights

[9] We obtained the adjustment weights from Guide to JTPA Performance Standards for Program Years 1992-1993, 1994-1995, 1996-1997, and 1998-1999 published by Social Policy Research Associates (see footnote in table 1 for full cites).

[10] For more explanations on the process of estimating the coefficients in the regression model, see U.S. Department of Labor (1987), Barnow (1992) or Social Policy Research Associates (1999).

would face a 'negative' performance standard on the employment measure (the adjustment, $100\Sigma_i\beta_i$, is greater than the baseline level implying $M_0<0$), meaning that it would not be penalized even if none of its terminees were employed at termination. Although this example is extreme, Table 1 reveals that many of the weights can lower the employment standard by 10 percent or more.

Table 2 focuses on the employment measure and presents summary statistics on the distribution across agencies of the actual adjustment to the baseline ($\Sigma_i\beta_i x_i$) by program year. Line 1, for example, says that the ER standards in 1993 varied across training agencies from 37 percent (86-49) to 74 percent (86-12) to suggesting that a training agency's enrolment pool—which is a choice variable—could greatly influence its standard. The adjusted performance standards for the earnings measure (not reported) show the same degree of variation.

The likely impact of PAW on enrolment is difficult to assess on theoretical grounds alone. On the one hand, the magnitude of the changes in the weights implies that the enrollee intake composition may have a significant impact on the standard. Meeting the standard was an issue in practice and could have financial consequences.[11] In fact, over the period 1993-1998, on average about 23% of training agencies failed to meet the employment rate standard, 6% failed to meet the earnings standard, and 6% failed to meet to meet both standards.[12] On the other hand, the number of demographic subgroups, and

[11]The award for the successful training agency averaged about seven percent of its budget. In some states, the highest awards amounted to about sixty percent of the training agency's budget. The reader who is interested in the details of the incentives confronting JTPA training agencies should see Courty and Marschke (2003).

[12] If the performance standards were set too high, so that all training agencies would fail no matter how they tried, then the ability to modify a standard using the enrolment composition would not matter much, and one would not expect to see enrolment choices

thus the number of implied shadow prices, increases exponentially with the number of factors. For example, there were 10 factors active in 1993 for the employment at termination measure (Table 1) which required the agency to distinguish among 1024 subgroups. In addition, the PAW varied across performance measures. As a result, PAW introduced very complex trade-offs and may have had little consequence in practice. In the end, whether PAW influenced intake choices is an empirical issue.

Table 1 shows that there are significant changes in the adjustment weights over time. For example, to meet its employment standard in program years 1992 or 1993, an agency that enrolled no 'high-school dropouts' would have to achieve an employment rate 18.4 percent higher than an agency that enrolled only 'high-school dropouts' (assuming that all other characteristics are equal across the agencies). In program year 1998 or 1999, however, the difference drops to 6.6 percent, an order of magnitude of about three. In addition, some adjustment factors eventually disappear from the adjustment worksheets and new factors are introduced.[13]

To write a micro model of enrolment and for the empirical work as well, it is more convenient to work with demographic subgroups instead of demographic characteristics. There is a simple correspondence between subgroups and characteristics. In our example, the two factors determine four demographic subgroups (black female, black male, and so on). Denote $s=(s_{bf},s_{bm},s_{nf},s_{nm})$ the enrolment vector measured in percentage of overall population over demographic subgroups where s_{bf}, for example, represents the percentage of enrollees

---

affected by PAW. This is also true if performance standards were set too low so that all training agencies exceeded their standards whether they enrolled purposefully or not.

[13] This lifecycle phenomenon of adjustment weights was observed earlier in JTPA's history by Barnow and Constantine (1988) who attribute it to increased proficiency due to learning by the training agencies in selecting enrollees on the basis of factors not included in the model.

who are black and female. We can rewrite the performance standard as

$$M_0(n)= m_0-(\omega_{bf}s_{bf}+\omega_{bm}s_{bm}+\omega_{nf}s_{nf}+\omega_{nm}s_{nm}) \qquad (2)$$

where $\omega_{bf}=\beta_b+\beta_f$, captures the decrease in standard due to increasing the fraction of black female by one percent, and can be interpreted as the 'shadow price' for that demographic subgroup. The other coefficients are similarly derived, $\omega_{bm}=\beta_b$, $\omega_{nf}=\beta_f$, $\omega_{nm}=0$. In the rest of this paper, we also call the ω adjustment weights, keeping in mind the distinction between the β in (1) and ω in (2).

3-Theoretical Predictions

We present in the appendix a microeconomic model of the training agency choice of enrollee population. We derive predictions on how the agency should respond to changes in the adjustment weights: how enrolment decisions and performance outcomes should change for different demographic subgroups. This section discusses the intuition behind the model and summarizes its predictions.

To simplify, we assume there is a single performance measure and I distinct demographic subgroups. The cost of training is assumed subgroup-specific and increasing and convex in the number of enrollees. Similarly, average performance decreases with the number of enrollees from a specific subgroup. These assumptions are consistent with the following interpretation. Applicants differ within a subgroup. Some applicants are easier to train and are more likely to achieve successful outcomes than others. It is optimal for the training agency to select first the most promising applicants. If the training agency hires more applicants of a given subgroup it will hire those who cost more to serve (cost is increasing) and who are less likely to perform well (performance increases at a decreasing rate). These

assumptions are reasonable if there is some heterogeneity within demographic subgroups that is observed by the agency. Cream-skimming becomes possible because the agency enrolls those applicants, within a demographic subgroup, who are likely to perform well on the measure, irrespectively of how well they perform on the true objective of job training.

The training agency allocates its budget across the demographic subgroups to maximize its award and may also have its own preferences over enrolment choices. We first show that under general assumptions about the cost and performance outcome functions, the training agency responds to an increase in the adjustment weight of demographic subgroup i by enrolling more applicants of subgroup i and fewer applicants of subgroup $k \neq i$. The proposition holds independently of the training agency's own preferences over enrollee choices. We then consider the impact of a change in the adjustment weight on the average subgroup performance outcome. As the adjustment weight of subgroup i increases, the number of enrollees of subgroup i increases and the average performance outcome of subgroup i decreases. The reason is simply that to increase its enrolment of applicants of type i, the training agency has to enroll less attractive applicants. Marginal enrollees achieve lower performance outcomes than average ones.[14]

Under additional assumptions on the model's primitives, we can derive general predictions on the impact of any changes in the performance weights. Specifically, the increase in the number of enrollees from subgroup i (Δn_i) is greater than the same change for subgroup k (Δn_k) if subgroup i's adjustment weight increases by a larger amount (or decreases by a lower amount) than subgroup k. Formally, $\Delta n_i > \Delta n_k$ if and only if $\delta_i > \delta_k$ for any i and k, where δ_i denotes the change in weight i. The result also applies to changes in the

[14] To simplify, we assume that the training treatment is constant across groups.

fraction of enrollees and this constitutes the focus of our empirical investigation (Hypothesis H1). The result on average performance outcomes also generalizes to any change in weight: the change in average performance of subgroup i is lower than the change in average performance of subgroup k if $\delta_i > \delta_k$ (Hypothesis H2). The remainder of this paper tests hypothesis H1 and H2.

4-Data and Empirical Strategy

The variables we wish to explain with our analysis, measured at the level of demographic subgroup-agency-year, are the enrolment shares and the performance outcomes. To compute these variables, we use data from the Standardized Program Information Report which are collected by the U.S. Department of Labor and distributed by the W.E. Upjohn Institute of Employment Research. Appendix 2 explains in detail how we constructed our panel data of demographic subgroups, agencies, and years. Consistent with the JTPA incentives, we form the subgroups based on the terminees in the program year, not the enrollees. We do this because the adjustment model modifies the standards based on the characteristics of the program year's terminees.[15]

Recall that using all 24 factors would generate more than 16 millions subgroups. Since the JTPA enrollee population is much smaller (for the enrolment analysis, for example, we have information on 682,515 terminees over the 6 program years), we eliminate all the

[15] Of course, the terminee population closely resembles the enrollee population. We are explicitly assuming that the training agencies anticipate the effects of its enrolment decisions on the performance standards which is the case when the standard remains unchanged. This is reasonable because the average length of training (a few months) is short relative to the period during which the weights remain constant (two years). We have considered the possibility of delays in the enrolment responses after the three changes in the PAW that took place in our sample period and this did not change the results.

subgroups for which we have no or few enrollees. In the end, we select 13 factors and construct 1,670 different subgroups for which we have information over all 6 years in at least one agency. This yields an average of 291 subgroups per agency-year. Table 3 presents descriptive statistics for our main variables (PAW, enrolment shares, and performance outcomes).

Table 1 and 2 demonstrate that there is much variation across years in our explanatory variables. Table 4 shows that there is also much variation from year to year in the enrolment size of the demographic subgroups identified by the DOL adjustment factors. We investigate whether this variation in enrolment can be explained by the changes in adjustment weights as predicted by (H1).

Multi-dimensionality

Each of the four standards had its own adjustment weights that could potentially influence the enrollee intake choice (H1) and also the performance outcomes (H2). Under JTPA, states were responsible for designing the incentive contracts using the four measures proposed by the DOL. Although these contracts vary greatly from state to state (different emphasis on the different measures and different choice of the award function), we can leverage three patterns that are common to all contracts to cope with the multi-dimensional nature of the incentive system. To start, the employment measure received a disproportionate emphasis in determining the award (Courty and Marschke, 2003). Moreover, awards were largely allocated for meeting standards and training agencies were more likely to fail the employment standards. Third, the PAW for the two welfare measure standards apply only to welfare subgroups and therefore should not influence the choice of non-welfare enrollees.

Given these considerations, we proceed as follows. We initially test H1 and H2 using the adjustment weights on the adult ER performance measure. In focusing on the employment measure, this approach follows the policy evaluation literature (e.g., Anderson et al.1993) and is justified by the first two characteristics of the incentive contracts mentioned above. Later, we introduce the adjustment weights on the WE standard. Since we do not have information on the contracts, we employ a general specification for how ER and WE could influence enrolment and outcomes that allows for interaction effects. This first set of analyses is valid under the assumption that the change in weights that apply to the welfare measures are independent from the change in weights that apply to the two measures we consider. As a robustness check, we reproduce the previous analyses without the welfare subgroups. The third characteristic of the incentive system implies that H1 and H2 hold for this subset of the sample even if the above assumption does not hold.

Exogeneity of the changes in the PAW

The PAW were changed three times in our sample period (see Table 1). In the empirical analysis, we assume that these changes are exogenous to contemporaneous enrolment decisions. Several arguments support this assumption. Recall that the PAW were computed as coefficient estimates of a regression of performance outcomes on demographic factors using performance data (from all training agencies) in the previous two year cycle. The DOL regression model used to compute the PAW was unstable and this was due to multi-colinearity between factors. Consistent with this view, Table 1 shows that the choice of demographic factors varied greatly over time (only 9 out of 24 demographic factors were used throughout our period). This choice was partially driven

by the concern to keep the PAW positive (since all the selected factors represented priority target subpopulations) and by current political considerations. The change in the PAW are exogenous if they are mainly driven by the arbitrary choice of the factors included and the sample realization of the two year cohort used to compute the regression coefficients.

But changes in the PAW could also be driven by changes in labor market conditions and/or changes in enrolment strategies. We argue that this is not an important issue for our empirical exercise, and if anything, it can only create an under-estimation of the agency responses. Consider first the later point. The concern is endogeneity of the PAW through strategic inter-temporal enrolment behavior. The behavior of all agencies as a group influences changes in the weights, because the DOL used the information collected on past enrolment choices and outcomes to update the weights, but an individual agency can be assumed to maximize the current period award myopically since the impact of its enrolment decisions on future weights is negligible. Consider next changes in labor market conditions. To start, assume that these changes are conditionally uncorrelated (e.g. random walk or permanent changes). Such changes would influence the PAW (through the regression model) but this would not introduce an endogeneity problem since the change in next period labor market conditions is uncorrelated with the current change in PAW. The only concern are trends in labor market conditions. Such trends could bias the inference against our hypothesis. Assume for example that the labor market potential of a subgroup starts to degrade. This increases that subgroup's PAW but the increase under-compensates for the continuing degradation in the subgroup's potential so we would under-estimate the enrolment response relative to the response that would take place with

a truly exogenous change in PAW.

Empirical strategy

Denote s_{iat} as the share of enrollees of demographic subgroup i in agency a in year t and w_{it} the adjustment weight, common to all agencies, for subgroup i and year t.[16] H1 implies an increasing relation between changes in relative weights and changes in relative shares of subgroups. We test this relation using the three changes in adjustment weights that took place in our sample period (the weights changed at the end of 1993, 1995, and 1997).

We propose different specifications to test H1 that are variations around the following approach. Assuming that the increasing relation implied by H1 is linear and does not vary across subgroups, agencies, or years gives

$$(s_{iat}-s_{iat'})-(s_{kat}-s_{kat'})=\gamma[(w_{it}-w_{it'})-(w_{kt}-w_{kt'})] \quad \text{for all i,k,a,t,t'} \qquad \text{(H1')}$$

where the parameter of interest γ is positive. Instead of comparing pairs of demographic subgroup-years, which does not naturally fit a regression framework, we aggregate this hypothesis to obtain a relation that can be estimated using a fixed-effect regression framework.[17] Formally, H1' is averaged over subgroups k and years t' to obtain

$$s_{iat}= -(s_a- \gamma w_a)+(s_{ai}-\gamma w_{ai})+(s_{at}-\gamma w_{at})+\gamma w_{it} \quad \text{for all i,a,t}$$

where s_a denotes the average share in agency a across all years and subgroups, s_{ai} the average i share in agency a across all years and similarly for s_{at} and the w averages. We can rewrite

---

[16] s_{iat} is defined as $\frac{n_{iat}}{\sum_k n_{kat}}$ where n_{iat} is the number of enrollees of group i in agency a and year t.

[17] Alternatively we could choose a subgroup to serve as the reference subgroup against which we compare all other subgroups but the choice of the reference subgroup is arbitrary.

this relation as a difference in difference (time and subgroup) equation at the agency level

$$s_{iat} = \alpha_a + \alpha_{ai} + \alpha_{at} + \gamma w_{it} \quad \text{for all i,a,t.}$$

The observed shares could vary randomly because they are measured with error (which is the case in our application since only a representative sample of 62 percent of total population is included in our dataset).[18] We obtain the following empirical model

$$s_{iat} = \alpha + \alpha_{ai} + \alpha_{at} + \gamma w_{it} + \varepsilon_{ait} \quad (3)$$

where α is a constant, α_{ai} is a subgroup-agency fixed effect and α_{at} is an agency-time fixed effect. We assume ε_{iat} is normal, mean zero, and distributed independently across training agencies.

The theory makes no prediction on $(\alpha,\alpha_{ai},\alpha_{at})$ but predicts that γ should be positive. Specification (3) tests an averaged version of H1. We interpret γ as the average effect over all training agencies, all year changes, and subgroups. We cluster the errors at the training agency level to permit arbitrary forms of autocorrelation and heteroscedasticity within training agency panels.

To test H2, we follow a similar procedure. We estimate the performance of each subgroup holding constant agency-subgroup and agency-time fixed effects

$$m_{iat} = \theta + \theta_{ai} + \theta_{at} + \theta w_{it} + \upsilon_{ijt} \quad \text{for all i,a,t} \qquad (4)$$

where as before θ_{ai} and θ_{at} allows for agency-subgroup and agency-time fixed effects. The parameter of interest is θ, which our model predicts is negative. As with (3), we assume υ_{ijt} is normal, mean zero, and distributed independently across training agencies and we cluster the errors at the training agency level.

[18] Alternatively, we could derive the econometric model following a random utility approach, assuming that agencies have group preferences that vary randomly over time.

In all specifications reported to test H1 and H2, we weight each subgroup-agency-year observation by the subgroup-agency share of the entire terminee population.[19] We have also considered two variant specifications and the results were not affected (not reported): one with equal weights and another with weights proportional to the subgroup's share relative to its agency population. In addition, we have considered specifications where, constructing the subgroups, we exclude those enrollees who are terminated in the first four months of each two year cycle. Our reasoning is that enrollees entering a new two year cycle may have been enrolled to optimize the previous cycle's weights (see footnote 16).[20]

5-Results

5-1 Tests of H1 and H2 for the ER Adjustment Weights

Table 5 reports the results from our estimation of the enrolment decision model, equation (3). In all specifications the dependent variable is the subgroup's termination share. The right-hand side of the regression includes the subgroup's weight for the employment standard (ER) in addition to the α_{ai} and α_{at}.

Model 1 produces a positive and statistically significant estimate of the ER weight coefficient, a finding that is consistent with H1. To give the reader an idea of the magnitude of the impact of the weight change on enrollee choice, we include the standardized coefficients. Literally interpreted, our result says that a one standard deviation in a

---

[19] That is, we weight each observation by $\frac{\sum_t n_{iat}}{\sum_a \sum_k \sum_t n_{kat}}$ where n_{iat} is the number of enrollees in subgroup i in agency a in year t.

[20] We chose four months, because the average enrolment duration is between four and five months long. Five months into the new cycle, we reason, enrollees will be terminating in the cycle in which they were intended to be terminated.

subgroup's performance weight relative to the average ER weight increases the subgroup's enrolment share by about .1 percent relative to the average agency subgroup. This response, however, is measured at the subgroup level which is the correct unit of analysis to understand agency behavior, but is of limited relevance from an economic or policy point of view. To assess the economic significance of this response, consider the following thought experiment. Assume a coefficient on a demographic characteristic, e.g. female, is increased by a standard deviation relative to the average coefficient. The enrolment share of all female subgroups will increase by about .1 percent relative to the average subgroup. Since there are on average 291 subgroups per agency in our sample (see Appendix 2), the overall increase in the share of females will be 14.7 percent (0.00101*291/2, because half of the subgroups are female on average). The female PAW alone can have a large impact on the composition of the enrollee population. But there are approximately a dozen demographic characteristics in play in the analysis suggesting that changes in PAW do have a significant influence on enrolment.

Model 3 includes both the subgroup's ER and WE adjustment weights. The coefficient estimate on the employment weight (in raw and standardized form) changes little from column 1. This finding supports our assumption that there is little interaction between the different measures of the incentive system.

Table 6, column 1, reports the results of the employment outcome estimation (equation 4). We find a statistically significant and negative coefficient estimate on the employment weight, as predicted under H2. The magnitude of the estimate suggests that a one standard deviation increase in a subgroup's ER weight relative to the average ER weight decreases the subgroup's relative employment rate by about 2 points. This result remains when we add the earnings weight as an explanatory variable. To assess the economic implication of this result,

consider the thought experiment discussed above. Increasing the female weight by one standard deviation decreases the performance of females relative to the average subgroup performance by 2 points. This figure seems reasonable to us considering that this change in weight is associated with a 14.7 percent increase in the relative share of females. The quality of the marginal enrollee within a subgroup decreases as more enrollees are drawn from this subgroup, which is consistent with the hypothesis that agencies cream-skim the best enrollees within each subgroup. Still, this figure is small relative to the variation in performance across demographic subgroups. Table 3 shows that the standard deviation in the employment outcome across all subgroups and years is 43 points.[21] Therefore the potential to cream-skim across subgroups (which can be curbed with the PAW) is of an order of magnitude greater than within subgroups (which is unaffected by the PAW). The PAW can eliminate the incentive to cream-skim across-subgroups leaving a residual incentive to cream-skim within subgroup which is second order.

5-2 Additional Measures and Robustness

The previous analysis is valid under the assumption that the ER has received the most emphasis in the incentive scheme, as has been argued in the literature, and consistent with the observation that the failure rate is much larger for the ER measure, or under the alternative assumption that the variation in the different enrolment incentives associated with

[21] What is relevant for cream-skimming across subgroups in the absence of PAW are the predictable differences across subgroups at the agency level. To capture this, we first take a year average of the subgroup performance at the agency level (this eliminates the unpredictable component of performance that is irrelevant in cream-skimming), then compute the standard deviation in subgroup performance at the agency level, and finally take the average across all agencies. We obtain an average standard deviation of 43.6 which is very similar to the above figure.

each set of weights are orthogonal to one another. Still, multi-dimensional incentives may matter. We address this issue in two ways. First, we test H1 and H2 for the adjustment weights on the WE measure but the results are inconclusive. Second, we consider the impact of the weights on both the ER and WE measures on the sub sample of non-welfare recipients and the logic is that the weights on the other two measures (the welfare ones) should not influence the enrolment incentives among the non-welfare sub populations. The results on the WE weights obtained from this sub-sample are consistent with H1 and H2 and we propose a possible interpretation that reconciles these two new sets of results.

Adjustment Weights on the Earnings Performance Measure (WE)

Table 5, columns 2 and 3, show the impact of the WE adjustment weights on the enrolment decision. Both columns show no impact which goes against H1. Table 7 reports the results of the earnings outcome specification (model (4) applied to WE). Whether we estimate the model with just the earnings weight or both the earnings and employment weights, the estimated coefficient on the earnings weight is statistically insignificant against H2.[22] These two results could be because the WE measure plays a lesser role in the incentive system or because the agencies have less discretion to select enrollees who are likely to perform well on the WE measure.

Interestingly, the coefficient estimate corresponding to the employment weight in table 7, column 2, is positive and significant (p value 0.001). While there are many potential explanations for this finding, it is consistent with the existence of a trade-off under multi-

[22] The number of observations used in this analysis is smaller than in the employment analysis because, consistent with the JTPA definition of the earnings measure, we use only the enrollees who are employed (by the employment measure definition) in the calculation of the earnings outcome.

dimensional incentives: the kinds of enrollees within a subgroup that produce higher employment outcomes, reduce earning outcomes.

Non-Welfare Recipients

The training agency's decision to enroll adult non-welfare recipients is less complicated than the decision to enroll welfare ones because non-welfare recipients' characteristics enter into the determination of only the two standards that have been the focus of this analysis. If the welfare measures play an important role, we should obtain a cleaner test of H1 and H2 when we limit the analysis to the non-welfare adults. Therefore, Table 8 shows the results of the previous analyses excluding the welfare recipients.

Two points should be made. First, the power of the significance tests is smaller after we exclude welfare recipients, which constitute about 40 percent of the adult population. This is partly responsible for why we observe that the coefficient estimates on the employment weight in the enrolment share (model 1) and outcome regressions (models 2 and 3) are insignificant. Second, the impact of the earnings weight in the regressions is greater when we exclude welfare recipients. In the enrolment share regression, though the coefficient estimate on earnings weight remains insignificant (by conventional significance standards) it is positive (as predicted under H1). The coefficient estimate on earnings weight in the earnings outcome regression (model 3) is now both negative and significant as predicted under H2. The standardized coefficient corresponding to this estimate is about -6 suggesting that a one standard deviation increase in the WE weight relative to the average WE weight reduces the relative subgroup earnings per week by about $6.

Taken together, these two new sets of results suggest that although H2 does not hold for

the entire sample, it does hold for the subset of non-welfare recipients. This may be because agencies have much more discretion to select applicants who are likely to perform well on the earnings measure, when they have to choose among non-welfare recipients, than they do for welfare recipients, who have on average lower levels of human capital. Also the earnings measure is calculated only off employed terminees. Because welfare recipients are less likely to be employed their prospective earnings might not be of such concern in the enrolment decision.

5-Summary and Conclusions

The recent introduction of performance incentives in several branches of the public service sector, such as in job training, education, and health, has raised concerns as to their impact on enrolment decisions. In particular, rewarding public agencies based on measurable outcomes such as employment outcomes, test scores, or health outcomes may lead to student-tracking in education or the neglect of the hard-to-serve in job training and of the chronically ill in health care. To retain control over the recipient population, some policy-makers have proposed adjusting the measures that are used to assess performance, effectively setting different ‘shadow prices’ for different subgroups of clients, but little evidence exists about the effectiveness of these methods in practice.

In the context of a large government job training program, we investigate the influence of enrolment incentives on case workers’ choice of intake population. Job training agencies in this program are rewarded for improving the labor market performance of the clients they serve but the reward function also depends on the enrolment choice. The main objective of the enrolment incentives is to level the playing

field, so that a training agency enrolling less able applicants has to meet a lower level of performance, effectively setting a system of shadow prices that correct for the challenge that each demographic subgroup presents.

Our empirical analysis establishes two sets of results. First, we measure the impact of changes in the relative shadow prices on changes in the relative fraction of different demographic subgroups. We find that changes in the incentive for enrolling members of a subgroup significantly change the fraction of enrollees from this subgroup. This is good news for those who wish to use PAW in job training and in other public services to attenuate the negative distributional consequences of performance-based incentive systems. One should keep in mind that the effectiveness of PAW elsewhere, however, will depend on the nature of the heterogeneity among participants and in the ability of the designer to identify dimensions over which cream-skimming takes place.

Second, we demonstrate the existence of within-subgroup heterogeneity. Case workers increase the number of enrollees from a specific subgroup by enrolling at the margin applicants that perform worse on the measure. That is, case-workers appear to be cream-skimming: they use their private information about applicant heterogeneity within subgroups. In contrast with the literature, which focuses on the impact of incentives on enrolment at the training agency level, we demonstrate that private information carries through even within the demographic subgroups defined by PAW. We show, however, that the potential for cream-skimming within subgroups is second order relative to across subgroup cream-skimming.

In this paper, we took the DOL methodology as given and investigated whether training agencies respond to exogenous changes in the enrolment incentives. An

important issue that we did not address is to evaluate the choice of the DOL's methods for determining the PAW. Did the DOL methodology achieve a reduction in cream-skimming or some other objective (e.g. channeling resources toward the subgroups with the highest earning impact)? We leave for future research the issue of determining how to set the PAW to achieve a given objective and to evaluate the impact of the PAW set by the DOL on cream-skimming.

References

Anderson, Kathryn, Richard Burkhauser, and Jennie Raymond. 1993. "The Effect of Creaming on Placement Rates under the Job Training Partnership Act." *Industrial and Labor Relation Review* 46(4): 613-24.

Barnow, Burt. 1992. "The Effects of Performance Standards on State and Local Programs: Lessons for the Job Opportunities and Basic Skills Programs." In *Evaluating Welfare and Training Programs, ed.* Charles Manski and Irwin Garfinkel, 277-309. Cambridge, Mass.: Harvard University Press.

Barnow, Burt and Jill Constantine. 1988. *Using Performance Management to Encouraging Services to Hard-to-Serve Individuals in JTPA.* National Commission for Employment Policy Research Report.

Barnow, Burt and Carolyn J. Heinrich. 2008. "One Standard Fits All? The Pros and Cons of Performance Standard Adjustments." Working paper, Johns Hopkins University and the University of Wisconsin.

Barnow, Burt and Jeffery Smith. 2004. "Performance Management of U.S. Job Training Programs: Lessons from the Job Training Partnership Act." *Public Finance and Management* 4(3): 247-87.

Berne, Robert. 1989. *Relative School-Level Performance Measures for Elementary and Middle Schools in New York City*. New York: Urban Research Center, New York University.

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

Brooks, Arthur C. 2002. "The Use and Misuse of Adjusted Performance Measures." *Journal of Policy Analysis and Management* 19(2): 323-34.

Courty, Pascal, Carolyn Heinrich, and Gerald Marschke. 2005. "Setting the Standard in Performance Measurement Systems." *International Public Management Journal* 8(3): 1-27.

Courty, Pascal and Gerald Marschke. 2003. "Performance Funding in Federal Agencies: A Case Study of a Federal Job Training Program." *Public Budgeting and Finance* 23(3): 22-48.

Cragg, Michael. 1997. “Performance Incentives in the Public Sector: Evidence from the Job Training Partnership Act.” *Journal of Law, Economics, and Organization* 13(1): 147-68.

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

Dranove, David, Daniel Kessler, Mark McClellan and Mark Satterthwaite. 2003. "Is more Information Better? The Effects of Report Cards on Cardiovascular Providers and Consumers." *Journal of Political Economy* 11: 555-88.

Gore, Al. 1993. *Creating a Government That Works Better & Costs Less. Report of the National Performance Review.* Washington, D.C.: U.S. Government Printing Office.

Heckman, James and Carolyn Heinrich. (eds.) Forthcoming. *Performance Standards in a Government Bureaucracy: Analytic Essays on the Performance Standards*

Systems in Job Training Programs. Kalamazoo, MI: W. E. Upjohn Institute for Employment Research.

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

Heckman, James and Jeffrey Smith. 2004. "The Determinants of Participation in a Social Program: Evidence from JTPA." *Journal of Labor Economics* 22(2): 243-98.

Heckman, James, Jeffrey Smith, and Christopher Taber. 1996. "What Do Bureaucrats Do? The Effects of Performance Standards and Bureaucratic Preferences on Acceptance in the JTPA Program." In *Advances in the Study of Entrepreneurship, Innovation, and Growth, Volume 7, Reinventing Government and the Problem of Bureaucracy*, ed. Gary Libecap, 191-218. Greenwich, Conn.: JAI Press.

Heinrich, Carolyn. 2004. "Improving Public-Sector Performance Management: One Step Forward, Two Steps Back?" *Public Finance and Management* 4(3): 317-51.

Jacob, Brian and Steven Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics.* 118(3): 843-78.

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

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): 348-58.

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

Osborne, David, and Ted Gaebler. 1992. *Reinventing Government: How the Entrepreneurial Spirit Is Transforming the Public Sector*. Reading, Mass.: Addison-Wesley.

Rubenstein, Ross, Amy Ellen Schwartz, and Leanna Stiefel. 2003. "Better than Raw: A Guide to Measuring Organizational Performance with Adjusted Performance Measures." *Public Administration Review* 63(5): 607-15.

Siedlecki, Joe and Christopher J. King. 2005. *Approaches to Adjusting Workforce Development Performance Measures.* Ray Marshall Center for the Study of Human Resources, Lyndon B. Johnson School of Public Affairs, University of Texas at Austin, Occasional Brief Series, 1(2).

Social Policy Research Associates. 1993. *Guide for Setting JTPA Performance Standards for Program Year 1992-1993*. Menlo Park, Calif.: Social Policy Research Associates.

Social Policy Research Associates. 1995. *Guide to JTPA Performance Standards for Program Years 1994 and 1995*. Menlo Park, Calif.: Social Policy Research Associates.

Social Policy Research Associates. 1997. *Guide to JTPA Performance Standards for Program Years 1996 and 1997*. Menlo Park, Calif.: Social Policy Research Associates.

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

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

Stiefel, Leanna, Amy E. Schwartz, Ross Rubenstein, and Jefrrey Zabel. 2005. *Measuring School Performance and Efficiency: Implications for Practice and Research.* NY: Eye on Education, Inc.

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. 2004. Workforce Investment Act: States and Local Areas Have Developed Strategies to Assess Performance, but Labor Could Do More to Help. GAO-04-657. Washington, D.C.: U.S. General Accounting Office.

U.S. Department of Labor. 1987. *Guide to Setting JTPA Title II-A Performance Standards for PY 87*. Washington, D.C.: Employment and Training Administration.

Wholey, Joseph and Harry Hatry. 1992. “The Case for Performance Monitoring.” *Pubic Administration Review* 52(6): 604-10.

Table 1: Baseline Levels and Adjustment Factors and Weights, Program Years 1992-1999

	Follow-up Entered Employment Rate (ER, %)				Follow-up Weekly Earnings (WE, $/week)			
	1992-93	1994-95	1996-97	1998-99	1992-93	1994-95	1996-97	1998-99
Baseline level (m_0)	86	88	88	77	266	301	361	376
Adjustment Factors (x)	Adjustment Weights (β)							
Female*	0.072	0.056	0.052	0.050	0.425	0.443	0.602	0.683
55 years old & over	--	0.118	0.105	0.130	1.126	0.774	0.484	0.61
age 30 to 54*	--	0.098	--	--	--	--	--	--
Black*	0.064	0.086	0.035	0.027	0.270	0.325	0.226	0.177
Other minority*	--	--	--	--	--	0.100	0.042	0.065
Minority male*	--	--	--	0.026	--	--	0.279	0.306
High school dropout*	0.184	0.084	0.073	0.066	0.271	0.276	0.24	0.145
Post high school attendees*	--	-0.066	-0.032	-0.008	-0.415	-0.659	-0.235	-0.334
High school dropout under 30	--	--	0.02	0.015	--	--	--	0.088
Handicapped	0.083	0.09	0.075	0.096	0.367	0.558	0.28	0.315
UI or UC claimant	--	--	-0.037	-0.022	-1.062	-0.361	-0.127	-0.081
Long-term AFDC recipient*	0.151	0.234	0.025	0.018	--	--	--	0.086
Cash welfare recipient*	--	--	0.054	0.031	--	--	0.093	0.072
SSI recipient	--	--	0.091	0.133	--	--	0.027	0.265
Offender*	--	0.057	--	--	--	--	--	--
Limited English speaking	--	--	--	--	--	--	0.259	0.251
Basic skills deficient	--	--	0.034	0.037	--	--	0.193	0.286
Reading skills below 7th grade	--	0.032	--	--	0.148	0.344	--	--
Lacking significant work history*	0.074	0.059	0.050	0.055	0.292	0.144	0.150	0.098
Unemployed 15 weeks or more*	0.111	0.103	0.086	0.073	--	0.242	0.091	0.076
Not in the labor force*	0.122	0.113	0.103	0.108	--	--	--	0.044
GA/RCA recipient	0.137	0.05	--	--	--	--	--	--
Veteran (Vietnam era veteran)	0.160	0.135	0.030	0.081	--	--	--	--
Homeless	--	--	--	0.043	0.595	0.602	--	0.136

*An adjustment factor that is included in our analysis.

Notes:

1. A factor was excluded from out analysis if either (1) the factor described hardly any or almost all individuals (i.e, its mean fell outside the [.1,.9] interval) or (2) factor information was missing for more than 10% of the observations (16% of observations lacked information on the variable Basic skills deficient, and 11% of observations lacked information on the variable Reading skills below 7th grade).
2. The data in this table come from *Guide for Setting JTPA Performance Standards for Program Year 1992-3* (1993), *Guide to JTPA Performance Standards for Program Years 1994 and 1995* (1995), *Guide to JTPA Performance Standards for Program Years 1996 and 1997* (1997), and *Guide to JTPA Performance Standards for Program Years 1998 and 1999* (1999), Menlo Park, Calif.: Social Policy Research Associates.

Table 2: Adjustment to the Baseline Level ($\Sigma_i \beta_i x_i$) for the Employment Rate Standard[(1)]

Program Year	Mean	Std. Dev.	Min.	Max	Number of Agencies
1993	25.405	4.793	11.760	49.428	639
1994	27.620	4.832	14.098	49.362	627
1995	27.250	5.045	13.992	50.293	665
1996	18.061	3.745	9.250	39.600	634
1997	17.415	3.781	5.400	36.072	663
1998	17.426	3.632	4.050	31.913	610
1993-1998	22.225	6.344	4.050	50.293	3838

(1) We compute for each agency the adjustment to the baseline level $\Sigma_i \beta_i x_i$ using the agency's actual enrollee population. The summary statistics reported here are based on the distribution of $\Sigma_i \beta_i x_i$ across agencies.

Table 3: Summary statistics

	Mean	SD	Min	25 Pctl	50 Pctl	75 Pctl	Max
Dependent Variables							
Enrollee Share	0.003	0.008	0.000	0.000	0.000	0.003	1.000
Employment Rate (%)	67.7	42.8	0.0	0.0	100	100	100
Earnings Outcome ($/week)	313.4	164.7	0.0	220.0	280.2	365.5	6997.9
Independent Variables							
Employment Weight	0.221	0.133	-0.066	0.130	0.202	0.202	0.286
Earnings Weight	0.527	0.418	-0.659	0.253	0.581	0.841	1.430

Table 4: Summary Statistics for Adjustment Factors

	Percentage of Terminees By Program Year						
Adjustment Factors	1993	1994	1995	1996	1997	1998	Total
Female*	63.03	65.80	66.50	67.56	66.88	65.64	65.81
55 years old & over	2.12	2.05	1.99	1.78	1.86	2.16	2.00
age 30 to 54*	56.45	56.32	56.52	57.25	57.67	57.65	56.91
Black*	29.52	30.04	31.85	33.26	33.55	34.61	31.91
Other minority*	18.58	19.49	20.58	22.16	21.03	23.08	20.64
Minority male*	17.49	16.99	17.64	17.95	17.91	19.67	17.85
High school dropout*	21.05	22.08	20.66	20.51	19.59	20.51	20.79
Post high school attendees*	21.26	22.28	23.69	24.26	24.51	23.81	23.21
High school dropout under 30	9.50	9.80	9.23	9.03	8.32	8.56	9.12
Handicapped	13.66	8.22	7.69	7.23	7.12	6.54	8.57
UI or UC claimant	14.26	10.62	9.40	9.56	9.34	9.36	10.57
Long-term AFDC recipient*	15.87	15.81	16.47	16.16	15.55	13.43	15.63
Cash welfare recipient*	39.61	42.91	42.25	40.08	37.28	32.33	39.41
SSI recipient	3.00	3.37	3.30	3.34	3.58	3.49	3.33
Offender*	13.05	13.46	13.57	14.82	16.06	17.26	14.52
Limited English speaking	5.10	4.68	4.18	4.48	4.10	4.84	4.58
Basic skills deficient	57.50	58.69	55.59	54.68	54.61	56.42	56.28
Reading skills below 7th grade	14.31	15.62	13.60	13.41	12.50	13.78	13.95
Lacking significant work history*	35.06	34.81	35.40	36.61	34.98	34.17	35.18
Unemployed 15 weeks or more*	41.49	36.54	33.36	31.26	31.46	31.16	34.62
Not in the labor force*	29.85	32.63	35.45	37.16	33.00	29.89	32.98
GA/RCA recipient	5.52	5.93	4.95	4.19	3.40	2.98	4.63
Veteran (Vietnam era veteran)	9.52	7.83	7.44	7.00	6.98	6.39	7.64
Homeless	4.38	2.22	2.42	2.63	2.57	2.40	2.81
Total Number of Terminees	138533	136834	121086	112822	110648	95653	715576

*Thirteen adjustment factors included in the analysis (see Table 1, Note 1)

Table 5
Determinants of Subgroup Enrolment Share

	(1)		(2)		(3)	
Independent Variable	Coef. Est.	Stand'ized Coef.	Coef. Est.	Stand'ized Coef.	Coef. Est.	Stand'ized Coef.
Employment Weight	**0.00759**	**0.00101**			**0.00773**	**0.00103**
	(0.00219)				**(0.00222)**	
	0.001				***0.014***	
Earnings Weight			0.00007	0.00003	-0.00042	-0.00018
			(0.00059)		(0.00059)	
			0.900		*0.865*	
Constant	0.00799		0.00777		0.00798	
	(0.00006)		(0.00001)		(0.00006)	
	0.000		*0.000*		*0.256*	
Observations	738689		738689		738689	
R-squared	0.49		0.49		0.49	

Notes:
1. Standard errors in parentheses.
2. P values in italics.
3. Errors clustered on training agencies.
4. All models include fixed effects (see equation 3 in text) .
5. All models are weighted by the subgroup-agency share of the entire terminee population.

Table 6 **Determinants of Subgroup Employment Outcome (ER)**				
	(1)		(2)	
Independent Variable	Coef. Est.	Stand'ized Coef.	Coef. Est.	Stand'ized Coef.
Employment Weight	**-14.86782**	**-2.03719**	**-15.42228**	**-2.11317**
	(3.42407) *0.000*		**(3.44840)** *0.000*	
Earnings Weight			1.36506	0.56749
			(1.34891) *0.312*	
Constant	1.84150		1.86583	
	(0.16014) *0.000*		(0.16243) *0.000*	
Observations	164488		164488	
R-squared	0.46		0.46	

Notes:
1. Standard errors in parentheses.
2. P values in italics.
3. Errors clustered on training agencies.
4. All models include fixed effects (see equation 4 in text).
5. All models are weighted by the subgroup-agency share of the entire terminee population.

Table 7 **Determinants of Subgroup Weekly Earnings Outcome (WE)**				
	(1)		(2)	
Independent Variable	Coef. Est.	Stand'ized Coef.	Coef. Est.	Stand'ized Coef.
Employment Weight			52.52690	6.92964
			(15.31844) *0.001*	
Earnings Weight	**3.02610**	**1.25155**	**-0.58678**	**-0.24268**
	(6.52741) *0.643*		**(6.39005)** *0.927*	
Constant	3.93647		6.66870	
	(0.35794) *0.000*		(0.93327) *0.000*	
Observations	122467		122467	
R-squared	0.52		0.52	

Notes:
1. Standard errors in parentheses.
2. P values in italics.
3. Errors clustered on training agencies.
4. All models include fixed effects (see equation 4 in text).
5. All models are weighted by the subgroup-agency share of the entire terminee population.

Table 8
Analysis omitting welfare recipients

	Dependent Variable					
	Enrolment Share		Employment Outcome		Earnings Outcome	
Independent Variable	Coef. Est.	Stand'ized Coef.	Coef. Est.	Stand'ized Coef.	Coef. Est.	Stand'ized Coef.
Employment Weight	0.00493	0.00052	-3.51101	-0.36899	37.22452	3.79134
	(0.00459) *0.284*		(5.44246) *0.519*		(23.48833) *0.114*	
Earnings Weight	0.00315	0.00129	-1.35605	-0.55022	-14.19245	-5.70765
	(0.00196) *0.109*		(1.75199) *0.439*		(8.09520) *0.080*	
Constant	0.01473		2.20412		1.18350	
	(0.00015) *0.000*		(0.22983) *0.000*		(1.16103) *0.309*	
Observations	366682		92187		71800	
R-squared	0.45		0.42		0.49	

Notes:
1. Standard errors in parentheses.
2. P values in italics.
3. Errors clustered on training agencies.
4. All models include fixed effects (see equations 3 and 4 in the text).
5. All models are weighted by the subgroup-agency share of the non-welfare recipient population.

Appendix 1: Derivations of H1 and H2

There are I demographic groups. Any applicant belongs to one and only one group. The training agency enrolls n_i applicants of demographic group i. The enrolment vector is denoted $n=(n_1,..,n_I)$ and the total number of enrollees is $N=\Sigma_i n_i$. The cost of training n_i enrollees of group i is $c_i(n_i)$ and the aggregate performance outcome for group i is $m_i(n_i)$ with $m_i(0)\geq 0$. Since more enrollees imply higher outcomes, we assume that $m'_i(n_i)>0$. As discussed in Section 3, we assume that $c'_i(n_i)>0$, $c''_i(n_i)\geq 0$ and $m''_i(n_i)<0$.

The aggregate performance outcome is the sum of performance outcomes over all groups, $M(n)=\Sigma_i m_i(n_i)$. The performance standard adjusts a baseline level m_0 for the enrollee composition

$$M_0(n)= m_0-\Sigma_i \beta_i n_i$$

where β_i is the adjustment weight for demographic group i. We denote β the vector of adjustment weights. The training agency is rewarded on the basis of excess performance

$$\Delta(n)=M(n)-M_0(n).$$

The training agency cares about the performance award and may also have its own preference over enrollees. The training agency has objective function $U(n,\Delta)$ where the first argument captures agency preferences over enrollee choices. To simplify, we consider the following functional form,

$$U(n,\Delta)=\Sigma_i \alpha_i n_i+\Delta$$

where α_i is a real number that captures the marginal preference attributed to demographic group i. The overall level of α defines how the training agency is willing to compromise its own preferences over enrolment for higher performance award. The training agency chooses n to maximize $U(n,\Delta(n))$ subject to the budget constraint $\Sigma_i c_i(n_i)\leq B$.

The designer may change one or more weights at a time. In general, the designer changes adjustment weight i by δ_i where $\delta=(\delta_1,..,\delta_I)$ is the vector of changes in weights. Denote $\bar{\delta}=1/N\sum_i \delta_i$ and the adjustment weight on measure i by $\beta_i+\varepsilon\delta_i$ where $\varepsilon\in[0,1]$.

Results

Proposition 1 derives a general result in the case where a single weight is changed, $\delta=(0..0,\delta_i=1,0..0)$. Proposition 2 considers any change in the vector of weights.

Proposition 1: (a) $dn_i/d\beta_i\geq 0$ and $dn_j/d\beta_i\leq 0$ for $j\neq i$ and these inequalities are strict for any interior solution ($n_i>0$). (b) $d[m_i(n_i)/n_i]/d\beta_i\leq 0$ and the inequality is strict for any interior solution.

Denote $n_i(\varepsilon)$ the number of enrollees of group i as a function of ε, $\Delta n_i=n_i(1)-n_i(0)$ the change in the number of enrollees of group i, and $\Delta(n_i/N)$ the same change measured in percentage terms. Similarly, we define $\Delta[m_i(n_i)/n_i]$ as the change in average performance. Proposition 2 derives general predictions on the impact of *any* change in the performance weights.

Proposition 2: Assume $c'_i(n)=c'_k(n)=c$ and $m''_i(n)=m''_k(n)=m$. (a) $\Delta n_i>\Delta n_k$ iff $\delta_i>\delta_k$. (a') a) $\Delta(n_i/N)>\Delta(n_k/N)$ iff $\delta_i>\delta_k$. (b) $\Delta[m_i(n_i)/n_i]<\Delta[m_k(n_k)/n_k]\leq 0$ if $\delta_i>\bar{\delta}>\delta_k$. (c) If $m_i(0)=0$ and $\beta_i=\beta$ then $\Delta[m_i(n_i)/n_i]<\Delta[m_k(n_k)/n_k]\leq 0$ if $\delta_i>\delta_k$.

Proposition 2 holds if the cost and performance measure functions have a linear and quadratic structure respectively, $c_i(n_i)=c_{o,i}+c_1 n_i$ and $m_i(n_i)= m_{0,i}+m_{1,i}n_i+m_2 n_i^2$. It does not say

anything about the direction of the change in the number of enrollees of group i or k. The total number of enrollees of group i could increase or decrease and similarly for group k. The proposition makes a prediction on the relative change in enrollees. The assumptions stated in Proposition 2 are necessary and sufficient for claim (a). Without these assumptions, one

cannot sign $\dfrac{\frac{dn_i}{d\varepsilon}}{\frac{dn_k}{d\varepsilon}} = \dfrac{\sum_j \frac{c_j'}{\lambda c_j'' - m_j''}(\delta_i c_j' - \delta_j c_i')}{\sum_j \frac{c_j'}{\lambda c_j'' - m_j''}(\delta_k c_j' - \delta_j c_k')}$ in general.

Proofs

We first derive a general result that is used in the proofs of both propositions. Denote λ the Lagrange multiplier on the budget constraint. In any interior solution ($n_i>0$), the first order condition says

$$m'_i(n_i)+\alpha_i+\varepsilon\delta_i+\beta_i =\lambda c'_i(n_i).$$

Take derivative of the first order condition and budget constraint with respect to ε.

$$\begin{cases} m_i'' \dfrac{dn_i}{d\varepsilon} + \delta_i = \lambda c_i'' \dfrac{dn_i}{d\varepsilon} + c_i' \dfrac{d\lambda}{d\varepsilon} \\ \sum_j c_j' \dfrac{dn_j}{d\varepsilon} = 0 \end{cases}$$

Compute the value of $\dfrac{d\lambda}{d\varepsilon}$ as

$$\frac{d\lambda}{d\varepsilon} = \frac{\sum_j \frac{\delta_j c_j'}{\lambda c_j'' - m_j''}}{\sum_j \frac{c_j^{2\prime}}{\lambda c_j'' - m_j''}}.$$

Plugging back into the first order condition gives

$$\frac{dn_i}{d\varepsilon} = \frac{\sum_j \frac{c_j'}{\lambda c_j'' - m_j''}(\delta_i c_j' - \delta_j c_i')}{(\lambda c_i'' - m_i'')\sum_j \frac{c_j^{2\prime}}{\lambda c_j'' - m_j''}} \quad . \qquad \text{(A)}$$

Proof of Proposition 1

(a) Set $\delta_i=1$ and $\delta_j=0$ for $j\neq i$ in expression (A) and conclude using the identity $dn_j/d\alpha_i = dn_j/d\varepsilon$.

(b) Taking derivative of the average performance of group i with respect to α_i

$$\frac{d}{d\alpha_i}\left[\frac{m_i(n_i)}{n_i}\right] = \frac{(m_i' n_i - m_i)}{n_i^2}\frac{d}{d\alpha_i} n_i$$

But since m_i is concave, we have $m'(n)n \leq m(n)-m(0)$, and the assumption $m(0)\geq 0$ implies $m_i' n_i - m_i < 0$. We conclude that $d[m_i(n_i)/n_i]/d\alpha_i < 0$. QED

Proof of Proposition 2

(a) Under the assumptions stated in proposition 2, expression (A) becomes

$$\frac{dn_i}{d\varepsilon} = \frac{-1}{m''}\left(\delta_i - \bar{\delta}\right).$$

$\delta_i > \delta_k$ implies $\frac{dn_i}{d\varepsilon} > \frac{dn_k}{d\varepsilon}$ and since $\frac{dn_i}{d\varepsilon}$ is linear in δ_i we have $\Delta n_i/\Delta n_k > 1$.

(b) We have, $(d/d\varepsilon)[m_i(n_i)/n_i] = -K_i(n_i)\left(\delta_i - \bar{\delta}\right)$ where K_i is a positive function.

(c) We have, $(d/d\varepsilon)[m_i(n_i)/n_i] = -K\left(\delta_i - \bar{\delta}\right)$ where K is a positive constant. QED

Appendix 2: Demographic Subgroup and Variable Construction

The use of the DOL PAW methodology was not mandated in JTPA and states could opt out of using them. We contacted all state agencies that had been in charge of administering JTPA and asked them whether they had used the PAW methodology during the time period of our study, program years 1993 through 1998. Of the 33 states that supplied this information, 29 indicated they used the methodology and 4 indicated they did not. We include in our analysis only the 463 training agencies residing in the 29 states that used the PAW.

Construction of the demographic subgroups: For our empirical analysis we use demographic subgroups rather than demographic characteristics. There were 24 adjustment factors used in the DOL's adjustment model during our time period. All factors are binary (e.g. male/female). We omitted from our analysis the very small demographic subgroups by dropping the nine factors for which the factor's minority realization represented 10 percent or fewer JTPA participants at any program year during our study time period (for example, because only 3 percent of participants were SSI recipient for six program years, we omitted the"SSI recipient" factor).[23] We omitted two more factors due to missing demographic information on program participants for these factors.[24] In the end, we used 13 adjustment factors in the analysis. These factors are marked with a start (*) in Table 1. These 13 adjustment factors yield 8,192 ($=2^{13}$) demographic subgroups for each of 463 training agencies, and thus 3,792,896 ($=8{,}129\times463$) possible subgroup-agency combinations for each program year.

Many of these 3,792,896 subgroup-agency cells were empty, prompting us to further limit the data. We excluded from the analysis all subgroup-agency combinations that had zero terminees in each of the six program years. Applying this criterion led us to drop 96% of the 3,792,896 possible subgroup-agency combinations. The final panel data includes 1,670 different subgroups for 463 agencies. The number of subgroups vary across agencies (Min=2, Max=1073) and there are on average 291.04 subgroups per agency. There are 134,755 ($=463\times291.04$) subgroup-agency observations by program year. The final analysis for enrolment share used 738,689 observations which is less than the total number of observations (808,530=134,755$\times$6 program years) due to missing data on local economic conditions.

Construction of the enrolment shares: The shares of subgroup terminees and the average performance outcomes were computed for each subgroup-agency-program year cell using data from the Standardized Program Information Report (SPIR). In the 463 agencies where the PAW were used, we have complete demographic information for 682,515 adult terminees,

[23] Nine factors excluded from the analysis are 55 years old & over, High school dropout under 30, Handicapped, UI or UC claimant, SSI recipient, Limited English speaking, GA/RCA recipient, Veteran, and Homeless.

[24] Our cutoff for inclusion in the analysis was a 90 percent data availability rate. Thus because information about being "basic skills deficient" was reported for only 84% of participants and possessing "reading skills below 7th grade" was reported for only 89 percent of participants, those factors were omitted from our analysis.

which accounts for 63% of the entire JTPA adult population during the 1993-1998 period.[25] This subsample appears to be representative of the entire enrollee population, however. For example, our sample includes 66% female, 32% black, 21% high school dropouts, and 40% welfare recipients and these figures are almost identical to the ones corresponding to the JTPA population (66% female, 32% black, 22% high school dropouts, and 37% welfare recipients).

Construction of the performance outcomes: Since under JTPA the follow-up performance outcomes were measured for only a subset of all terminees, the samples for the employment and earnings outcomes analysis (H2) are smaller than the sample for the enrolment analysis (H1).[26] SPIR reports a follow-up employment outcome for 44% of terminees (N=297,352) and a follow-up weekly earnings outcome for 72% of the follow-up employment outcome sample (N=213,176). We construct the subgroups for the outcome analysis using the same method as above and obtain 164,488 subgroups-agency-year observations for the employment outcome and 122,467 subgroups-agency-year observations for the earnings outcome.

[25] The percentages of terminees included in our analysis (relative to the entire population) for program years 1993, 1994, 1995, 1996, and 1997 were 53%, 65%, 65%, 67%, 65%, and 66%, respectively.

[26] To save money, JTPA administrators estimated each training agency's overall performance from the performance of a sample of terminees drawn randomly from the training agency's terminee population.