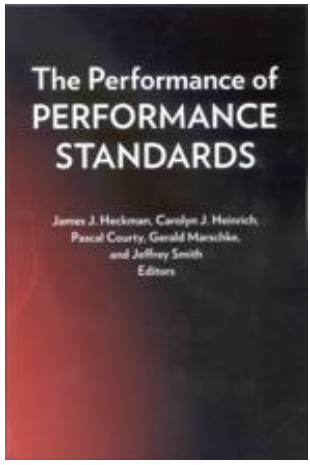
Do Short-Run Performance Measures Predict Long-Run Impacts?



James J. Heckman
University of Chicago

Carolyn J. Heinrich
University of Wisconsin

Jeffrey Smith
University of Michigan

Chapter 9 (pp. 273-304) in:
The Performance of Performance Standards
James J. Heckman, Carolyn J. Heinrich, Pascal Courty, Gerald Marschke, and Jeffrey Smith, eds.
Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, 2011
DOI: 10.17848/9780880993982.ch9

The Performance of Performance Standards

James J. Heckman
Carolyn J. Heinrich
Pascal Courty
Gerald Marschke
Jeffrey Smith
Editors

2011

W.E. Upjohn Institute for Employment Research
Kalamazoo, Michigan

Library of Congress Cataloging-in-Publication Data

The performance of performance standards / James J. Heckman . . . [et al.], editors.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-88099-292-3 (pbk. : alk. paper)
ISBN-10: 0-88099-292-1 (pbk. : alk. paper)
ISBN-13: 978-0-88099-294-7 (hardcover : alk. paper)
ISBN-10: 0-88099-294-8 (hardcover : alk. paper)
1. Government productivity. 2. Performance standards. 3. Civil service—Personnel management. I. Heckman, James J. (James Joseph)
JF1525.P67P476 2011
352.6'7—dc22

2011007877

W.E. Upjohn Institute for Employment Research
300 S. Westnedge Avenue
Kalamazoo, Michigan 49007-4686

Cover design by Alcorn Publication Design.
Index prepared by Diane Worden.
Printed in the United States of America.
Printed on recycled paper.

9
Do Short-Run Performance Measures Predict Long-Run Impacts?

James J. Heckman
Carolyn J. Heinrich
Jeffrey Smith

This chapter culminates the analysis in this volume by examining two closely related questions.[1] The first of these is posed in the title: Do performance measures based on short-run outcomes predict long-run program impacts? If they do, then performance management systems like those in JTPA and WIA will provide incentives that enhance the economic efficiency of program operations. Put differently, if existing performance measures predict long-term impacts, then their use provides some benefits to weigh against the costs documented in earlier chapters. The second question concerns the efficiency costs of cream skimming induced by the performance standards. As noted in Chapter 3, depending on the relationship between the performance measures and net program impacts, cream skimming may be efficiency increasing (a positive relationship), efficiency decreasing (a negative relationship), or neutral (no relationship).

We address these questions in two different ways. The two analyses build on different identifying assumptions but both utilize the experimental data from the National JTPA Study (NJS) introduced in Chapter 6. The two analyses represent different ways of dealing with the fact that, absent additional assumptions, experimental data do not provide impacts for individuals, only average impacts for groups. Both strategies have important limitations, which we discuss in detail later on in the chapter.

Both methods yield the same basic findings. First, the short-run labor market outcomes commonly used as performance measures do

not predict long-run impacts. Indeed, in some cases we find a perverse relationship, indicating that the performance measures actually provide an incentive for program staff to move away from, rather than toward, economic efficiency. Second, we find little evidence of an efficiency cost associated with cream skimming; if anything, it may provide a small efficiency gain.

NJS DATA

We use data gathered as part of the NJS, an experimental evaluation of the JTPA program described in Chapters 2 and 4, for the analyses in this chapter. The experiment was conducted at 16 of the more than 600 JTPA training centers (which we will also refer to as sites). Table 9.1 lists the sites that volunteered to participate in the experiment and provides some descriptive statistics. Columns one through three indicate the racial/ethnic composition of the adult participant population during the study, while the fourth column indicates adult participants' average years of schooling. The fifth and sixth columns display unemployment and poverty rates.

The final three columns indicate the fraction of participants assigned to each of the three experimental treatment streams, based on the services recommended for them prior to random assignment. The classroom training in occupational skills (CT-OS) stream includes individuals who were recommended to receive CT-OS and possibly other services not including subsidized on-the-job training (OJT) at private firms. The OJT treatment stream includes individuals recommended to receive OJT and possibly other services not including CT-OS. The other services stream is a residual category that, with only a few exceptions, includes individuals not recommended to receive either CT-OS or OJT. As illustrated in Exhibit 3.17 of Orr et al. (1996), individuals in the CT-OS stream usually received classroom training whether in the form of basic education or CT-OS or both. Those in the OJT stream often did not enroll; when they did enroll they tended to receive OJT or, somewhat less often, job search assistance. Individuals in the "other" treatment stream received a wide variety of services.

Table 9.1 Descriptive Statistics for the 16 Sites in the National JTPA Study

	Fraction of participants that are:			Avg. yrs. of schooling for participants	Unemp. rate	Poverty rate	Fraction of participants assigned to:		
Site	White	Black	Hispanic				CT-OS stream	OJT stream	Other services stream
Corpus Christi, TX	23.3	10.4	65.5	11.2	10.2	13.4	34.3	51.5	14.1
Cedar Rapids, IA	87.8	7.6	1.3	11.6	3.6	6.0	60.0	35.4	4.6
Coosa Valley, GA	82.1	17.1	0.6	10.7	6.5	10.7	36.1	38.1	25.7
Heartland of FL	50.2	45.7	2.8	11.4	8.5	11.3	28.9	27.1	44.0
Fort Wayne, IN	72.3	23.7	2.8	11.5	4.7	5.9	6.4	66.2	27.3
Jersey City, NJ	6.3	68.6	20.3	11.5	7.3	18.9	46.0	35.7	18.3
Jackson, MS	13.9	85.5	0.3	12.2	6.1	12.8	57.9	35.5	6.6
Larimer County, CO	77.9	1.8	17.0	12.2	6.5	5.9	29.6	7.1	63.3
Decatur, IL	68.1	31.9	0.0	11.8	9.2	7.8	14.4	79.1	6.5
Northwest MN	81.3	1.8	10.9	11.4	8.0	11.1	25.6	74.0	0.4
Butte, MT	86.6	0.3	5.0	11.7	6.8	7.5	26.6	40.1	33.3
Omaha, NE	38.6	53.4	4.2	11.7	4.3	6.7	77.4	18.9	3.7
Marion, OH	95.6	2.3	0.9	11.3	7.0	7.2	48.8	41.8	9.4
Oakland, CA	8.0	68.3	6.8	12.4	6.8	16.0	49.6	7.9	42.6
Providence, RI	33.6	33.9	24.6	11.3	3.8	12.1	32.3	13.0	54.7
Springfield, MO	96.1	1.8	0.0	11.9	5.5	10.1	17.7	74.6	7.7

SOURCE: Race/ethnicity and years of schooling for adult participants come from calculations by the authors using the National JTPA Study data. Race/ethnicity categories do not necessarily sum to one due to the omission of "other." Unemployment rates are from Orr et al. (1996, Exhibit 3.3) and are unweighted annual averages for 1987–1989. Poverty rates come from Orr et al. (1996, Exhibit 3.2) and are for 1979. The treatment stream recommendation fractions for adults come from Kemple, Doolittle, and Wallace (1993, Table 7.1).

The site selection strategy for the evaluation excluded sites with small enrollments for cost reasons. Attempts to gain external validity among larger sites by selecting sites at random failed due to high refusal rates, as described by Doolittle and Traeger (1990) and Hotz (1992). Without random site selection, external validity in the strict sense clearly fails. At the same time, Table 9.1 makes clear that the 16 sites represent a diverse mix in terms of participant demographics, local economic conditions, and service mix. Doolittle and Traeger (1990, Section 5) compare the 16 experimental sites to the population of all JTPA sites and find that, on average, the two groups look much alike. In our view, these patterns make our results suggestive, rather than either definitive or irrelevant, when generalized to the JTPA program more broadly.

At the experimental centers, persons who applied to and were accepted into the program were randomly assigned to either a treatment group allowed access to JTPA services or to a control group denied access to JTPA services for the next 18 months. A short survey at the time of random assignment collected background information on demographic characteristics, educational attainment, work history, past training receipt, current and past transfer program participation, and family income and composition. This survey was self-administered with assistance from program staff; it achieved a response rate well over 90 percent as well as only modest item nonresponse conditional on survey response. We use variables from this baseline survey to define our subgroups (and for the participant descriptive statistics in Table 9.1).

In addition, follow-up surveys collected information on employment and earnings around 18 months after random assignment and, for a random subsample, at around 30 months after random assignment. The response rates for the two surveys were 83 and 77 percent, respectively, with little difference between the experimental treatment and control groups (see Appendix A of Orr et al. [1994]). Both the program and the experimental analysis divided participants into four groups based on age and sex: adult males and females aged 22 and above and male and female out-of-school youth aged 16–21 (the NJS did not examine the component of JTPA serving in-school youth). We examine only adult males and females in this chapter due to the small samples available for the two youth groups.

We use the data on wages, earnings, and employment from the follow-up surveys to construct the performance measures and outcome vari-

ables. Our outcome variables consist of earnings and employment for 18 or 30 months after random assignment. For our analyses using percentiles, we use all observations with valid values of earnings over the 18 months after random assignment. For the analyses using subgroup variation in experimental impacts, we trim the top 1 percent of the earnings values. The employment variables measure the fraction of months employed, where we code an individual as employed in a month if they have positive earnings in that month.

The JTPA performance measures we analyze are hourly wage and employment at termination from the program and weekly earnings and employment 13 weeks after termination. In most states at this time, program staff members obtained these outcomes via telephone surveys of participants. We do not have access to the telephone survey data for our sample and instead use program termination dates from JTPA administrative data combined with data from the follow-up surveys on job spells to construct the performance measures. Because program administrators did not necessarily contact participants on the exact date of termination or follow-up (and to allow for some measurement error in the timing of the self-reported job spells), we count all job spells within 30 days on either side of the termination date (or 13 weeks after termination, as appropriate) in constructing the performance measures. We measure employment based on the presence or absence of a job spell within this window. For the wage measure, we use the highest hourly wage within the window for persons holding more than one job. For the earnings measure, we take the average weekly earnings on all jobs over the 61-day window. Following the definition of the corresponding official performance measures, we calculate hourly wages and weekly earnings for employed persons only.

For more information on the NJS experimental data, see the official impact reports in Bloom et al. (1997) and Orr et al. (1996), the official implementation reports in Doolittle and Traeger (1990) and Kemple, Doolittle, and Wallace (1993), and related papers on the design and the data by Hotz (1992), Smith (1997), Kornfeld and Bloom (1999), and Heckman and Smith (2000). For discussions of interpretational issues see Heckman, Smith, and Clements (1997), Heckman, Smith, and Taber (1998) and Heckman et al. (2000).

ECONOMETRIC ANALYSIS STRATEGIES: NOTATION AND MOTIVATION

Ideally, we would like to relate individual program impacts to individual values of the performance measures. Unfortunately, as discussed in, e.g., Heckman (1992); Heckman, Smith, and Clements (1997); Heckman and Smith (1998); and Djebbari and Smith (2008), without additional assumptions, even experimental data do not allow us to generate individual-level impact estimates.

To consider this issue more carefully, we return to the notation defined in Chapter 3. Recall that $Y^1_{a,i}$ denotes a labor market outcome for person i in some period a given treatment, where the 1 superscript denotes treatment. Similarly, $Y^0_{a,i}$ denotes a labor market outcome in the same period given no treatment, implying that the impact for individual i in period a equals $Y^1_{a,i} - Y^0_{a,i} = \Delta_{a,i}$. In this chapter, we distinguish between two periods: the short run, denoted by s, and the long run, denoted by l. Both periods begin at the time the individual decides to participate or not. In terms of this notation, we would ideally like to relate $Y^1_{s,i}$ and $\Delta_{l,i}$. Finally, recall that S denotes the set of individuals treated.

Experimental data consist of the marginal distributions of outcomes in the treated and untreated states, that is, $f(Y^0_a)$ and $f(Y^1_a)$. Experimental data to not identify the joint distribution of outcomes, $f(Y^0_a, Y^1_a)$, and therefore do not identify individual impacts. Experimental data do identify mean impacts for subgroups of individuals defined by characteristics not affected by the treatment (which usually means those observed prior to random assignment). Letting g denote some particular subgroup (such as those with exactly 12 years of schooling) out of a set G, we can construct the impact estimate for the subgroup by taking a mean difference between the treated and untreated units in subgroup g. More formally, we estimate the subgroup impact $\Delta_{a,g} = E(Y^1_a \mid G = g) - E(Y^0_a \mid G = g)$ by replacing the conditional expectations with the corresponding sample means.[2]

The next two sections describe the strategies we employ to deal with the lack of individual impact estimates. The first strategy imposes additional, nonexperimental, assumptions on the data that allow us to construct individual impact estimates. The second strategy relies solely

on subgroup variation in the experimental impacts and, as such, requires no additional assumptions.

ECONOMETRIC ANALYSIS STRATEGIES: RANK PRESERVATION

Our first econometric strategy builds on the assumption of rank preservation outlined in Heckman, Smith, and Clements (1997).[3] We assume that the joint distribution of treated and untreated outcomes takes a very simple form: the counterfactual for each quantile of the treated outcome distribution consists of the corresponding quantile of the untreated outcome distribution. Thus, for example, the counterfactual outcome for the median treated person consists of the outcome of the median untreated person. Note that under this assumption, cream skimming on Y_a^0 implies the same choices as cream skimming on Y_a^1. We can think of the simple world defined by the rank preservation assumption as a "one factor" world in which those who do well in the treated state also do well in the untreated state and those who do poorly in the treated state also do poorly in the untreated state.

This assumption may seem quite unusual, but in fact it nests the widely (though often implicitly) used common effect model in which $\Delta_{a,i} = \Delta_a$. In the common effect world, the treatment has the same effect on all participants. In this world, the treated outcome distribution has the same shape as the untreated outcome distribution but its location differs by the common treatment effect. For example, if the untreated outcomes have a normal distribution with mean 100 and variance 20, and the common treatment effect equals 10, then the treated outcomes have a normal distribution with mean 110 and variance 20. Moreover, in the common effect world, quantiles of the treated and untreated outcome distributions again form counterfactuals for one another. The rank preservation assumption relaxes the assumption of an equal treatment effect for all participants while keeping the link between the quantiles of the two outcome distributions. It therefore nests the common effect model as a special case.

More formally, following Heckman, Smith, and Clements (1997), if each individual has the same rank in the distributions of Y_a^0 and Y_a^1,

then we can associate a Y_a^0 with each Y_a^1; continuity of the two distributions implies a unique association. The assumptions of rank preservation plus continuity allow us to construct Δ_a as a function of Y_a^0 (or, what is the same thing, of Y_a^1). We operationalize this idea by taking percentile differences across the treated and untreated outcome distributions.[4] Let $Y_a^{0,j}$ denote the jth percentile of the Y_a^0 distribution, with $Y_a^{1,j}$ the corresponding percentile in the Y_a^1 distribution. Thus, we estimate $\Delta_a(Y_a^{0,j}) = Y_a^{1,j} - Y_a^{0,j}$. Our data include mass points at zero earnings in both the treated and untreated distributions. For the corresponding percentiles we simply assign an impact of zero; because all of the outcomes equal zero in the lower percentiles of the two distributions, order does not matter. Thus, the lack of a unique association in this part of the distribution poses no problems in our application.

ECONOMETRIC ANALYSIS STRATEGIES: SUBGROUP VARIATION IN EXPERIMENTAL IMPACTS

Our second identification strategy relies solely on the exogenous variation in treatment status induced by the experiment. As noted above, as a result of random assignment, we can construct unbiased mean impact estimates for subgroups defined by variables observed prior to random assignment.

To implement this strategy, we form 43 subgroups based on the following characteristics measured at the time of random assignment: race, age, education, marital status, time since most recent employment, receipt of Aid to Families with Dependent Children (AFDC—the predecessor to the current Temporary Aid to Needy Families program), receipt of Food Stamps, and training center. Individuals with complete data belong to eight subgroups, while we include those with incomplete data in as many subgroups as their data allow. Using a regression framework, we construct mean-difference experimental impact estimates for each subgroup.[5] We adjust these estimates by dividing through by the fraction enrolled in each subgroup to reflect the fact that a substantial fraction of persons (41 percent of adult males and 37 percent of adult females) in the treatment group dropped out and did not participate in JTPA.[6] We construct the subgroup average performance measures by

simply averaging the individual performance measures over the members in each subgroup.

RESULTS BASED ON THE RANK PRESERVATION ASSUMPTION

Figures 9.1 and 9.2 present estimates of $\Delta_a(Y_a^{0,j})$ constructed under the rank preservation assumption. Self-reported earnings in the 18 months after random assignment constitute the outcome variable. The horizontal axis in each figure indicates percentiles of the treated and untreated (i.e., control) outcome distributions. The vertical axis indicates the difference in outcomes at each percentile.

We begin with the estimates for adult women in Figure 9.1, for whom the sample size is the largest. First, we observe zero impacts through the 20th percentile. This region corresponds to persons with zero earnings in the 18 months after random assignment in both the treated and untreated states under the rank preservation assumption. Second, we observe a relatively constant positive treatment effect of around $800 over the interval from the 20th to the 90th percentile. Third, we note a discernible increase in the estimated impact in the final decile. Assuming roughly equal costs among participants at different percentiles, the pattern in Figure 9.1 suggests that cream skimming beyond the 20th percentile has little effect on the economic efficiency of JTPA. However, a policy of targeting services at the bottom two deciles entails clear costs. To the extent that the untreated outcome proxies for the performance measures, Figure 9.1 suggests only a very modest (and very nonlinear) positive relationship between the performance measures and the impacts.

Figure 9.2 for adult men tells a similar tale. We observe a relatively flat relationship over the range from the 10th to the 50th percentile, after which it dips and then rises again. Given the wide standard errors (and the smaller region of zero impacts at the lowest percentiles) we can say with some (but not overwhelming) confidence that cream skimming, in regard to adult males, also likely has little effect, either positive or negative, on efficiency. And, to the extent that the untreated outcomes

Figure 9.1 Quantile Treatment Effects, Adult Males

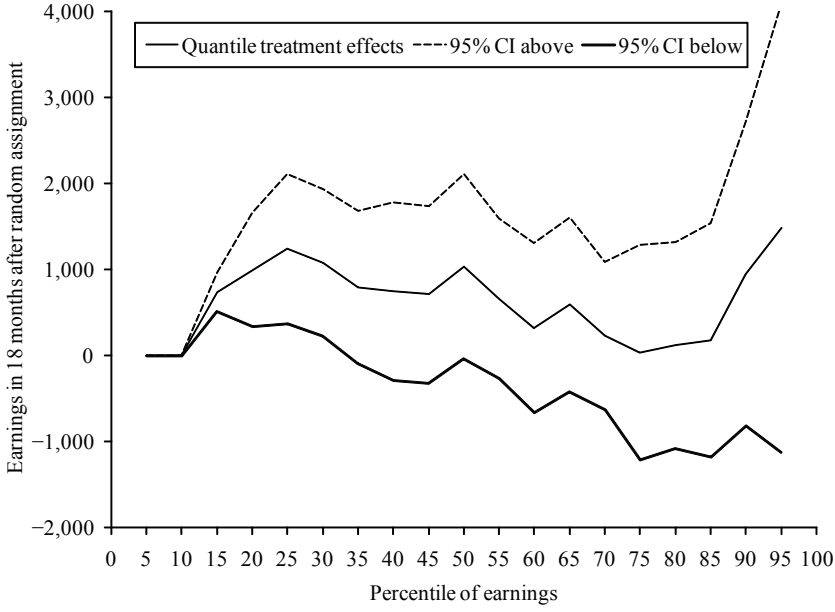
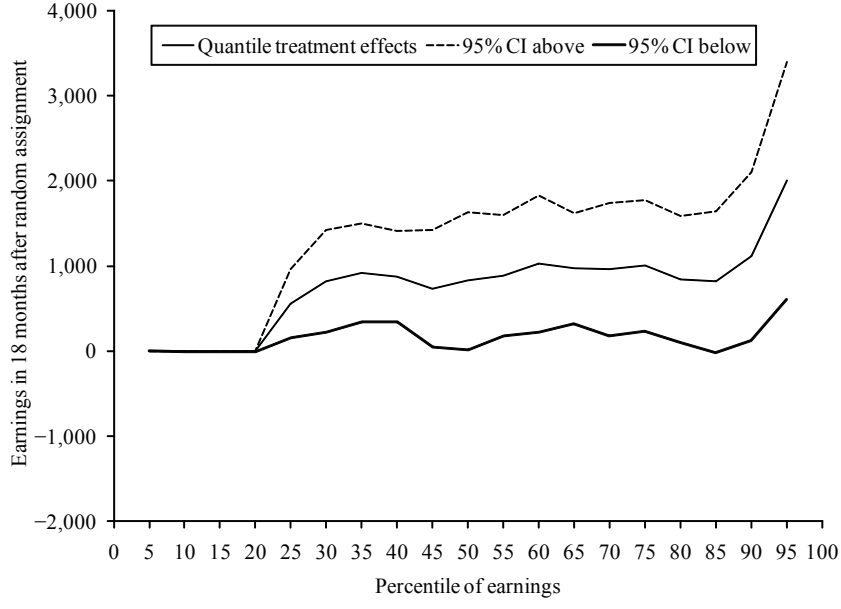


Figure 9.2 Quantile Treatment Effects, Adult Females



proxy for the performance measures, we see little relationship between the two for adult males under the rank preservation assumption.

RESULTS BASED ON SUBGROUP VARIATION: UNIVARIATE

In this section we examine the experimental impact estimates for subgroups defined by individual baseline characteristics. Put differently, we examine the correlation between predictors of Y_a^1 and impacts conditional on values of those predictors. Caseworkers may use specific variables, such as labor force status, to help them forecast short-run outcomes as part of a strategy to select as participants individuals likely to do well on the performance measures. Moreover, the relationship between such characteristics and Δ_a is of interest in its own right.

Tables 9.2 and 9.3 summarize subgroup estimates of the impact of JTPA on the earnings and employment of adult females and adult males in the JTPA experiment, respectively. The first column in each table lists the values of each subgroup variable. Columns two through five present impacts on earnings in the 18 and 30 months after random assignment and on the fraction of months employed in the 18 and 30 months after random assignment. Note that the samples differ for the 18 and 30 month outcomes due to survey nonresponse in the second follow-up survey. We also present p-values from tests of the null of equal impacts among the subgroups defined by each variable. We present subgroup impacts conditional on labor force status (employed, unemployed, and out of the labor force), highest grade completed, AFDC receipt and month of last employment (if any), all measured at the time of random assignment. All of these variables predict the level of the 18-month and 30-month outcomes for participants.

For adult females, we reject the null of equal impacts among subgroups in 4 of the 16 possible cases. Two of the rejections (at the 5 percent level) occur for employment over 18 months and earnings over 30 months conditional on AFDC receipt, with larger impacts in each case for women receiving AFDC. As AFDC receipt is negatively related to Y_a^1, this finding suggests that cream skimming may be somewhat inefficient for adult women. The other two rejections occur for earnings and employment over 30 months conditional on month of last employment.

Table 9.2 Experimental Impact Estimates by Subgroup, Adult Females

Subgroup	Earnings impacts ($)		Employment impacts	
	18 months	30 months	18 months	30 months
	Labor force status			
P-value for equal impacts	0.3919	0.5745	0.4715	0.2286
Employed	1,223.78	1,487.38	0.0017	−0.0158
	(651.64)	(2,461.08)	(0.0135)	(0.0168)
Unemployed	507.42	428.84	0.0112	0.0184
	(507.92)	(1,715.10)	(0.0112)	(0.0128)
Out of the labor force	1,543.72	3,274.29	0.0274	0.0184
	(601.48)	(2,089.21)	(0.0160)	(0.0188)
	Education			
P-value for equal impacts	0.6890	0.4641	0.8149	0.4646
Highest grade completed < 10	1,029.22	−2227.56	0.0135	0.0175
	(643.40)	(2,577.38)	(0.0164)	(0.0182)
Highest grade completed 10–11	1,341.37	3,088.46	0.0289	0.0246
	(592.06)	(2,179.51)	(0.0147)	(0.0171)
Highest grade completed 12	460.29	1503.23	0.0129	−0.0053
	(469.73)	(1,711.16)	(0.0109)	(0.0129)
Highest grade completed > 12	971.20	795.14	0.0115	0.0209
	(816.54)	(2,997.34)	(0.0172)	(0.0211)
	AFDC receipt			
P-value for equal impacts	0.7224	0.0371	0.0277	0.2607
Not receiving AFDC	712.26	−947.01	0.0028	0.0026
	(392.05)	(1,462.17)	(0.0087)	(0.0105)
Receiving AFDC	924.57	3,624.35	0.0343	0.0211
	(451.07)	(1,631.02)	(0.0113)	(0.0127)
	Recent employment			
P-value for equal impacts	0.8614	0.0492	0.5708	0.0139
Currently employed	1,104.08	396.24	0.0138	0.0056
	(721.42)	(2,851.27)	(0.0151)	(0.0197)
Last employed 0–2 months ago	594.01	979.22	0.0099	0.0060
	(713.69)	(2,485.38)	(0.0161)	(0.0181)
Last employed 3–5 months ago	171.44	−7,677.17	−0.0063	−0.0589
	(953.91)	(3,485.31)	(0.0199)	(0.0220)
Last employed 6–8 months ago	1,874.38	975.22	0.0451	0.0502
	(1,175.53)	(3,721.12)	(0.0263)	(0.0305)

Table 9.2 (continued)

Subgroup	Earnings impacts ($)		Employment impacts	
	18 months	30 months	18 months	30 months
	Recent employment			
Last employed 9–11 months ago	1,679.73	5,244.59	0.0310	0.0636
	(1,311.91)	(4,437.63)	(0.0305)	(0.0382)
Last employed ≥ 12 months ago	1,304.36	4,919.73	0.0341	0.0347
	(587.15)	(2,020.46)	(0.0155)	(0.0180)
Never employed	610.59	−2,490.44	0.0335	−0.0059
	(609.42)	(2,736.46)	(0.0168)	(0.0191)

NOTE: Monthly earnings are based on self-reports with top 1 percent trimming. Estimates are adjusted for program dropouts in the treatment group. Earnings impacts are calculated using all sample members with valid observations for self-reported monthly earnings during each period. The sample includes 4,886 valid observations for the 18-month period after random assignment and 1,147 valid observations for the 30-month period after random assignment. Heteroskedasticity-consistent standard errors appear in parentheses.

SOURCE: Heckman, Heinrich, and Smith (2002).

We have trouble interpreting these estimates, which do not reveal any obvious systematic pattern.

Considering the estimates in Table 9.2 more broadly, we see three patterns. First, we lack the data to precisely estimate most of the subgroup impacts. Second, the point estimates often suggest very different impacts by subgroup. Third, the subgroup impact estimates often change substantially between 18 and 30 months. Taken together, these findings leave us with a lot of uncertainty about the efficiency effects of cream skimming. At the same time, it seems unlikely that caseworkers, who receive little feedback about the long-run labor market outcomes of participants at either the individual or aggregate level, have more information about these patterns than we do. Thus, any efforts to select participants based on these observed variables will likely have little systematic relationship to impacts, a conclusion quite consistent with the finding in Bell and Orr (2002) and Lechner and Smith (2007) that caseworkers cannot predict impacts.

For adult males, statistically significant differences in impacts among subgroups defined by our set of characteristics emerge only once, for impacts on 18-month earnings conditional on labor force sta-

Table 9.3 Experimental Impact Estimates by Subgroup, Adult Males

Subgroup	Earnings impacts ($)		Employment impacts	
	18 months	30 months	18 months	30 months
	Labor force status			
P-value for equal impacts	0.0407	0.3469	0.2679	0.6517
Employed	2,839.24	6,328.20	0.0300	0.0005
	(1,145.51)	(4,143.22)	(0.0166)	(0.0194)
Unemployed	718.84	3,021.68	0.0056	0.0180
	(710.16)	(2,339.51)	(0.0105)	(0.0125)
Out of the labor force	−2,193.85	−2,725.72	−0.0163	0.0289
	(1,658.81)	(4,693.28)	(0.0262)	(0.0281)
	Education			
P-value for equal impacts	0.6077	0.7939	0.9587	0.7206
Highest grade completed < 10	680.26	1,713.46	0.0114	0.0403
	(1,193.62)	(3,935.62)	(0.0203)	(0.0225)
Highest grade completed 10–11	−64.77	−270.18	0.0120	0.0134
	(1,020.79)	(3,516.67)	(0.0163)	(0.0188)
Highest grade completed 12	1,438.13	552.70	0.0030	0.0105
	(793.68)	(2,729.26)	(0.0119)	(0.0141)
Highest grade completed > 12	−92.00	4,886.81	0.0116	0.0201
	(1,238.21)	(4,155.34)	(0.0172)	(0.0221)
	AFDC receipt			
P-value for equal impacts	0.5948	0.5794	0.3813	0.6678
Not receiving AFDC	722.73	2,933.22	0.0122	0.0161
	(556.43)	(1,810.58)	(0.0085)	(0.0099)
Receiving AFDC	−232.18	−274.82	−0.0132	0.0306
	(1,706.56)	(5,495.50)	(0.0278)	(0.0322)
	Recent employment			
P-value for equal impacts	0.5995	0.6193	0.9112	0.7010
Currently employed	2,668.20	3,053.96	0.0176	−0.0134
	(1,230.61)	(4,174.11)	(0.0178)	(0.0212)
Last employed 0–2 months ago	816.36	6,126.54	0.0168	0.0205
	(1,091.14)	(3,637.23)	(0.0152)	(0.0180)
Last employed 3–5 months ago	−425.61	1,248.64	0.0037	0.0119
	(1,162.99)	(3,794.83)	(0.0176)	(0.0209)
Last employed 6–8 months ago	−5.65	−790.27	−0.0135	0.0312
	(1,824.51)	(5,453.91)	(0.0256)	(0.0296)

Table 9.3 (continued)

Subgroup	Earnings impacts ($)		Employment impacts	
	18 months	30 months	18 months	30 months
	Recent employment			
Last employed 9–11 months ago	1,191.58	−4,914.81	0.0163	0.0098
	(2,328.58)	(7,657.02)	(0.0384)	(0.0478)
Last employed ≥ 12 months ago	525.44	3,885.63	0.0284	0.0475
	(1,333.79)	(4,722.38)	(0.0224)	(0.0257)
Never employed	−799.52	−6,377.68	0.0017	0.0145
	(1,606.04)	(6,242.27)	(0.0295)	(0.0319)

NOTE: Monthly earnings are based on self-reports with top 1 percent trimming. Estimates are adjusted for program dropouts in the treatment group. Earnings impacts are calculated using all sample members with valid observations for self-reported monthly earnings during each period. The sample includes 4,886 valid observations for the 18-month period after random assignment and 1,147 valid observations for the 30-month period after random assignment. Heteroskedasticity-consistent standard errors appear in parentheses.

SOURCE: Heckman, Heinrich, and Smith (2002).

tus. In this case, the largest impacts appear for men employed at the time of random assignment. Employment at random assignment correlates positively with Y_a^1. As for adult women, the insignificant coefficients vary substantially among subgroups, and exhibit patterns that are difficult to interpret, such as nonmonotonicity as a function of months since last employment or years of schooling, as well as substantial changes from 18 to 30 months. Combined with the general lack of statistically significant subgroup impacts, the pattern of estimates represents weak evidence of at most a modest efficiency gain to cream skimming for adult males. For both men and women, of course, the costs of service provision may vary among subgroups as well, so that the net impacts may differ in either direction from the gross impacts reported here.

Other results in the literature that make use of the experimental data from the NJS echo the findings in Tables 9.2 and 9.3. Bloom et al. (1993, Exhibits 4.15 and 5.14) present subgroup impact estimates on earnings in the 18 months after random assignment. Orr et al. (1996, Exhibits 5.8 and 5.9) present similar estimates for 30-month earnings using a somewhat different earnings measure than we use here.[7] Both consider a different set of subgroups than we do. Only a couple of significant

subgroup impacts appear at 18 months. At 30 months, the only significant subgroup differences found by Orr et al. (1996) among adults are for adult men, where men with a spouse present have higher impacts.[8] Overall, the absence of many statistically significant subgroup differences, combined with the pattern of point estimates, makes the findings in Bloom et al. (1993) and Orr et al. (1996) consistent with our own findings. There exists little evidence of substantial efficiency gains or losses from picking participants on the basis of X and, even if such potential gains or losses exist, neither we nor, in all probability, the caseworkers, have any real knowledge of them.

RESULTS BASED ON SUBGROUP VARIATION IN EXPERIMENTAL IMPACTS: REGRESSION

We now turn to our multivariate regression analysis of the relationship between subgroup impacts and subgroup average performance measures. Table 9.4 presents estimates of the relationship between experimental impacts on earnings and on employment and various performance measures based on short-term labor market outcomes. We estimate separate regressions for each outcome (earnings and employment for 18 and 30 months) and for each performance measure.[9]

The four columns of estimates in Table 9.4 correspond to cumulated earnings and employment impacts for 18 and 30 months after random assignment. Each cell in the table presents the regression coefficient associated with the column's dependent variable and the row's independent variable, the estimated (robust) standard error of the coefficient, the p-value from a test of the null hypothesis that the population coefficient equals zero and the R^2 for the regression. We do not report the estimated constant terms from the regressions to reduce clutter. For example, the first row of the first column reveals that a regression of subgroup earnings impacts for the 18 months after random assignment on the subgroup average hourly wage at termination from the JTPA program yields an estimated coefficient of −$577.61 on the hourly wage, with a standard error of $304.00, a p-value of 0.0645, and an overall R^2 of 0.0809.

Four striking findings emerge from Table 9.4. First, and most important, we find many negative relationships between short-run performance indicators and experimental impact estimates at the subgroup level. In many cases the short-run outcome measures utilized in the JTPA performance standards system have a perverse relationship with the longer-run earnings and employment impacts that constitute the program's goals. The only evidence supporting the efficacy of short-run outcome measures comes from the employment-based performance measures for adult men, which are positive and statistically significant at the 10 percent level in three cases. These same performance measures have negative coefficients in seven out of eight cases for adult women.

Second, we find low R^2 values throughout. The short-term performance measures have only a very weak relationship with impacts on earnings and employment for 18 or 30 months. Third, moving from performance measures based on outcomes at termination from the program to longer-term measures based on outcomes three months after termination usually weakens the relationship between the performance measure and program impacts. In particular, the R^2 values nearly always decline and the estimated coefficients sometimes become less positive or more negative. Fourth, the performance measures often do worse (in terms of the fraction of variance explained) at predicting impacts for 30 months after random assignment than at predicting impacts for 18 months after random assignment. This indicates that the low predictive power of the performance measures in our analysis does not result from reductions in work activity during the periods of program participation (the so-called lock-in effect), which for some participants constitutes a nontrivial chunk of the 18 months after random assignment.

In sum, the regression analysis yields three clear conclusions. First, short-run performance measures do a very poor job of predicting long-run impacts, in terms of explained variation. In general, performance measures only weakly related to program goals accomplish little as rewards and punishments often get assigned based on noise. In terms of the discussion in Chapter 3, the JTPA performance measures do not solve the principal-agent problem by providing incentives for impact maximization. Moreover, they clearly fail to provide cheap, quick proxies for econometric impact evaluations. Second, the point estimates often suggest a negative relationship, indicating that the JTPA performance standards system may have provided an incentive for reduced

Table 9.4 Relationship between Δ and Y_1^1 in JTPA: Earnings and Employment Impacts

Performance standard measure	Earnings impact ($) measured over: 18 months after random assignment	Earnings impact ($) measured over: 30 months after random assignment	Employment impact measured over: 18 months after random assignment	Employment impact measured over: 30 months after random assignment
	Adult females			
Hourly wage at time of termination	−577.61	−1,729.66	−0.018	−0.010
	(304.00)	(1,280.64)	(0.008)	(0.011)
	$p = 0.0645$	$p = 0.1842$	$p = 0.0202$	$p = 0.3559$
	$R^2 = 0.0809$	$R^2 = 0.0426$	$R^2 = 0.1246$	$R^2 = 0.0208$
Weekly earnings at time of follow-up	−3.74	−12.05	−0.000	−0.000
	(8.78)	(36.54)	(0.000)	(0.000)
	$p = 0.6726$	$p = 0.7432$	$p = 0.2728$	$p = 0.3277$
	$R^2 = 0.0044$	$R^2 = 0.0026$	$R^2 = 0.0293$	$R^2 = 0.0234$
Employment at time of termination	−117.72	−2,065.61	−0.023	−0.029
	(941.92)	(3,928.63)	(0.023)	(0.033)
	$p = 0.9012$	$p = 0.6019$	$p = 0.3213$	$p = 0.3767$
	$R^2 = 0.0004$	$R^2 = 0.0069$	$R^2 = 0.0246$	$R^2 = 0.0196$
Employment at time of follow-up	1,513.28	−1,873.03	−0.067	−0.024
	(1,482.04)	(6,236.83)	(0.037)	(0.053)
	$p = 0.3132$	$p = 0.7655$	$p = 0.0767$	$p = 0.6521$
	$R^2 = 0.0248$	$R^2 = 0.0022$	$R^2 = 0.0745$	$R^2 = 0.0050$

	Adult males			
Hourly wage at time of termination	465.41	−1,405.68	0.003	−0.005
	(394.76)	(1,653.30)	(0.005)	(0.010)
	$p = 0.2452$	$p = 0.4001$	$p = 0.4914$	$p = 0.6230$
	$R^2 = 0.0328$	$R^2 = 0.0173$	$R^2 = 0.0116$	$R^2 = 0.0059$
Weekly earnings at time of follow-up	6.74	−20.76	0.000	−0.000
	(7.42)	(31.79)	(0.000)	(0.000)
	$p = 0.3690$	$p = 0.5174$	$p = 0.9921$	$p = 0.3274$
	$R^2 = 0.0197$	$R^2 = 0.0103$	$R^2 = 0.0000$	$R^2 = 0.0234$
Employment at time of termination	2,542.99	3,673.71	0.005	−0.059
	(1,384.72)	(5,869.08)	(0.017)	(0.034)
	$p = 0.0737$	$p = 0.5349$	$p = 0.7559$	$p = 0.0850$
	$R^2 = 0.0778$	$R^2 = 0.0097$	$R^2 = 0.0024$	$R^2 = 0.0723$
Employment at time of follow-up	2,579.24	18,716.00	0.050	0.021
	(2,486.91)	(9,842.28)	(0.028)	(0.061)
	$p = 0.3058$	$p = 0.0643$	$p = 0.0848$	$p = 0.7338$
	$R^2 = 0.0256$	$R^2 = 0.0810$	$R^2 = 0.0707$	$R^2 = 0.0029$

NOTE: The actual JTPA performance measures are defined as follows: "Hourly wage at placement" is the average wage at program termination for employed adults. "Weekly earnings at follow-up" are the average weekly wage of adults employed 13 weeks after program termination. "Employment rate at termination" is the fraction of adults employed at program termination. "Employment rate at follow-up" is the fraction of adults who were employed 13 weeks after program termination. In our analysis, employment rates were calculated based on the presence or absence of a job spell within 30 days before or after each reference date (termination or follow-up). Hourly wages were calculated based on the highest reported hourly wage for all job spells reported within 30 days before or after each reference date. Weekly earnings were calculated by averaging the product of hourly wages and hours worked per week across all reported job spells within 30 days before or after each reference date weighted by the fraction of the 61-day window spanned by each job spell.

SOURCE: Heckman, Heinrich, and Smith (2002).

efficiency. Third, we can say little about the efficiency cost to cream skimming other than that the data do not make a loud statement in either direction given our sample size and subgroups.

PUTTING THE RESULTS IN CONTEXT

The findings presented in this chapter do not represent an anomaly in the literature, but rather tell much the same story as the other studies that perform similar analyses. Table 9.5 summarizes six other studies that examine the relationship between performance standards measures based on short-run outcome levels and long-run program impacts; these six studies include, to the best of our knowledge, all of the published studies of this type as well as two that appeared only as government reports.[10] For each study, the table provides the citation, the particular employment and training program considered, the data used for the analysis, the impact measure used (for example, earnings from 18 to 36 months after random assignment), the impact estimator used (for example, random assignment), the particular performance measures considered (for example, employment at termination), and the findings.

Four studies, Gay and Borus (1980), Cragg (1997), Barnow (1999), and Burghardt and Schochet (2001), reach conclusions very similar to our own. The other two studies, Friedlander (1988) and Zornitsky et al. (1988), obtain more mixed results. The most positive of the studies, Zornitsky et al. (1988), examines the AFDC Homemaker/Home Health Aide Demonstration, which provided a homogeneous treatment to relatively homogeneous clients. This program represents a very different context from multitreatment programs serving heterogeneous populations such as JTPA and WIA. Moreover, this demonstration program, with its focus on the skills for a particular high-demand occupation, most likely did not lead to much postprogram human capital investment. As noted in Chapter 3, such investments tend to weaken the relationship between the short-run performance measures and long-run impacts. Taken together, these studies generally support our finding from the JTPA data that performance standards based on short-run outcome levels likely do little to encourage the provision of services to those who benefit most from them in employment and training programs.

LIMITATIONS OF OUR ANALYSIS

The analysis in this chapter focuses on one particular caseworker response to the imposition of a performance management system based on short-run outcomes: changes in who gets accepted into the program. In a model similar to the one presented in Chapter 3, caseworkers attempt to forecast both impacts and performance outcomes using the information available at the time of the acceptance decisions. In the case of the subgroup regression analysis, our interpretation assumes that caseworkers use observed characteristics to forecast both impacts and performance outcomes and then act on those forecasts. The performance management system causes them to put more weight onto the performance outcome forecast in making decisions about whom to serve.

Two important assumptions lurk in the shadows behind this interpretation. First, we must assume that mean impacts and mean performance at the subgroup level do not differ between a world with performance standards and a world without them. This assumption could easily fail if, for example, service allocations conditional on characteristics change with the introduction of performance management.[11] Second, we must also assume that mean impacts and mean performance at the subgroup level do not differ between applicants and participants.[12] Caseworkers see and make choices about applicants, while we have data only on individuals accepted into the program, as indicated by their reaching random assignment. The data from the Corpus Christi site in the NJS considered in Heckman, Smith, and Taber (1996) indicate that only about one-third of applicants reach random assignment, which leaves plenty of scope for differences between applicants and participants in the relationship that we estimate.

A final and very important limitation resides in the inability of our analyses in this chapter (or indeed, in this book) to say anything about the effect of the performance standards on the technical efficiency (or productivity) of the local JTPA training centers. By way of illustration, consider the subgroup regression analysis and suppose that having a performance standards system increases both the mental and physical effort levels (more "working smart" and less on-the-job leisure) of program staff. Suppose that this extra effort increases the impact of the

Table 9.5 Evidence on the Correlation Between Y_1 and Δ from Several Studies

Study	Program	Data	Measure of impact
Gay and Borus (1980)	Manpower Development and Training Act (MDTA), Job Opportunities in the Business Sector (JOBS), Neighborhood Youth Corps Out-of-School Program (NYC/OS), and the Job Corps	Randomly selected program participants entering programs from December 1968 to June 1970 and matched (on age, race, city, and sometimes neighborhood) comparison sample of eligible nonparticipants.	Impact on Social Security earnings in 1973 (from 18 to 36 months after program exit).
Zornitsky et al. (1988)	AFDC Homemaker-Home Health Aid Demonstration	Volunteers in the seven states in which the demonstration projects were conducted. To be eligible, volunteers had to have been on AFDC continuously for at least 90 days.	Mean monthly earnings in the 32 months after random assignment and mean monthly combined AFDC and food stamp benefits in the 29 months after random assignment.
Friedlander (1988)	Mandatory welfare-to-work programs in San Diego, Baltimore, Virginia, Arkansas, and Cook County	Applicants and recipients of AFDC (varies across programs). Data collected as part of MDRC's experimental evaluations of these programs.	Postrandom assignment earnings (from UI earnings records) and welfare receipt (from administrative data).

Impact estimator	Performance measures	Findings
Nonexperimental "kitchen sink" Tobit model	Employment in quarter after program, before-after (four quarters before to one quarter after) changes in weeks worked, weeks not in the labor force, wage rate, hours worked, income, amount of unemployment insurance received, and amount of public assistance received.	No measure has a consistent, positive, and statistically significant relationship to the estimated impacts across subgroups and programs. The before-after measures, particularly weeks worked and wages, do much better than employment in the quarter after the program.
Experimental impact estimates	Employment and wages at termination. Employment and welfare receipt three and six months after termination. Mean weekly earnings and welfare benefits in the three and six month periods after termination. These measures are examined both adjusted and not adjusted for observable factors including trainee demographics and welfare and employment histories and local labor markets.	All measures have the correct sign on their correlation with earnings impacts, whether adjusted or not. The employment and earnings measures are all statistically significant (or close to it). The welfare measures are correctly correlated with welfare impacts but the employment measures are not unless adjusted. The measures at three and six months do better than those at termination, but there is little gain from going from three to six.
Experimental impact estimates	Employment (nonzero quarterly earnings) in quarters 2 and 3 (short term) or quarters 4 to 6 (long term) after random assignment. Welfare receipt in quarter 3 (short-term) or quarter 6 (long-term) after random assignment.	Employment measure is positively correlated with earnings gains but not welfare savings for most programs. Welfare indicator is always positively correlated with earnings impacts, but rarely significantly so. It is not related to welfare savings. Long-term performance measures do little better (and sometimes worse) than short-term measures.

Table 9.5 (continued)

Study	Program	Data	Measure of impact
Cragg (1997)	JTPA (1983–87)	NLSY	Before-after change in participant earnings.
Barnow (1999)	JTPA (1987–89)	NJS	Earnings and hours worked in month 10 after random assignment.
Burghardt and Schochet (2001)	Job Corps	Experimental data from the National Job Corps Study	The outcome measures include receipt of education or training, weeks of education or training, hours per week of education or training, receipt of a high school diploma or GED, receipt of a vocational certificate, earnings, and being arrested. All are measured over the 48 months following random assignment.

Impact estimator	Performance measures	Findings
Generalized bivariate Tobit model of preprogram and postprogram annual earnings	Fraction of time spent working since leaving school in the preprogram period. This variable is strongly correlated with postprogram employment levels.	Negative relationship between work experience and before-after earnings changes.
Experimental impact estimates	Regression-adjusted levels of earnings and hours worked in month 10 after random assignment.	At best a weak relationship between performance measures and program impacts.
Experimental impact estimates	Job Corps centers divided into three groups: high performers, medium performers, and low performers based on their overall performance rankings in program years 1994, 1995, and 1996. High and low centers were in the top and bottom third nationally in all three years, respectively.	No systematic relationship between the performance groups and the experimental impact estimates.

SOURCE: Heckman, Heinrich, and Smith (2002); Barnow and Smith (2004).

program for all participants by $100 over 18 months. In our regressions of estimated subgroup mean impacts on estimated subgroup mean performance levels, this extra $100 shows up in the intercept, not in the slope coefficient, with the result that we do not interpret it as the effect of the performance standards. To our knowledge, the only evidence of the effect of performance management on technical efficiency in the context of an active labor market program comes from the United Kingdom, where Burgess et al. (2004) find evidence of such effects for small work teams but not for large ones. This lack of evidence comes as a real surprise, given that the literature on performance incentives in private firms, well summarized in Prendergast (1999), focuses almost exclusively on productivity effects.

SUMMARY AND CONCLUSIONS

This chapter presents several empirical analyses designed to address the questions laid out in the introduction: First, do short-run performance measures predict long-run impacts? Second, what are the efficiency costs of cream skimming? We describe the identifying assumptions underlying our analyses as well as their limitations.

Taken as a whole, our empirical analysis reaches two important conclusions. First, the limited evidence we have suggests that whatever cream skimming occurs in JTPA produces only modest efficiency gains or losses. In other words, though we must acknowledge the noisiness of the evidence, our results suggest at most a modest efficiency cost associated with eschewing cream skimming in favor of a focus on the most hard-to-serve among those applying to the program. Second, the JTPA performance standards do not promote efficiency because the short-run outcomes they rely on have essentially a zero correlation with long-run impacts on employment and earnings. This surprising result comports with the findings in several other studies that have estimated this relationship.

Notes

1. This chapter presents results from Heckman, Heinrich, and Smith (2002) and borrows in places from their text.
2. In addition to simple mean differences, we can also use regression analysis to obtain experimental impact estimates. Doing so may generate more precise estimates if the exogenous conditioning variables included in the regression soak up a lot of the residual variance.
3. This concept has a variety of names in the published literature, including "perfect positive dependence" and "perfect positive rank correlation." We use "rank preservation" here because it is short and seems to be gaining ground in the most recent literature.
4. See, e.g., Heckman, Smith, and Clements (1997); Bitler, Gelbach, and Hoynes (2008); and Djebbari and Smith (2008) for more details on this estimator, including the construction of the standard errors.
5. This correction amounts to using the simple Wald instrumental variables estimator with treatment status as an instrument for enrollment. See, e.g., the discussions in Heckman, Smith, and Taber (1998) and Heckman, LaLonde, and Smith (1999, Section 5.2) on the properties and origin of this estimator.
6. An alternative strategy would generate predicted individual impacts by including interaction terms between baseline covariates and the treatment group dummy in an impact regression; see Barnow (1999) for an application.
7. Their earnings measure combines self-report data with data from Unemployment Insurance earnings records. For more details, see the discussion in Orr et al. (1996).
8. Orr et al. (1996, Exhibits 5.19 and 5.20) also present subgroup impact estimates for male and female youth. As expected given the small sample sizes, they find no statistically significant differences in estimated impacts among the subgroups.
9. To improve statistical efficiency, we use the inverse of the robust standard errors from the corresponding impact estimation as weights in each regression. Recall that the dependent variable here is the impact; its estimated standard error is thus an estimate of the variance of the error term for that impact, which represents one observation in our regression. Viewed in this way, the procedure amounts to doing weighted least squares in the presence of heteroskedasticity, where the extent of the heteroskedasticity is indicated by differences among subgroups in the estimated standard errors of the impacts.
10. We thank Tim Bartik of the W.E. Upjohn Institute for Employment Research for providing us with copies of two of the unpublished papers.
11. To see this, consider a simple case with two groups, A and B, two services, Classroom Training (CT) and Job Search Assistance (JSA), and one short-run performance measure, P. For group A, CT yields impact 100 and performance 20 while JSA yields impact 40 and performance 40. In contrast, for group B, CT yields impact 30 and performance 10 while JSA yields impact 40 and performance 40. Suppose further that without the performance management system, program staff

would maximize impacts by assigning group A to CT and group B to JSA. In contrast, with the performance management system, they maximize performance by assigning both A and B to JSA. Thus, the introduction of the performance standards system induces a substantial efficiency loss. Unfortunately, our regression analysis applied to data collected from this imaginary program after the introduction of performance standards would not reveal the efficiency loss. This follows from the fact that in the world with the performance standards system, the correlation between subgroup mean impacts and subgroup performance equals zero.

12. To see the issue, consider a simple example. In this example our program has just one service: CT. Among the applicants, some individuals have a (H)igh impact of CT because they get along well with the instructor, others would have a (L)ow impact because they do not. At the same time, applicants also differ in their job search behavior following CT. Some individuals, call them (F)ast, take the first job they find after completing CT while other individuals, call them (S)low, search longer but find a higher paying job in the end, as standard search theory would predict. Together H/L and F/S define four groups. Assume that these four groups each constitute one-quarter of the applicants and that the program has sufficient resources to serve half of the applicants. To make the example concrete, we assign the following values: H-F individuals have impact 100 and performance 50, H-S individuals have impact 120 and performance 20, L-F individuals have impact 50 and performance 50, and L-S applicants have impact 80 and performance 10. In a world without performance standards, caseworkers serve only H individuals, while in a world with performance standards, caseworkers serve only F individuals. In the applicant population, impact and performance outcomes have a negative correlation, indicating an efficiency loss from selection into the program based on performance rather than impacts. In the participant population, impact and performance have a zero correlation because, by construction, performance equals 50 for all the participants regardless of their impact. This example clearly violates the assumption of the same relationship between impacts and performance among participants and applicants. It also demonstrates that failure of this assumption can lead to a misleading conclusion about the efficiency effects of cream skimming and about whether short-term outcomes predict long-term impacts.

References

Barnow, Burt. 1999. "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): 118–141.

Barnow, Burt, and Jeffrey Smith. 2004. "Performance Management of U.S. Job Training Programs." In *Job Training Policy in the United States*, Christopher O'Leary, Robert Straits, and Stephen Wandner, eds. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, pp. 21–56.

Bell, Stephen, and Larry Orr. 2002. "Screening (and Creaming?) Applicants to Job Training Programs: The AFDC Homemaker–Home Health Aide Demonstrations." *Labour Economics* 9(2): 279–301.

Bitler, Marianne, Jonah Gelbach, and Hilary Hoynes. 2008. "Distributional Impacts of the Self-Sufficiency Project." *Journal of Public Economics* 92(3–4): 748–765.

Bloom, Howard, Larry Orr, Stephen Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes Bos. 1997. "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study." *Journal of Human Resources* 32(3): 549–576.

Bloom, Howard, Larry Orr, George Cave, Stephen Bell, and Fred Doolittle. 1993. *The National JTPA Study: Title IIA Impacts on Earnings and Employment at 18 Months*. Bethesda, MD: Abt Associates.

Burgess, Simon, Carol Propper, Marisa Ratto, and Emma Tominey. 2004. "Incentives in the Public Sector: Evidence from a Government Agency." CMPO Working Paper No. 04/103. Bristol, UK: Centre for Market and Public Organisation, Bristol Institute of Public Affairs, University of Bristol.

Burghardt, John, and Peter Schochet. 2001. *National Job Corps Study: Impacts by Center Characteristics*. Princeton, NJ: Mathematica Policy Research.

Cragg, Michael. 1997. "Performance Incentives in the Public Sector: Evidence from the Job Training Partnership Act." *Journal of Law, Economics, and Organization* 13(1): 147–168.

Djebbari, Habiba, and Jeffrey Smith. 2008. "Heterogeneous Program Impacts: Experimental Evidence from the PROGRESA Program." *Journal of Econometrics* 145(1–2): 64–80.

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

Friedlander, Daniel. 1988. *Subgroup Impacts and Performance Indicators for Selected Welfare Employment Programs*. New York: Manpower Demonstration Research Corporation.

Gay, Robert, and Michael Borus. 1980. "Validating Performance Indicators for Employment and Training Programs." *Journal of Human Resources* 15(1): 29–48.

Heckman, James. 1992. "Randomization and Social Program Evaluation." In *Evaluating Welfare and Training Programs*, Charles Manski and Irwin Garfinkel, eds. Cambridge, MA: Harvard University Press, pp. 201–230.

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

Heckman, James, Neil Hohmann, Jeffrey Smith, and Michael Khoo. 2000. "Substitution and Dropout Bias in Social Experiments: A Study of an Influ-

ential Social Experiment." *Quarterly Journal of Economics* 105(2): 651–694.

Heckman, James, Robert LaLonde, and Jeffrey Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics,* Vol. 3A, Orley Ashenfelter and David Card, eds. Amsterdam: North-Holland, pp. 1865–2097.

Heckman, James, and Jeffrey Smith. 1998. "Evaluating the Welfare State." In *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial*, Steinar Strøm, ed. Econometric Society Monograph No. 31. Cambridge: Cambridge University Press, pp. 241–318.

———. 2000. "The Sensitivity of Experimental Impact Estimates: Evidence from the National JTPA Study." In *Youth Employment and Joblessness in Advanced Countries*, David Blanchflower and Richard Freeman, eds. NBER Comparative Labor Markets Series. Chicago: University of Chicago Press, pp. 331–356.

Heckman, James, Jeffrey Smith, and Nancy Clements. 1997. "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies* 64(4): 487–535.

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

———. 1998. "Accounting for Dropouts in Evaluations of Social Programs." *Review of Economics and Statistics* 80(1): 1–14.

Hotz, V. Joseph. 1992. "Designing an Evaluation of the Job Training Partnership Act." In *Evaluating Welfare and Training Programs*, Charles Manski and Irwin Garfinkel, eds. Cambridge, MA: Harvard University Press, pp. 76–114.

Kemple, James, Fred Doolittle, and John Wallace. 1993. *The National JTPA Study: Site Characteristics and Participation Patterns*. New York: Manpower Demonstration Research Corporation.

Kornfeld, Robert, and Howard Bloom. 1999. "Measuring Program Impacts on Earnings and Employment: Do Unemployment Insurance Wage Reports from Employers Agree with Surveys of Individuals." *Journal of Labor Economics* 17(1): 168–197.

Lechner, Michael, and Jeffrey Smith. 2007. "What Is the Value Added by Case Workers?" *Labour Economics* 14(2): 135–151.

Orr, Larry, Howard Bloom, Stephen Bell, Fred Doolittle, Winston Lin, and George Cave. 1996. *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study*. Washington, DC: Urban Institute Press.

Orr, Larry, Howard Bloom, Stephen 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.

Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature* 37(1): 7–63.

Smith, Jeffrey. 1997. "Measuring Earnings Levels among the Poor: Evidence from Two Samples of JTPA Eligibles." Working paper. London, Ontario, Canada: University of Western Ontario.

Zornitsky, Jeffrey, and Mary Rubin. 1988. *Establishing a Performance Management System for Targeted Welfare Programs*. Report No. 88-14. Washington, DC: National Commission for Employment Policy.