The Sensitivity of Experimental Impact Estimates
Evidence from the National JTPA Study

James J. Heckman and Jeffrey A. Smith

8.1 Introduction

The experimental estimates of the impact of youth training funded under the Job Training Partnership Act (JTPA) from the recent National JTPA Study (NJS) resulted in large budget cuts in the JTPA program. The experiment, which included only 16 of the more than 600 JTPA training centers, found negative and statistically significant impacts on the earnings of male youths in the 18 months after random assignment and negligible impacts on the earnings of female youths. In response to these estimates, Congress cut funding for the youth component of JTPA from $540 million in 1994 to only $110 million in 1995, a cut of over 80 percent.

In light of the dramatic changes in JTPA resulting from the NJS impact estimates, it is of interest to consider their sensitivity to issues of construction and interpretation. In this paper, we address the following questions: (1) How sensitive are the estimates to the set of training centers included in the evaluation? (2) Does it matter how the impact estimates from the individual training centers in the evaluation are combined? (3) How sensi-

James J. Heckman is the Henry Schultz Distinguished Service Professor of Economics and director of the Center for Social Program Evaluation at the University of Chicago, a senior fellow of the American Bar Foundation, and a research associate of the National Bureau of Economic Research. Jeffrey A. Smith is associate professor of economics at the University of Western Ontario and a faculty research fellow of the National Bureau of Economic Research.

This research was supported by National Science Foundation grants SBR 91-11-445 and SBR 93-21-048 and by grants from the Russell Sage Foundation and the Social Science and Humanities Research Council of Canada. The authors thank David Blanchflower, Richard Freeman, Lawrence Katz, Gerd Ronning, and NBER workshop participants for helpful comments and Robert LaLonde for useful suggestions and a very close reading of a draft of this paper.

tive are the estimates to the treatment of outliers in the earnings data? (4) How sensitive are the estimates to the construction of the earnings measure used in the evaluation? (5) How sensitive are the estimates to the manner in which dropouts from the experimental treatment group are handled? (6) How sensitive are the conclusions of the evaluation to the manner in which substitution by control group members into alternative sources of training similar to that provided by JTPA is dealt with?

We find the following: (1) The dispersion in impacts across centers is large enough that choosing a different set of centers could have produced a fundamentally different pattern of impact estimates. (2) Combining the centers in the NJS in a manner that takes account of the fact that some centers dropped out of the experiment early leads to negative and statistically significant impact estimates for female youth. (3) The magnitude and statistical significance of the male youth estimates depend on how outliers in the earnings data are handled. (4) The different methods used to construct the earnings variables in the two official NJS impact reports lead estimated impacts on important subgroups to change by up to $1,000 and to switch signs. (5) Taking account of the 40 percent of experimental treatment group members who drop out substantially increases the magnitude of the impact estimates. (6) Substitution by control group members in the NJS is empirically important, and taking account of it in the construction and interpretation of the estimates requires recourse to nonexperimental evaluation methods. Estimates of the impact of JTPA classroom training that account for both treatment group dropout and control group substitution present a substantially more positive picture of the effects of JTPA youth training than the unadjusted experimental impact estimates.

Our work has an important methodological motivation. No social program has ever been the subject of multiple experimental evaluations. It is well known in the literature that factors such as those we consider in this paper can have a major influence on the estimates obtained from nonexperimental evaluations, even holding constant the data sources and econometric methods employed. The prime example of such sensitivity is the multiple evaluations of JTPA's predecessor, the Comprehensive Employment and Training Act (CETA). Dickinson, Johnson, and West (1987) show that the widely divergent estimates in these evaluations resulted in large part from seemingly minor choices in the construction of the estimates.

While analysts using data from a social experiment do not have to choose a nonexperimental evaluation method, they must still make many choices regarding how to construct, report, and interpret their estimates. Some of these choices, such as the selection of locations at which to evaluate the program, are more problematic in experiments than in nonexperimental analyses. Others, such as what to do about control group members who obtain close substitutes for the experimental treatment, are unique to experiments. The claim that experiments are superior to nonexperimental methods because they produce "one number" is false. Correcting this mistaken view by showing the sensitivity of experimental estimates to the numerous choices that must be made to produce them is one of the primary goals of this paper.

The strategy we adopt makes use of the data at hand. An important reason why multiple experimental evaluations have never been conducted for the same program is that experimental evaluations are quite expensive. For example, the NJS cost around $30 million. Thus, rather than answering the sensitivity question directly by conducting and reporting on the results from multiple experimental evaluations, we examine the sensitivity of estimates constructed from a single experimental data set to alternative choices regarding construction and interpretation.[1]

The remainder of this paper proceeds as follows. Sections 8.2 and 8.3 describe the institutional structure of the JTPA program and characterize the data from the NJS, respectively. Section 8.4 examines the effects of variation in the set of JTPA training centers included in the analysis. Section 8.5 examines how the method used to combine the data from the individual training centers affects the impact estimates. Section 8.6 considers the effects of alternative methods for handling outliers in the earnings data, and section 8.7 shows how the construction of the earnings measure affects the impact estimates. In section 8.8 we consider ways to adjust and reinterpret the impact estimates to take account of treatment group members who drop out and control group members who obtain close substitutes to the experimental treatment. In section 8.9 we summarize our findings and discuss their implications for the interpretation of the NJS estimates, for our understanding of the effectiveness of training programs for disadvantaged youth, and for future evaluations of employment and training programs.

8.2 The JTPA Program

The JTPA program was, until recently, one of the largest federal training programs in the United States. With an annual budget of around $1 billion, JTPA provided employment and training services to several hundred thousand economically disadvantaged persons each year. The JTPA program was highly decentralized, with more than 600 JTPA training centers across the United States. While JTPA was a major provider of training in most areas, it was usually not the only provider of subsidized training to the disadvantaged. The federal government provided the funding and set

[1] There is some evidence of this type from earlier experiments. An important example is the long-standing debate on the effect of a negative income tax on marital stability. Both Cain and Wissoker (1990) and Hannan and Tuma (1990) use data from the U.S. negative income tax experiments, but they come to dramatically different conclusions.

the broad outlines of the JTPA program but left local training centers a substantial amount of flexibility in determining whom to serve and how to serve them.

Devine and Heckman (1996) show that the JTPA-eligible population included nearly everyone below the poverty line and many persons above it. Because its budget allowed JTPA to serve only about 3 percent of those eligible for it each year, program operators had wide latitude in choosing whom to serve. Moreover, even if program operators had picked at random from among the eligible (and they did not), they would have ended up with widely different participant populations due to the heterogeneity across training centers in the characteristics of the eligible population.[2]

Local operators also had control over what services to offer to JTPA participants. The most common services provided by JTPA were classroom training in occupational skills, subsidized on-the-job training at private firms, and job search assistance. Less common were basic education (typically GED preparation) and work experience. The relative proportions of trainees receiving each type of training, as well as the form, content, and duration of training within each type, varied widely across centers (Kemple, Doolittle, and Wallace 1993).

8.3 The National JTPA Study

Our data come from the National JTPA Study, a recent experimental evaluation of the JTPA program commissioned by the U.S. Department of Labor.[3] Due to the high fixed costs of setting up random assignment at a given training center, the NJS includes only 16 centers. The original design called for a random sample of training centers, but these plans had to be abandoned when most of the centers initially contacted refused to participate. In the end, it was necessary to approach over 200 training centers in order to find 16 willing to take part in the experiment (Doolittle and Traeger 1990). In addition, training centers in large urban areas and training centers serving fewer than 500 persons per year were excluded for cost and sample size reasons, respectively. Random assignment took place between 1987 and 1989, with the exact dates varying across training centers. A total of 20,601 persons were randomly assigned, of whom 2,558 were male youths and 3,132 were female youths.

8.4 Selection of Training Centers

In this section we examine the sensitivity of the overall experimental impact estimates to the set of centers included in the evaluation. The selec-

[2] Smith (1997a) documents this heterogeneity for four centers in the NJS at which data on the eligible population were collected.

[3] Doolittle and Traeger (1990) and Bloom et al. (1993) describe the NJS in detail.

tion of training centers is more problematic in experimental than nonexperimental evaluations. The high fixed costs of setting up random assignment limit the number of centers. In addition, centers often refuse to participate because of the political and budgetary costs associated with random assignment[4] or because of the increased recruitment necessary to fill the experimental control group.

We present experimental impact estimates for each center in the NJS. In order for the set of centers included in the evaluation to affect the overall impact estimates, it must be the case that the impact differs across centers. While the point estimates do vary widely, formal statistical tests do not reject the null hypothesis of equal impacts across centers. In light of this, we perform a simulation analysis that shows the effect of the variability we do observe in center-level impacts on the overall impact estimate. The simulation provides strong evidence of the sensitivity of the overall impact estimates for youths to the set of included centers. In addition, the statistical significance of the overall estimate for male youths is very sensitive to center selection.

We also consider whether the variation across centers in the estimated impact of JTPA can be traced to specific factors operating at the center level, such as the center's administrative structure or the local labor market. We find little evidence for the importance of center-level factors.

8.4.1 Impacts by Training Center in the National JTPA Study

Table 8.1 presents experimental estimates of the mean impact of JTPA training on earnings in the first 18 months after random assignment for male and female youths in the NJS.[5] Two important patterns emerge from these estimates. First, the point estimates differ substantially across training centers within each demographic group. For example, for male youths, the center-specific impact estimates range from a low of −$6,554.68 at center 2 to a high of $4,432.61 at center 8. Second, only one estimate in table 8.1 is statistically distinguishable from zero at the 5 percent level, though some of the extreme positive and negative estimates are statistically distinguishable from one another.

Statistical tests of the equality of impacts across training centers do not reject that hypothesis at conventional levels. Two *F*-tests were carried out for each group, one with covariates included in the regression used to estimate the impacts and one without. We use the same regression specification as in the official report of Bloom et al. (1993). For male youths, the

[4] The U.S. Department of Labor spent nearly $1 million on payments to centers to cover the budgetary costs of participating in the NJS. Doolittle and Traeger (1990) note that ethical and public relations difficulties with random assignment and the denial of services to control group members were the concerns cited most often by centers declining to participate in the NJS.

[5] The centers are not identified by name due to an agreement between the centers in the NJS and the U.S. Department of Labor.

Table 8.1 **Experimental Estimates of Impact on Self-Reported Earnings in First 18 Months after Random Assignment**

Training Center[a]	Male Youths	Female Youths
Center 1	−2,364.73	−440.97
	(1,304.04)	(701.85)
Center 2	−6,554.68	−929.94
	(3,048.81)	(1,772.11)
Center 3	531.76	−106.22
	(1,335.79)	(681.21)
Center 4	−153.10	−806.40
	(2,385.74)	(1,514.22)
Center 5	−644.58	−626.51
	(1,084.94)	(796.94)
Center 6	−1,645.31	−1,418.03
	(1,607.93)	(893.77)
Center 7	1,501.57	−460.12
	(1,280.66)	(784.37)
Center 8	4,432.61	1,491.85
	(3,037.43)	(1,542.14)
Center 9	−1,278.52	333.92
	(3,266.78)	(2,491.02)
Center 10	−2,611.03	789.43
	(2,599.02)	(2,069.87)
Center 11	−1,570.64	−2,489.35
	(2,064.25)	(1,665.46)
Center 12	−1,958.44	−377.75
	(2,252.27)	(1,065.39)
Center 13	318.39	775.10
	(2,044.43)	(1,323.22)
Center 14	−1,150.37	1,090.70
	(1,208.29)	(860.41)
Center 15	−2,265.58	985.84
	(1,525.25)	(945.40)

Source: National JTPA Study 18 Month Impact Sample.

Note: The self-reported earnings variable used here includes the Bloom et al. (1993) hand imputations for outliers. These impact estimates are regression-adjusted using the same specification as in Bloom et al. (1993); results differ slightly from those in Bloom et al. because we were unable to exactly replicate their construction of some of the covariates. Numbers in parentheses are estimated standard errors.

[a] Only 15 training centers are listed because youth were not randomly assigned at one of the 16 centers in the NJS.

p-values are 0.3945 and 0.3940 with and without covariates, respectively. For female youths, they are 0.7284 and 0.3162, respectively.

In thinking about the lack of statistically significant estimates at the individual centers, and the failure to reject the null of equal impacts, it is important to note that the available sample sizes are rather small, particularly given the large variance in earnings (even conditional on covariates) in the JTPA participant population. The average center sample size is 117 for male youths and 153 for female youths.

One good reason to think that small sample sizes, and not actual equality of impacts across centers, underlie the lack of statistically significant findings is that such findings appear in other evaluations of employment and training programs for the disadvantaged with larger sample sizes. For example, the experimental evaluation of California's Greater Avenues to Independence (GAIN) program reported in Riccio, Friedlander, and Freedman (1994) reveals earnings impacts (over the 36 months after random assignment) for AFDC single family heads ranging from $260 in Los Angeles County to $3,113 in Riverside County. With an average sample size of over 3,000 per county, the larger estimates in the GAIN study are statistically distinguishable from zero and from the smaller estimates. Similar differences across centers are found in the experimental evaluation of the National Supported Work (NSW) program described in Hollister, Kemper, and Maynard (1984), where again the sample sizes per center are larger than for youths in the NJS.

8.4.2 Effect of Center Selection on Variability of Overall Impact Estimates

This subsection examines the effect of variation in center-level impacts from the NJS on the overall impact estimates. We conduct a simulation in which we calculate overall impact estimates based on random samples of 15 centers drawn, with replacement, from the NJS data. The data from the centers in each random sample are combined to produce overall impact estimates for male and female youths. In formal terms, we treat the estimated impacts from the NJS training centers as providing a nonparametric estimate of the distribution of center-level impacts for the population of JTPA training centers. Because we use the nonrandom sample of JTPA training centers participating in the NJS, our results likely understate the variability in overall impacts that would be obtained from repeated random sampling from the *population* of JTPA training centers.

Table 8.2 reports characteristics of the distribution of overall impacts obtained when 100,000 samples of 15 centers are randomly drawn from the observed distribution. The top panel reports percentiles of the distribution of overall impacts from the 100,000 samples, as well as the mean and standard deviation of the overall impact estimates. The figures reveal remarkable variability in the overall impacts obtained from random samples of 15 centers. For female youths, the interquartile range is around $380. For male youths, it is over $600. Looking in the tails, the variation in estimates is particularly large for female youths, for whom the 5th percentile estimate is −$647.80 while the 95th percentile estimate is $312.45. This variability is large relative to the overall experimental impact estimates reported in Bloom et al. (1993).

Table 8.2 **Sensitivity of Experimental Impact Estimates to Set of Training Centers Included**

	Male Youths	Female Youths
Parameters of the Distribution of Overall Impact Estimates from 100,000 Random Samples of 15 Training Centers		
1st Percentile	−1,969.51	−885.33
5th Percentile	−1,663.78	−647.80
25th Pecentile	−1,227.08	−341.91
Median	−920.00	−146.10
75th Percentile	−613.19	42.32
95th Percentile	−158.22	312.45
99th Percentile	167.91	503.27
Mean	−917.81	−154.36
Standard deviation	457.96	292.52
Characteristics of the Distribution of Overall Impact Estimates from 100,000 Random Samples of 15 Training Centers		
Fraction negative and significant at 1%	.2846	.0142
Fraction negative and significant at 5%	.5265	.0626
Fraction negative and significant at 10%	.6490	.1149
Fraction negative	.9767	.6990
Fraction positive and significant at 1%	.0000	.0006
Fraction positive and significant at 5%	.0001	.0050
Fraction positive and significant at 10%	.0003	.0127
Fraction positive	.0233	.3010

Source: National JTPA Study 18 Month Impact Sample.

Note: Each set of 15 training centers is drawn at random from the NJS data, with replacement. The self-reported earnings variable used here includes the Bloom et al. (1993) hand imputations for outliers. This analysis uses simple mean-difference experimental impact estimates.

The bottom panel of table 8.2 summarizes the sign and statistical significance of the overall impact estimates obtained from the 100,000 random samples of centers. The overall estimates are essentially always negative for male youths. For female youths, they are negative 70 percent of the time and positive 30 percent of the time. Varying the set of included training centers has strong effects on the statistical significance of the overall estimates. In almost half the samples, the negative overall impact estimate for male youths is not statistically significant at the 5 percent level. Given the common practice of treating statistically insignificant estimates as zero, these findings are very important.

8.4.3 Do Center-Level Factors Account for Heterogeneous Impacts?

Linking the differing impacts across centers to specific factors associated with each center, such as their approach to treatment or their local economic conditions, serves two purposes. First, it may allow the problem of "external validity" that results from allowing centers to self-select into the evaluation to be overcome. Provided the support of the distribution of center characteristics affecting program impacts among centers in the evaluation spans the support in the population, the relevant characteristics can be conditioned on in the evaluation and then used in combination with the distribution of factors across centers for the program as a whole in generating estimated impacts for the population of centers. Second, such links have obvious policy relevance, particularly for factors controlled by center staff, such as the approach to treatment.

We investigate this question indirectly using the JTPA data. If center-level factors drive the differences in impact estimates, then the impact estimates across demographic groups should be correlated. That is, if centers that have strong local economies, or are run by private rather than public agencies, have higher (or lower) impacts, this should hold across demographic groups because these center characteristics are fixed for a given center. Thus positive correlations between the center-specific impacts for pairs of demographic groups provide evidence of the importance of center-level characteristics.

Table 8.3 displays estimated Pearson product-moment correlations and Spearman rank correlations between the center-level impact estimates for pairs of demographic groups in the NJS. In this case, we include all four NJS demographic groups because doing so increases the available evidence from one correlation to six. The table also displays *p*-values from tests of the null hypothesis that the true correlation is zero, along with the number of estimates used in calculating the correlation.

None of the estimated correlations in table 8.3 is statistically distinguishable from zero at the 5 percent level. All but one of the point estimates is below .3 in absolute value, and all but two are below .2. A few of the point estimates are negative. Overall, the table provides little evidence that center-level factors are important determinants of the impact of JTPA.

Another possible source of heterogeneous impacts at the center level is that certain centers perform well or poorly at providing certain treatment types. In the NJS, it is possible to produce experimental impact estimates that condition, not on the services actually received, which are determined after random assignment, but on the services for which potential participants are recommended by JTPA staff prior to random assignment. The three treatment streams based on recommended services are the classroom training in occupational skills (CT-OS) stream, the on-the-job training (OJT) stream, and the "other services" stream. We calculated the Pearson product-moment and Spearman rank correlations between center-level impact estimates within each treatment stream for each demographic

Table 8.3 Correlations of Experimental Impact Estimates across Training Centers

Demographic Group	Adult Males	Adult Females	Male Youths
Pearson Product-Moment Correlation			
Adult females	.1906 (.4704) [16]		
Male youths	.0835 (.7675) [15]	.0582 (.8368) [15]	
Female youths	−.3528 (.2160) [14]	.2473 (.3939) [14]	.5001 (.0686) [14]
Spearman Rank Correlation			
Adult females	.0235 (.9311) [16]		
Male youths	.1821 (.5159) [15]	.1321 (.6387) [15]	
Female youths	−.4066 (.1491) [14]	.0198 (.9465) [14]	.2835 (.3260) [14]

Source: National JTPA Study 18 Month Impact Sample.

Note: The self-reported earnings variable used here includes the Bloom et al. (1993) hand imputations for outliers along with the imputed values generated by Bloom et al. for adult female nonrespondents using information from state unemployment insurance earnings records. Training centers with fewer than 30 experimental sample members are excluded. The correlations are calculated using simple mean-difference experimental impact estimates. Numbers in parentheses are p-values from tests of the null hypothesis that the true correlation is zero. Numbers in brackets are numbers of impact estimates used to construct correlations. The estimated standard errors do not account for the fact that the impacts being correlated are themselves estimates. Doing so would make the estimated standard errors larger and therefore reinforce the conclusions drawn in the text.

group. The results of this analysis match those reported in table 8.3. If anything, the evidence for the treatment streams is even weaker, as the estimated correlations are more often negative than in table 8.3.

8.5 Pooling

In the preceding section, we considered the sensitivity of the experimental impact estimates to the set of training centers included in the evaluation under the assumption that the best way to combine the data across centers was to pool it into a single large sample of individuals. In this

Table 8.4 **Sensitivity of Experimental Impact Estimates to Method of Pooling Training Centers**

Weighting Variable	Male Youths	Female Youths
Center sample size	−893.05	−191.52
	(466.93)	(293.11)
	[.0548]	[.5156]
Inverse variance of estimated impact	−506.30	−78.61
	(429.26)	(276.95)
	[.2380]	[.7794]
Number of program year 1989 terminees	−660.89	−609.05
	(553.77)	(341.41)
	[.2340]	[.0750]

Source: National JTPA Study 18 Month Impact Sample.

Note: The self-reported earnings variable used here includes the Bloom et al. (1993) hand imputations for outliers. Simple mean-difference experimental impact estimates are used for each training center in computing the overall impacts. The overall estimates obtained when the training center estimates are weighted by the training center sample sizes differ slightly from the overall mean difference estimates obtained using the full sample of individuals because the ratio of control to treatment group members differs slightly across training centers. At some centers, random assignment ratios higher than 2:1 were used for short periods. Numbers in parentheses are estimated standard errors. Numbers in brackets are *p*-values from tests of the null hypothesis that the true impact is zero.

section, we examine the sensitivity of the experimental estimates to two alternative pooling methods.[6]

The desirability of using an alternative pooling method depends on how the impact of JTPA and the variance of the outcome variable, earnings, vary across training centers. If both the impact and the outcome variance are the same across centers, then there is no gain from doing anything other than combining the data from each training center into a single large sample of individuals. Doing so is equivalent to weighting the center-level impact estimates by the center sample sizes. Impact estimates produced in this way appear in the first row of table 8.4.

If the variance of earnings itself varies across centers, while the impact is constant or varies independently of the variance in earnings, then the efficiency of the estimates can be increased without adding any bias by calculating the overall impact as a weighted average of the impact estimates for the individual centers, with the weights inversely proportional to the variance of the impact estimate at each center. Impact estimates

[6] When, as in the NJS, the sample of centers is not randomly selected from the population of centers, the justification for combining the centers in any way to produce an overall impact estimate is unclear. Whatever estimate is obtained from doing so is not externally valid, which means that it is not a valid estimate of the impact of training at centers other than those included in the evaluation.

obtained in this way appear in the second row of table 8.4. For male youths, weighting cuts the magnitude of the impact estimate almost in half, indicating that point estimates for this group are lower (more negative) at centers where the variance of the impact estimate is relatively large. This effect is less pronounced for female youths.

Another type of weighting is useful if the representation of each center in the experimental sample differs from its representation in the overall JTPA participant population. To see why, suppose that large training centers are underrepresented in the experimental sample and that the impact of the program is bigger in large training centers due to economies of scale. In this case, simply combining the samples from the individual training centers results in an overall impact estimate that is biased downward relative to the true impact of JTPA on a randomly selected participant. Note that underrepresentation of particular centers in the experimental sample is not an issue if the impact of JTPA is the same at every training center, or if whether or not a training center is underrepresented is independent of its impact.

In the NJS data, the participating centers are not represented in proportion to the number of participants they serve because several centers dropped out of the experiment early. The third row of table 8.4 presents impact estimates constructed by weighting the center-specific impacts with weights proportional to the number of JTPA terminees at each center in program year 1989, where program year 1989 is selected because it overlaps with the period of random assignment at most of the centers.[7] This weighting has a large effect on the impact estimate for female youths, which becomes nearly as large in absolute value as that for male youths and statistically significant at the 10 percent level.

8.6 Treatment of Earnings Outliers

Unusually large earnings observations, or outliers, can have important effects on experimental impact estimates based on conditional means. Outliers may represent invalid values, or they may represent valid values with a very low probability of being observed. In either case, it may be desirable to adopt a systematic procedure to minimize their influence on the impact estimates.

Table 8.5 shows the sensitivity of the experimental impact estimates for youths in the NJS data to alternative methods of handling earnings outliers. The first row of the table presents impact estimates constructed using the raw earnings data. The second row presents the estimates from Bloom et al. (1993), in which the top 2 percent of the earnings values for each group were examined by hand for coding errors or inconsistencies and

[7] Program year 1989 runs from July 1989 to June 1990.

Table 8.5 Sensitivity of Experimental Impact Estimates to Method of Handling Earnings Outliers

	Male Youths	Female Youths
Unadjusted earnings data	−1,141.55 (492.05) [.02]	−72.78 (291.12) [.81]
Bloom et al. (1993) hand corrections	−867.33 (429.37) [.04]	−163.00 (262.90) [.53]
Top 1% trimmed within groups	−946.14 (411.37) [.02]	−119.47 (248.90) [.63]
Top 2% trimmed within groups	−805.66 (387.34) [.04]	−140.60 (238.31) [.56]
Top 3% trimmed within groups	−737.64 (374.44) [.05]	−141.75 (232.38) [.54]
Top 4% trimmed within groups	−656.59 (364.77) [.07]	−125.35 (223.84) [.58]
Top 5% trimmed within groups	−679.72 (355.87) [.06]	−119.21 (216.75) [.58]

Source: National JTPA Study 18 Month Impact Sample.

Note: These impact estimates are regression-adjusted using the same specification as in Bloom et al. (1993); results for the case using the hand corrections differ slightly from those in Bloom et al. because we were unable to exactly replicate their construction of some of the covariates. Estimates with trimming are obtained by dropping the indicated percentage of the earnings values from the top of the earnings distribution for each of the control and treatment groups for each demographic group in each month prior to calculating the impact estimates. Numbers in parentheses are estimated standard errors. Numbers in brackets are *p*-values from two-tailed tests of the null hypothesis that the true value of the coefficient is zero.

then corrected if necessary.[8] These hand corrections have a large effect for male youth, where they reduce the absolute value of the estimate by almost $300, or around 25 percent.

The remaining rows examine the alternative strategy of trimming off the top 1 to 5 percent of the raw earnings values in each month in each of the treatment and control groups prior to calculating the experimental impact

[8] The estimates in the first row of table 8.4 differ slightly from those in the second row of table 8.5 because the estimates in table 8.4 are not regression adjusted and because the random assignment ratio was changed from two treatment group members to each control group member at some centers for short periods. The latter causes the weighted (by the center sample sizes) average of the center-specific impact estimates to differ from the impact estimates obtained from the pooled sample of individuals.

estimates. The trimming procedure has the advantage that it is easier to replicate than the hand correction procedure. Estimates are obtained with trimming of the top 1, 2, 3, 4, and 5 percent of the earnings values. In the case where the outliers represent invalid values, the raw data can be thought of as a mixture of two distributions, one valid and one invalid, and the trimming acts to remove the invalid values. In the case where the outliers represent valid values, reporting estimates based on trimmed means can be justified on robustness grounds. For male youths, trimming has a marked effect on the magnitude of the impact estimates. The point estimate falls as the amount of trimming increases and ceases to be statistically significant at the 5 percent level when more than 3 percent of the observations are trimmed. There is little effect of trimming on the estimates for female youths.

8.7 Earnings Measures

We have assumed throughout this paper that earnings represent the outcome measure of interest in an evaluation. However, there are many alternative ways to measure earnings, and the specific measure chosen may affect the impact estimates obtained.[9] For example, earnings data from surveys may do a better job of capturing earnings in the underground economy but a poorer job of capturing regular earnings than administrative data from unemployment insurance (UI) records.

This section presents two pieces of evidence on the sensitivity of the NJS experimental impact estimates to the earnings measure used. The first piece of evidence is a comparison of 12-month impacts constructed using self-reported and UI administrative earnings data for a subsample of the NJS data with valid values for both measures. The second piece of evidence compares the 18-month impact estimates from the two official NJS impact reports submitted to the U.S. Department of Labor. Our evidence reveals surprising sensitivity of the experimental impact estimates for youth to seemingly modest changes in the construction of the earnings variable. This sensitivity is sufficient to affect the policy conclusions drawn from the NJS in some respects.

Table 8.6 compares 12-month impact estimates constructed using self-reported and UI administrative earnings data on a common sample. Confining the impact estimate to the first 12 months after random assignment

[9] A related issue that we do not address in detail here is whether other outcome measures should be included in an evaluation. The choice of whether to examine other outcome measures will affect the results of a cost-benefit analysis as benefits not measured are often not included. In some past evaluations, such as the nonexperimental evaluation of the Job Corps program described in Mallar et al. (1982), program impacts on factors other than earnings, such as crime, have been responsible for much of the overall benefit attributed to the program. Recent evaluations, such as the NJS, have tended to downplay these other outcomes.

Table 8.6 **Comparison of Experimental Impact Estimates Calculated Using Self-reported and Administrative Earnings Data**

Earnings Measure	Male Youths	Female Youths
Impact using self-reported data	−555*	6[a]
Impact using UI administrative data	−240[a]	21[a]
Difference in impacts	−315	−15
N	1,447	1,939

Source: Estimates drawn from Bloom et al. (1993, exhibit E.10).

Note: The estimates are calculated over the first 12 months after random assignment rather than the first 18 months after random assignment in order to maximize the number of observations with valid values for both earnings measures. The sample includes persons with valid values for both self-reported earnings and administrative earnings from state unemployment insurance (UI) records for the first year after random assignment. All persons at the Jersey City, New Jersey, and Marion, Ohio, training centers are excluded as UI earnings data are not available for those states.

[a]Not statistically significantly different from zero.

*Significant at the 10 percent level.

maximizes the size of the sample with valid values for both measures. For male youths, the estimate constructed using the self-reported earnings data is nearly twice as large as that constructed using the UI data and is statistically significant at the 10 percent level. For female youths, the difference is essentially zero. The findings for adults (not reported here) match those for male youths and reinforce the conclusion that which of the two earnings measure is used makes a difference in the resulting impact estimates.[10]

Table 8.7 compares the experimental impact estimates for the first 18 months after random assignment presented in the official 18- and 30-month impact reports submitted to the U.S. Department of Labor (Bloom et al. 1993; Orr et al. 1995) These estimates are broken down by the experimental treatment streams described earlier, which divide the sample based on the services recommended by JTPA staff prior to random assignment. The two sets of estimates differ in their construction in a number of important ways. In particular, in the 30-month impact report (1) persons with fewer than 18 months of self-reported earnings data had the remaining months filled in with UI earnings data when the UI data were available,

[10] Bloom et al. (1993) examine and reject explanations based on recall bias in self-reported earnings, missing UI earnings at centers near state borders, and measurement problems at specific centers for the differences in mean earnings between the self-reports and the UI administrative data. Smith (1997b) argues, based on comparisons with other earnings measures in the NJS and with other samples of similar populations, that the difference in impacts results from an apparent upward bias in the survey-based earnings measure. Inflating the means of both the treatment and control groups by a common factor increases the absolute value of the experimental impact estimate.

Table 8.7 **Comparison of Experimental Impact Estimates in Official NJS 18- and 30-Month Impact Reports**

Treatment Stream	18-Month Report	30-Month Report	Ratio of Estimates[a]
	Male Youths[b]		
CT-OS	−380	795 (899)	−.48
OJT	−2,392*	−1,814 (1,082)	1.32
Other services	−1,976*	249 (721)	−7.94
	Female Youths		
CT-OS	−792	174 (376)	−4.55
OJT	762	321 (892)	2.37
Other services	−271	−130 (759)	2.08

Source: Bloom et al. (1993, exhibits 6.7 and 6.12) for 18-month impact report. Orr et al. (1995, exhibit 5.17) for 30-month impact report.

Note: Table reports impact per enrollee calculated using Bloom (1984) estimator. No standard errors are reported for the estimates in the 18-month impact report. Numbers in parentheses are estimated standard errors for estimates from the 30-month impact report.

[a] Estimates from the two impact reports differ due to changes in sample composition, rescaling of self-reported overtime earnings in the 30-month impact report, and use of rescaled data from matched unemployment insurance earnings records in the 30-month impact report. See the text for more details.

[b] Male youth results refer to the full sample for the 18-month impact report and to the subsample of persons without a self-reported arrest between their sixteenth birthdays and the date of random assignment for the 30-month impact report.

*Significant at the 10 percent level.

(2) the UI earnings data were rescaled up by the ratio between the mean self-reported and UI earnings for each demographic group, (3) some persons who were excluded from the 18-month evaluation because they were randomly assigned late in 1989 were included, (4) only male youths without self-reported arrests between their sixteenth birthdays and the time of random assignment were included because it was found that the negative impact of the program for this group reported in the 18-month evaluation was concentrated among those with self-reported arrests, and (5) the overtime component of the self-reported earnings measure was scaled down in light of evidence of an upward bias in the reporting of this component of earnings. The most important of these factors are the use and rescaling of the UI data because they affect the largest fraction of the sample.

The table shows surprisingly large differences in impact estimates across official reports. The largest percentage effects are for female youths, where the ratio of the two estimates is at least 2.0 for all three treatment streams. The point estimate for the CT-OS treatment stream reverses sign and changes by nearly $1,000 between the two reports. For male youths, the statistical significance of the estimates in the 18-month report disappears in the 30-month report, and the CT-OS and other services stream estimates change sign. The estimates in table 8.6 make clear that much of the change in the male youth estimates results from changes in the earnings measure, not from the arbitrary restriction of the sample to persons without self-reported arrests prior to random assignment.[11]

8.8 Treatment Group Dropout and Control Group Substitution

In this section we discuss two issues that can have important effects on the impact estimates reported in an experimental evaluation and, more important, on the interpretation of those estimates. The first is treatment group members who drop out of the program prior to receiving treatment. The second is control group members who obtain substitutes for the experimental treatment from other sources.

8.8.1 Treatment Group Dropout

In order to reduce costs and minimize the disruption of normal JTPA operating procedures, random assignment took place at the JTPA office after recommendation for services, rather than at the service provider location prior to the start of services. This led to a substantial dropout problem in the NJS data, with around 40 percent of the treatment group never enrolling in JTPA (see Heckman, Smith, and Taber 1998).

In the presence of dropouts, the treatment group earnings distribution mixes the distributions of earnings for persons who have and have not received the treatment, instead of providing a clean estimate of the distribution of earnings conditional on treatment. The literature offers three strategies for dealing with dropouts. The first consists of reinterpreting the impact estimates as estimating the impact of "assignment to the treatment group"—sometimes called "intent to treat"—rather than of actual receipt of treatment. While the impact of assignment to treatment is often of interest in medical contexts, it is less interesting in the case of training programs.

The second strategy, developed in Bloom (1984), assumes that dropouts in the treatment group experience the same outcome they would have ex-

[11] Similar variability across the official reports is observed for the adult groups (Heckman, LaLonde, and Smith 1999). E.g., a reader of Bloom et al. (1993) would conclude that on-the-job training has the largest impact on adult women, while a reader of Orr et al. (1995) would conclude that "other services" is the best treatment for that group.

perienced had they been in the control group. Under this assumption, an estimate of the impact of treatment on the treated can be obtained by dividing the experimental impact estimate by one minus the fraction of the treatment group that drops out. This strategy is often plausible but fails when dropouts receive partial treatment, as they appear to in the NJS case.[12] Heckman, Smith, and Taber (1998) propose alternative identifying assumptions based on prior knowledge of the impact of partial treatment or prior knowledge of the ratio of the mean earnings in the untreated (control) state of persons who would and would not have been dropouts had they been randomly assigned to the experimental treatment group.

Heckman, Smith, and Taber (1998) find large differences between the unadjusted NJS experimental impact estimates and those provided by the Bloom (1984) method and their alternative identifying assumptions. For example, Bloom et al. (1993) report that the impact of assignment to treatment on the earnings of male youths aged 16–21 in the 18 months after random assignment is −$854, while the estimate of the effect of treatment on the treated for this group obtained using the Bloom (1984) method is −$1,356. At the same time, differences between the Bloom (1984) estimates and the alternatives in Heckman, Smith, and Taber (1998) are modest, given reasonable assumptions about the effectiveness of partial treatment or about the ratio of earnings of control group members who would and would not have been dropouts.

8.8.2 Control Group Substitution

Control group substitution arises in the evaluation of many training programs because there are often multiple programs serving the same clientele and because these programs often contract out to service providers who offer the same services to the general public. Cave and Quint (1991) find substitution in their evaluation of the Career Beginnings program, Puma et al. (1990) find it in their evaluation of the Food Stamp Employment and Training Program, and Riccio et al. (1994) find it in their evaluation of the GAIN program.[13]

Heckman, Hohmann, Smith, and Khoo (1999) document the importance of control group substitution in the NJS. Table 8.8, taken from their paper, shows the percentage of the treatment and control group members recommended to receive classroom training prior to random assignment (the CT-OS treatment stream) that actually received classroom training

[12] Doolittle and Traeger (1990) estimate that about half of the persons in the treatment group who did not formally enroll in JTPA, and who are therefore counted as dropouts in the official reports, received some form of JTPA services following random assignment. In most cases, these services were fairly minimal such as counseling or job search assistance.

[13] Heckman, LaLonde, and Smith (1999) provide evidence on control group substitution for a large number of social experiments.

Table 8.8 **Percentages of Experimental Treatment and Control Groups Receiving Classroom Training Services**

	Male Youths	Female Youths
Treatment group	55.7	58.6
Control group	34.5	40.1
Difference	22.2	18.5
p-Value for difference[a]	0.00	0.00

Source: National JTPA Study 18 Month Rectangular Sample. Statistics taken from Heckman, Hohmann, Smith, and Khoo (1999, table II).

Note: Sample consists of all persons in the CT-OS treatment stream in the NJS with valid values of earnings and training in the 18 months after random assignment. The CT-OS treatment stream consists of persons recommended to receive classroom training by JTPA staff prior to random assignment. The training measure used here includes only classroom training. Some persons in each group received other training services but not classroom training.

[a] p-Values are from tests of the null hypothesis that the difference between the percentages receiving training in the two groups is zero.

in the 18 months after random assignment, along with the p-value from a test of the null hypothesis of equality of the two percentages. For both youth groups, the data reveal substantial substitution into alternative classroom training services by controls, as well as a high rate of dropping out among the treatment group.[14] Substitution and dropping out also characterize the other two treatment streams in the NJS, though the rates of substitution are less because some other JTPA services, such as subsidized on-the-job training at private firms, are not widely available from alternative sources.

There are three standard methods for handling control group substitution. The first reinterprets the experimental impact estimate as estimating the marginal impact of the additional training provided by the program being evaluated relative to that received by the control group, rather than estimating the impact of training relative to no training. When the latter parameter is the object of interest, as it often is, this approach is unsatisfactory.

The other two methods use the experimental data to estimate the impact of training relative to no training. The second method relies on the assumption of either a common impact of training incidence (or of each hour of training) across persons or a varying impact of training incidence (or of each hour of training) whose idiosyncratic portion is either unknown to the person deciding to participate in training or not used in

[14] Some treatment group members in the CT-OS treatment stream enrolled in JTPA but received services other than classroom training. See Heckman, Hohmann, Smith, and Khoo (1999) for details.

making that decision. We call this (very strong) assumption A-1. Under assumption A-1,

$$\hat{\Delta} = \frac{\bar{Y}_t - \bar{Y}_c}{\bar{p}_t - \bar{p}_c}$$

estimates the mean impact of training relative to no training, where $\bar{Y}_t$ is mean earnings in the treatment group, $\bar{Y}_c$ is mean earnings in the control group, and $\bar{p}_t$ and $\bar{p}_c$ are either the fractions of the treatment and control group members receiving training or the mean hours of training received by treatment and control group members, respectively.

The third method uses the treatment group data to conduct a standard nonexperimental evaluation using the techniques in Heckman and Robb (1985), Heckman, LaLonde, and Smith (1999), and elsewhere. By comparing persons who receive training with persons who do not, these nonexperimental techniques also address both substitution and dropout.

Table 8.9 displays unadjusted experimental impact estimates, adjusted estimates based on assumption A-1, and estimates from two standard nonexperimental estimation procedures. The dependent variable is earnings in the first 12 months after training (for persons with a post-random-assignment classroom training spell) or the first 12 months after random assignment (for persons without such a spell). Using this dependent variable makes the estimates net of the opportunity costs (in terms of forgone earnings) of training and is consistent with the usual practice in nonexperimental evaluations. The sample for the experimental and adjusted experimental estimates consists of persons in the classroom training treatment stream with valid values of the dependent variable. The sample for the nonexperimental estimates consists only of treatment group members in the CT-OS treatment stream. Thus the nonexperimental comparison group consists of the treatment group dropouts.

The first panel of table 8.9 presents benchmark experimental impact estimates for this sample and dependent variable. The next panel presents estimates obtained under assumption A-1 expressed in terms of either training incidence or hours of training. The adjusted estimates are substantially larger in absolute value than the benchmark experimental estimates. Furthermore, the two versions of assumption A-1 yield very different estimates for male youths—$1,883 when A-1 applies to the incidence of training and $693 when A-1 applies to each hour of training.

The final panel presents nonexperimental estimates of the impact of training relative to no training. The first is the coefficient on a training receipt indicator from an OLS regression of earnings on the indicator and a vector of individual characteristics, while the second is the coefficient on the training indicator in the same regression with the difference between earnings before random assignment and earnings after training as the

Table 8.9 **Sensitivity of Impact Estimates to Method of Accounting for Control Group Substitution and Treatment Group Dropout**

	Male Youths	Female Youths
Unadjusted Experimental Estimate		
Experimental estimate	334.16	−52.61
	(510.57)	(290.74)
Adjusted Experimental Estimates Based on Assumption A-1[a]		
A-1 for training incidence	1,883.10	−253.90
	(2,877.28)	(1,403.12)
A-1 for each hour of training	693.31	−80.84
	(1,059.35)	(544.18)
Nonexperimental Impact Estimates Using the Experimental Treatment Group		
OLS[b]	1,653.61	1,645.79
	(542.13)	(309.24)
Difference in differences[c]	2,114.79	1,542.44
	(593.81)	(365.37)

Source: National JTPA Study 12 Month Post-Training Sample.

Note: The dependent variable in all cases except the difference-in-differences estimator consists of self-reported earnings in the 12 months after the first spell of classroom training following random assignment, for those with a classroom training spell, or the first 12 months after random assignment, for those without a classroom training spell. The sample for the unadjusted and adjusted experimental estimates consists of all NJS sample members in the CT-OS treatment stream with valid self-reported earnings and training data for the 12-month period indicated in the preceding sentence. The sample for the nonexperimental estimates consists of treatment group members meeting the same criteria. The measure of training includes only self-reported classroom training. Numbers in parentheses are estimated standard errors.

[a] The adjusted experimental estimates are constructed using the experimental treatment and control groups as described in the text. Assumption A-1 is that either training incidence (or each hour of training) has the same impact on everyone or the impact of training incidence (or each hour of training) varies but individuals do not know the idiosyncratic portion of their impact or do not use that information in deciding whether to take training. Reported estimates for the per hour case are at the mean hours of classroom training in the treatment group.

[b] The OLS estimates consist of the coefficient on an indicator variable for classroom training receipt in a regression of earnings on the training indicator and a vector of background variables. The comparison group for these estimates is the treatment group dropouts.

[c] The difference-in-differences estimates consist of the coefficient on an indicator variable for classroom training receipt in a regression of the difference between earnings before random assignment and earnings after random assignment or training on the training indicator and a vector of background variables. The comparison group for these estimates is the treatment group dropouts.

dependent variable. The two sets of nonexperimental estimates are quite close, and both sets are larger than the unadjusted experimental estimates.[15]

The lessons from this section are as follows. First, treatment group dropout and control group substitution are empirically important in the NJS. Second, taking account of them makes a difference in both the magnitude and the interpretation of the impact estimates. Moreover, doing so involves making the same type of nonexperimental assumptions that experiments attempt to avoid. We show that the impact estimates depend on which among the set of plausible assumptions is invoked in solving the substitution and dropout problems.

8.9 Summary and Conclusions

In this paper, we examine the sensitivity of the NJS experimental impact estimates for youth along several dimensions. Our analysis emphasizes that experimental impact estimates differ from nonexperimental estimates only in that they rely on random assignment. All of the normal issues that arise in any empirical evaluation, such as how to measure the outcome variable, what to do about outliers, and how to combine data from different training centers, arise in experiments just as they do in nonexperimental analyses. Other issues, such as treatment group dropout, control group substitution, and selection of the training centers to include in the evaluation, are unique to experiments or are more problematic in an experimental context.

We show that the magnitude and interpretation of the experimental estimates depend crucially on a number of these factors. We find the selection of which training centers to include in the evaluation and the construction and interpretation of estimates of the effect of training relative to no training in the presence of treatment group dropout and control group substitution to be the most important factors in the NJS youth data. In addition, we demonstrate the importance of the construction of the earnings variable used in the evaluation. The fact that the two official NJS impact reports submitted to the U.S. Department of Labor provide 18-month impact estimates for youth that change by over $1,000 in one case and that switch signs in several others illustrates this importance.

While our analysis does not indicate that experiments should be

[15] Heckman, Hohmann, Smith, and Khoo (1999) consider a number of other nonexperimental estimators. The nonexperimental estimates almost always exceed the unadjusted experimental estimates. At the same time, they emphasize that whether JTPA classroom training passes a cost-benefit test after taking account of substitution and dropout depends on assumptions about the longevity of training's impact on earnings and about the discount rate. For most demographic groups, plausible assumptions imply that JTPA classroom training produces a private benefit to its recipients but has negative net social benefits.

dropped in favor of a return to nonexperimental methods, it suggests the importance of examining the sensitivity of experimental impact results and the potential value of conducting multiple independent experimental evaluations of the same program. It also makes clear that experiments do not constrain the ability of an investigator to find what he or she wants to find as strongly as many advocates of experimentation hoped they would.

Our findings support moderation in the interpretation of the NJS youth results. The magnitudes of the impact estimates for male youths are sensitive along nearly every dimension we examine. The statistical significance of the negative male youth impact estimates is extremely fragile; it appears more likely that JTPA has a zero impact on male youths than a negative one. At the same time, the estimates for both youth groups are sensitive to the adjustments for control group substitution and treatment group dropout. Like Heckman, Hohmann, Smith, and Khoo (1999), we find that the effect of JTPA classroom training on earnings measured relative to no training, rather than relative to the available alternatives, is positive, though probably not positive enough to pass a social cost-benefit test.

Finally, the results presented in this paper emphasize the consistency of the JTPA impact estimates with earlier findings for other programs. For youths, the record of government training programs for the disadvantaged is almost uniformly negative.[16] Impacts on the earnings of dropouts in the NSW demonstration were negligible (Hollister et al. 1984). The CETA estimates for youth reported in Bassi (1984) are negative for males and negligible for females. Cave and Doolittle (1991) present experimental impact estimates from Jobstart, a youth program similar to the Job Corps but lacking its residential component. Its effect on earnings is negative for male youths and negligible for female youths. The one bright spot is the somewhat dated nonexperimental evaluation of the Job Corps by Mallar et al. (1982), which found a positive effect on participant earnings and criminal behavior sufficient to pass a cost-benefit test.[17] Unlike the other programs, the Job Corps involves a residential component, in which youth are removed from their neighborhoods to a separate camp with other Job Corps participants. It is also, unlike JTPA, quite expensive.

Though sensitive along several dimensions and, for JTPA classroom training, perhaps somewhat more positive than found for previous programs once adjusted for substitution and dropping out, the NJS impact estimates for youth fit comfortably into the pattern of several decades of research that finds very limited earnings effects for the types of services offered by JTPA.

[16] Heckman, Roselius, and Smith (1994), Heckman, LaLonde, and Smith (1999), Heckman, Lochner, Smith, and Taber (1997), and LaLonde (1995), among others, provide extended surveys of the literature on training.

[17] Their cost-benefit analysis does not include the deadweight costs associated with raising the funds for the program through taxation.

References

Bassi, Laurie. 1984. Estimating the effect of training programs with non-random selection. *Review of Economics and Statistics* 66 (1): 36–43.

Bloom, Howard. 1984. Accounting for no-shows in experimental evaluation designs. *Evaluation Review* 82 (2): 225–46.

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.

Cain, Glen, and Douglas Wissoker. 1990. A reanalysis of marital stability in the Seattle-Denver Income Maintenance Experiment. *American Journal of Sociology* 95:1235–69.

Cave, George, and Fred Doolittle. 1991. *Assessing Jobstart: Interim impacts of a program for school dropouts.* New York: Manpower Demonstration Research Corporation.

Cave, George, and Janet Quint. 1991. *Career beginnings impact evaluation: Findings from a program for high school students.* New York: Manpower Demonstration Research Corporation.

Devine, Theresa, and James Heckman. 1996. The structure and consequences of the eligibility rules for a social program: A study of the Job Training Partnership Act (JTPA). In *Research in labor economics,* ed. Solomon Polachek, 15:111–70. Greenwich, Conn.: JAI.

Dickinson, Katherine, Terry Johnson, and Richard West. 1987. An analysis of the sensitivity of quasi-experimental net impact estimates of CETA programs. *Evaluation Review* 11 (4): 452–72.

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

Hannan, Michael, and Nancy Tuma. 1990. A reassessment of the effect of income on marital dissolution in the Seattle-Denver experiment. *American Journal of Sociology* 95:1270–98.

Heckman, James, Neil Hohmann, Jeffrey Smith, and Michael Khoo. 1999. Substitution and dropout bias in social experiments: A study of an influential social experiment. *Quarterly Journal of Economics.* Forthcoming.

Heckman, James, Robert LaLonde, and Jeffrey Smith. 1999. The economics and econometrics of active labor market policies. In *Handbook of labor economics,* vol. 3, ed. Orley Ashenfelter and David Card. Amsterdam: North-Holland.

Heckman, James, Lance Lochner, Jeffrey Smith, and Christopher Taber. 1997. The effects of government policy on human capital investment and wage inequality. *Chicago Policy Review* 1 (2): 1–40.

Heckman, James, and Richard Robb. 1985. Alternative methods for evaluating the impact of interventions. In *Longitudinal analysis of labor market data,* ed. James Heckman and Burton Singer, 156–245. New York: Cambridge University Press.

Heckman, James, Rebecca Roselius, and Jeffrey Smith. 1994. U.S. education and training policy: A re-evaluation of the underlying assumptions behind the "new consensus." In *Labor markets, employment policy and job creation,* ed. A. Levenson and L. C. Solomon, 83–122. Boulder, Colo.: Westview.

Heckman, James, and Jeffrey Smith. 1996. Experimental and nonexperimental evaluation. In *International handbook of labour market policy and evaluation,* ed. Günter Schmid, Jacqueline O'Reilly, and Klaus Schömann, 37–88. London: Edward Elgar.

Heckman, James, Jeffrey Smith, and Christopher Taber. 1998. Accounting for dropouts in evaluations of social experiments. *Review of Economics and Statistics* 80 (1): 1–14.

Hollister, Robinson, Peter Kemper, and Rebecca Maynard, eds. 1984. *The National Supported Work demonstration.* Madison: University of Wisconsin Press.

Kemple, James, Fred Doolittle, and John Wallace. 1993. *The National JTPA Study: Final implementation report.* New York: Manpower Demonstration Research Corporation.

LaLonde, Robert. 1995. The promise of public sector-sponsored training programs. *Journal of Economic Perspectives* 9 (2): 149–68.

Mallar, Charles, David Long, Stewart Kerachsky, and Craig Thornton. 1982. *Evaluation of the impact of the Job Corps program: Third follow-up report.* Princeton, N.J.: Mathematica Policy Research.

Orr, Larry, Howard Bloom, Stephen Bell, Winston Lin, George Cave, and Fred Doolittle. 1995. *The National JTPA Study: Impacts, benefits and costs of Title II-A.* Bethesda, Md.: Abt Associates.

Puma, Michael, Nancy Burstein, Katie Merrell, and Gary Silverstein. 1990. *Evaluation of the Food Stamp Employment and Training Program: Final report.* Bethesda, Md.: Abt Associates.

Riccio, James, Daniel Friedlander, and Stephen Freedman. 1994. *GAIN: Benefits, costs and three-year impacts of a welfare-to-work program.* New York: Manpower Demonstration Research Corporation.

Smith, Jeffrey. 1997a. The JTPA selection process: A descriptive analysis. Manuscript.

———. 1997b. Measuring earnings levels among the poor: A comparison of two samples of JTPA eligibles. Manuscript.

This Page Intentionally Left Blank