Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme

Author(s): James J. Heckman, Hidehiko Ichimura and Petra E. Todd

Source: *The Review of Economic Studies*, Vol. 64, No. 4, Special Issue: Evaluation of Training and Other Social Programmes (Oct., 1997), pp. 605-654
Published by: Oxford University Press
Stable URL: https://www.jstor.org/stable/2971733
Accessed: 20-12-2024 12:52 UTC

Review of Economic Studies (1997) **64**, 605–654 0034-6527/97/00280605$02.00

Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme

JAMES J. HECKMAN
University of Chicago

HIDEHIKO ICHIMURA
University of Pittsburgh

and

PETRA E. TODD
University of Pennsylvania

First version received October 1994; *final version accepted June* 1997 (*Eds.*)

This paper considers whether it is possible to devise a nonexperimental procedure for evaluating a prototypical job training programme. Using rich nonexperimental data, we examine the performance of a two-stage evaluation methodology that (a) estimates the probability that a person participates in a programme and (b) uses the estimated probability in extensions of the classical method of matching. We decompose the conventional measure of programme evaluation bias into several components and find that bias due to selection on unobservables, commonly called selection bias in econometrics, is empirically less important than other components, although it is still a sizeable fraction of the estimated programme impact. Matching methods applied to comparison groups located in the same labour markets as participants and administered the same questionnaire eliminate much of the bias as conventionally measured, but the remaining bias is a considerable fraction of experimentally-determined programme impact estimates. We test and reject the identifying assumptions that justify the classical method of matching. We present a nonparametric conditional difference-in-differences extension of the method of matching that is consistent with the classical index-sufficient sample selection model and is not rejected by our tests of identifying assumptions. This estimator is effective in eliminating bias, especially when it is due to temporally-invariant omitted variables.

1. INTRODUCTION

This paper addresses the following question. It is *possible to devise a nonexperimental procedure for evaluating a prototypical job training programme that produces impact estimates and inferences about the programme that are very close to those produced from a randomized social experiment*? We combine nonexperimental data on persons who chose not to participate in the programme with data from a large scale social experiment to examine the performance of various matching methods, including new conditional difference-in-differences extensions of matching methods that we present here, in estimating an averaged version of the effect of treatment on the treated.

Matching methods pair programme participants with members of a nonexperimental control group who have similar observed attributes and estimate treatment impacts by subtracting mean outcomes of matched comparison group members from the mean outcomes of matched participants. We extend traditional matching methods by (a) incorporating exclusion restrictions across programme outcome and programme participation equations; (b) presenting weaker conditions under which matching is justified as an evaluation method than appear in the published literature; (c) incorporating prior information about the functional form of estimating equations, including additive separability between measured and unmeasured determinants of outcomes; (d) extending matching to a longitudinal context and producing a generalized difference-in-differences estimator that identifies parameters of interest under different, and generally weaker, assumptions than matching, and (e) providing a rigorous asymptotic distribution theory for the matching estimator under general conditions without invoking special assumptions about the distribution of the data.

Using data from the control group of an experiment in conjunction with data from several nonexperimental comparison groups, this paper tests the assumptions that justify our approach and the stronger assumptions that are traditionally invoked to justify matching methods. We reject the strong assumptions maintained in the literature but find support for the weaker assumptions that justify our generalized difference-in-differences extension of matching. Matching methods reduce the conventional measure of bias substantially for most groups but do not eliminate it entirely. Our generalized difference-in-differences estimator is generally more effective than conventional matching methods in removing bias from our data, especially when it is contaminated by temporally-invariant components of bias such as unobserved site and questionnaire effects.

We also address a second question. *What features of the nonexperimental data and of the matching method are essential in reducing the conventional measure of bias used in evaluation studies*? To address this question, we take as our point of departure the observation that ideal social experiments identify programme impacts by balancing many features of the data at the same time: (1) Participants and controls have the same distributions of unobserved attributes; (2) They have the same distributions of observed attributes; (3) The same questionnaire is administered to both groups, so outcomes and characteristics are measured in the same way for both groups; and (4) Participants and controls are placed in a common economic environment.

Features (2)–(4) can also be achieved in a nonexperimental evaluation. Resampling methods can be applied to nonexperimental data to mimic feature (2) of an experiment. Matching methods substantially reduce bias when features (3) and (4) characterize the nonexperimental data. When they do not, the conventional matching method can fail dramatically, as we demonstrate. However, our difference-in-differences extension of matching is more robust than conventional methods in data in which features (3) and (4) are absent.

The recent econometric literature on programme evaluation has emphasized feature (1)—elimination of selective differences in unobservables drawn from a common distribution for participants and experimental controls—as the principal benefit of randomized trials.[1] Our study suggests that this emphasis is misplaced. Features (2)–(4) are far more important to the success of the experimental method in evaluating the training programme we study than is feature (1). Selection bias, rigorously defined, is a *relatively* small part of bias as conventionally measured.

1. See, e.g. Ashenfelter (1978), Ashenfelter and Card (1985), and LaLonde (1986).

This paper emphasizes the interplay between data and method. Both matter in evaluating the impact of training on earnings for four different demographic groups. The consistency of our findings across these groups is striking. The effectiveness of any econometric estimator is limited by the quality of the data to which it is applied, and no programme evaluation method "works" in all settings. We produce some striking examples where estimators that perform well on good data perform poorly when applied to bad data. Failure to locate participants and comparison group members in the same labour market is a major source of evaluation bias; so is the failure to use the same definitions of outcome and explanatory variables, as often occurs when different surveys are administered to participants and comparison group members. Estimation methods also matter. Simple balancing of observables in the participant and comparison group samples goes a long way toward producing a more effective evaluation strategy.

This paper develops in the following way. Section 2 defines the evaluation problem and discusses the benefits of randomized experiments compared to nonexperimental methods. Section 3 shows how matching approximates randomization and presents the identifying assumptions that justify the method. Section 4 presents several extensions of the classical method of matching and the identifying assumptions that justify them. Section 5 summarizes the main findings from our previous empirical research. In that research, we use nonparametric methods to characterize the form of the evaluation bias that arises from using members of a comparison group of self-selected nonparticipants to proxy what participants would have experienced if they had not participated in the programme we study. We discuss the implications of this research for the design of a successful evaluation strategy. This evidence motivates our extensions of the method of matching. Section 6 reviews our evidence on the determinants of programme participation. Section 7 defines the main samples and models used in this paper. Section 8 demonstrates that the distributions of the matching variables are different for participant and comparison group samples. This finding has significant implications for understanding the sources of evaluation bias as conventionally measured. Section 9 decomposes conventional measures of evaluation bias into components due to (a) selection on unobservables, (b) failure to compare participants and controls at common values of matching variables, and (c) failure to weight the two groups comparably. We show that bias due to mismatching and misweighting of the data is numerically more important than bias due to selection on unobservables. Comparing the incomparable is a major source of evaluation bias. Yet the bias left over after adjusting for weighting and mismatching is still large compared to the experimentally-estimated treatment impact. Matching reduces bias but is not guaranteed to produce reliable estimates of programme impacts. Our difference-in-differences extension of the method of matching is usually more effective than conventional methods in the samples we analyse. Section 10 uses the experimental data to test the identifying assumptions invoked in the matching literature and our various extensions of it.

Section 11 defines a class of matching estimators and longitudinal and repeated cross section difference-in-differences estimators within a unified framework and Section 12 compares the empirical performance of alternative matching methods and our extensions of them. We measure the effectiveness of a nonexperimental estimator by how well it eliminates differences in earnings between a nonexperimental comparison group and a randomized-out control group.

We examine the features of the data that attenuate bias in our samples. Section 13 studies the effectiveness of different matching estimators when the probabilities of participation on which the matches are based are estimated with progressively coarser conditioning information. Several estimators perform moderately well for all demographic

groups when data on recent labour market histories are included in estimating the probability of participation, but not when earnings histories or labour force histories are absent.

Section 14 analyses comparison group samples drawn from SIPP data to assess the importance of controlling for geographical location and of using the same survey instrument to collect comparison group data. Our results indicate that geographical proximity and uniformity of the survey instrument across treatment and comparison group samples are necessary features of a successful evaluation study of earnings impacts. This evidence confirms the importance of local labour markets in determining wages—a point emphasized in the research of Blanchflower and Oswald (1994). The major source of bias arising from the application of nonexperimental estimators to evaluate training programmes that is reported in LaLonde (1986) arises from the mismatch of questionnaires and labour markets across treatment group and comparison group members, and not because of the failure of econometric estimators to eliminate selection bias.

Section 15 investigates the performance of matching estimators when dropouts or "no-shows" are used as a comparison group. Programme dropouts are located in the same labour market and are administered the same questionnaire as programme completers, so this group automatically satisfies two key requirements for a successful nonexperimental evaluation. Our evidence on the performance of "no shows" as a comparison group is mixed. Section 16 summarizes the evidence.

2. THE EVALUATION PROBLEM AND THE BENEFITS OF RANDOMIZED EXPERIMENTS

The evaluation problem is a missing data problem. At any time, persons may be in either one of two potential states but not in both.[2] The states associated with receiving treatment and not receiving treatment are denoted "1" and "0" respectively. Outcomes are (Y_1, Y_0). Let $D=1$ if a person is in state "1"; $D=0$ otherwise. The outcome observed for an individual is Y defined as

$$Y=DY_1+(1-D)Y_0.$$

This is the Fisher model (1951), the Roy Model (1951) or the switching regression model of Quandt (1972).[3] The gain from participating in the programme is $\Delta=Y_1-Y_0$. If we could simultaneously observe Y_1 and Y_0 for the same person, there would be no evaluation problem since one could construct Δ for everyone.

To cast the discussion in familiar econometric notation, write outcomes as a function of observables (X) and unobservables (U_1, U_0)

$$Y_1=g_1(X)+U_1, \tag{1a}$$

$$Y_0=g_0(X)+U_0. \tag{1b}$$

The conventional econometric approach maintains that $E(U_1|X)=0$ and $E(U_0|X)=0$ and further assumes that g_1 and g_0 are nonstochastic functions. For the familiar case of linear regression, the g functions specialize to $g_1(X)=X\beta_1$, and $g_0(X)=X\beta_0$. These functional form assumptions are not required to implement matching estimators, but we use them in certain semiparametric extensions of the method of matching.

2. For simplicity we consider the two outcome case. Extension to a multiple-outcome switching model is straightforward.

3. Statisticians sometimes call this the "Rubin model" after a clear exposition of Fisher's model of experiments presented by Rubin (1978).

The most commonly-used evaluation parameters are means.[4] One mean receives the most attention: *the mean effect of treatment on the treated.* This parameter is

$$E(Y_1 - Y_0 | X, D=1) = E(\Delta | X, D=1)$$
$$= g_1(X) - g_0(X) + E(U_1 - U_0 | X, D=1). \quad (2)$$

The matching methods discussed in this paper focus on estimating an averaged version of this parameter

$$M(S) = \frac{\int_S E(\Delta | X, D=1) dF(X | D=1)}{\int_S dF(X | D=1)}, \quad (3)$$

where S is a subset of the support of X given $D=1$.

This mean answers the question "How much did persons participating in the programme benefit compared to what they would have experienced without participating in the programme?" This parameter is the gross gain to participants from the programme. When compared with costs, this parameter is informative on the question of whether or not an existing programme's benefits exceed its costs and whether the programme should be kept or terminated, provided that the potential outcomes in the no treatment state for all persons are good approximations to the no-programme outcome state for both participants and nonparticipants.[5] It is a nonstandard parameter from the vantage point of conventional econometrics because it combines "structure" (the g_0 and g_1 functions) with the means of error terms (U_0 and U_1).[6]

Social experiments recover the conditional distribution of Y_0, $F_0(y_0 | D=1, X)$, if randomization is administered at a stage of the application and enrollment process at which persons would ordinarily be accepted into programme, if the attrition from the programme is random, and if randomization does not disrupt the programme.[7] The evaluation problem arises because ordinary observational data do not provide sample counterparts for the missing counterfactual Y_0 values for participants ($D=1$). Experiments supplement observational data by providing the information needed to form the sample counterpart of $E(Y_0 | D=1, X)$ and hence to construct parameter $M(S)$.

To see how randomization solves the evaluation problem, consider randomization among persons who have applied to and been provisionally accepted into a social programme. Persons in the $D=1$ population are selected to receive programme services by a random device. Let $R=1$ if an eligible provisionally-accepted applicant is randomized into the programme; $R=0$ otherwise. It is assumed that if $R=1$, persons accept admission into the programme and receive services and if $R=0$ they do not obtain programme services.

4. Heckman (1992), Heckman, Smith and Clements (first draft 1993, 1997), and Heckman and Smith (1997) discuss other parameters derived from the distribution of outcomes in the programme impacts.

5. Heckman and Smith (1997) and Heckman (1997) consider this parameter in the context of cost benefit analysis and present precise conditions under which it identifies an economically-interpretable parameter. Heckman and Smith (1997) present cost estimates for the programme evaluated here under different assumptions about the social opportunity cost of funds to finance the programme.

6. Heckman and Robb (1985, 1986) present conditions for identifying this parameter using instrumental variables. Their conditions apply to the general "variable treatment effect case" of equations 1(a) and 1(b). See also Heckman (1997) for the implicit behavioural assumptions invoked in using instrumental variables to estimate parameter (2) when responses to treatment are heterogeneous.

7. Randomization at eligibility generates the same information plus the information required to identify $\Pr(D=1 | X)$. See Heckman (1996).

Recall that it is not necessary to assume that $E(U_1 | X) = 0$ or $E(U_0 | X) = 0$, so X can fail to be exogenous in the conventional sense of that term. The lack of any requirement for exogeneity highlights both the unconventional nature of the parameter of interest and the benefits of randomization in estimating it.

We can write observed outcomes for the entire population as $Y = D[RY_1 + (1-R)Y_0] + (1-D)Y_0$, so

$$E(Y | X, D=1, R=1) = E(Y_1 | X, D=1) = g_1(X) + E(U_1 | X, D=1), \tag{4a}$$

$$E(Y | X, D=1, R=0) = E(Y_0 | X, D=1) = g_0(X) + E(U_0 | X, D=1). \tag{4b}$$

Randomized-out controls provide the data that can be used to estimate counterfactual (4b). Conditional mean (4a) can be consistently estimated using ordinary observational data on programme participants.

Subtract (4b) from (4a) to obtain

$$\begin{aligned} &E(Y | X, D=1, R=1) - E(Y | X, D=1, R=0) \\ &\quad = g_1(X) - g_0(X) + E(U_1 - U_0 | X, D=1) = E(\Delta | X, D=1). \end{aligned} \tag{5}$$

This parameter can be consistently estimated using sample counterparts to population means.[8] One randomization identifies an entire function $E(\Delta | X, D=1)$ over any subset of the support of X given $D=1$.[9]

3. MATCHING AS A SUBSTITUTE FOR EXPERIMENTS

Nonexperimental methods use data on members of a nonexperimental comparison group (for whom $D=0$) to infer counterfactual outcomes for participants. A widely-used method is matching. The conventional method of matching estimates parameter $M(S)$, using non-experimental data by assuming that conditional on X, (Y_1, Y_0) and D are independent

$$(Y_1, Y_0) \perp\!\!\!\perp D | X. \tag{A-1}$$

where "$\perp\!\!\!\perp$" denotes independence. See, e.g. Rosenbaum and Rubin (1983). If (A-1) is true, then

$$F(y_0 | X, D=1) = F(y_0 | X, D=0),$$

so conditional on X non-participant outcomes have the same distribution that participants would have experienced if they had not participated in the programme. As a consequence, if the mean exists,

$$E(Y_0 | X, D=1) = E(Y_0 | X, D=0) = E(Y_0 | X), \tag{6}$$

and the missing counterfactual mean can be constructed from the outcomes of non-participants.

If, in addition, it is assumed that

$$0 < \Pr(D=1 | X) < 1, \tag{A-2}$$

for all X, then (2) can be defined for all values of X. (A-1) and (A-2) together are called "strong ignorability" by Rosenbaum and Rubin (1983). Under these conditions,

8. Heckman, Ichimura, Smith and Todd (1996b, c) develop, justify and apply nonparametric methods for estimating this parameter.

9. Heckman (1996) shows how randomization acts as an instrumental variable.

experimental and non-experimental analyses identify the same parameters. It is clear that for the purposes of estimating (2) or $M(S)$, the weaker assumption

$$Y_0 \perp\!\!\!\perp D \mid X, \tag{A-3}$$

is enough to identify those parameters.[10]

In the language of Heckman and Robb (1985), matching assumes that selection is *on observables*. Conditional independence or mean independence are strong assumptions. There may be variables apart from X on which the analyst cannot condition that affect both Y_0 and D.[11] In this case selection is on unobservables, as defined by Heckman and Robb. Assumptions (A-1) or (A-3) impose behavioural assumptions that (Y_1, Y_0) or Y_0 do not determine D conditional on X. This rules out selection into the programme based on unobserved (by the analyst) outcomes. More precisely (A-1) implies that $\Pr(D=1 \mid X, Y_1, Y_0) = \Pr(D=1 \mid X)$ while (A-3) implies that $\Pr(D=1 \mid X, Y_0) = \Pr(D=1 \mid X)$. These assumptions are at odds with those invoked in many economic models of self selection such as the Roy model (see e.g. Heckman and Honoré (1990)) and assume either that agents do not act on potential outcomes in deciding to participate in the programme or else that econometricians have as much information about the programme being studied as the agents making decisions. See Heckman and Smith (1997) for a more extensive discussion of the implicit behavioural assumptions that justify the method of matching.

It is often difficult in practice (i.e. with samples of typical size) to match on high dimensional X. This is the matching version of the curse of dimensionality. Rosenbaum and Rubin derive an important practical result. Let $\Pr(D=1 \mid X) = P(X)$. They demonstrate that (A-1) and (A-2) together imply $(Y_1, Y_0) \perp\!\!\!\perp D \mid P(X)$ and hence $Y_0 \perp\!\!\!\perp D \mid P(X)$. This insight shows that matching can be performed on $P(X)$ alone, reducing a potentially high dimensional matching problem to a one dimensional problem, provided that $P(X)$ is known.

By aligning the distribution of observed characteristics in the $D=0$ population with that in the $D=1$ population, matching mimics one feature of randomized data. Randomization within the $D=1$ population ensures that the distributions of X for participants ($R=1$ and $D=1$) and non-participants ($R=0$ and $D=1$) are the same. But there are other features of randomized data that are not so easily achieved by applying the method of matching to nonexperimental data.

A major limitation of nonexperimental methods compared to experimental methods for estimating $M(S)$ is that they do not guarantee that the support for the comparison group equals the support for programme participants. This condition is obviously satisfied in data generated from an experiment, i.e. Support $(X \mid D=1, R=1) =$ Support $(X \mid D=1, R=0)$. The inability to find comparable comparison group members for programme participants is a major source of bias and $M(S)$ often cannot be identified for all subsets S in the support of X given $D=1$. If the support in an experiment differs from the support common to participant and comparison group members in a nonexperimental study, different parameters are implicitly defined and estimated in the two types of studies. Below

10. The assumption of conditional independence of Y_1 given X is useful if the parameter of interest is the mean impact of treatment on the untreated. In that case, estimates of Y_1 for persons with $D=0$ would be inferred from data on persons with $D=1$, instead of inferring estimates of Y_0 for persons with $D=1$ from data on persons with $D=0$ as in this paper. The mean treatment impact on a randomly assigned person can be obtained as the weighted average of the estimates of mean impact of treatment on the treated and on the untreated, under an assumption like (A-1) on both Y_1 and Y_0.

11. Even if one can condition on X in one sample, there is no guarantee that the same variables are available in other samples.

we present some empirical evidence on the importance of this source of noncomparability bias across experimental and nonexperimental studies of the same programme.

Second, both participants and controls in a randomized experiment are usually administered the same questionnaire so outcomes and personal characteristics are measured the same way for both groups. In contrast, observational studies often combine two separate data sets for participants and non-participants that are collected using different survey instruments and different survey definitions of the same economic concept, such as earnings.

Third, both participants and controls reside in the same local labour market. Matching methods are far more effective in recovering the parameter of interest when the comparison group and treatment group both reside in the same local labour markets. For the main body of nonexperimental data that we analyse, programme applicants and nonapplicants come from the same narrowly-defined geographical areas (cities). Both the levels and dynamics of earnings and employment are affected by the conditions of the local labour market in which persons are located.

Table 1 presents features of the nonexperimental comparison groups used in previous evaluations of major U.S. job training programmes. Rows 1 and 2 reveal that few studies have nonexperimental comparison group members located in the same labour markets as programme participants and many use samples that do not administer the same questionnaire to both participants and comparison group members. LaLonde's comparison groups suffer from both defects. A major conclusion of this paper is that placing comparison group members in the same economic environment and administering them the same questionnaire as participants substantially improves the performance of nonexperimental estimators.

4. EXTENSIONS OF MATCHING

Our companion paper extends the received matching framework of Rosenbaum and Rubin in several ways: (1) by developing an asymptotic distribution theory for kernel-based matching estimators both for the cases when $P(X)$ is known and when it is estimated; (2) by demonstrating how the efficiency of the matching estimator can be improved by exploiting exclusion restrictions in terms of variables that appear in outcome and participation equations; (3) by demonstrating how conventional functional form restrictions invoked in econometrics—like additive separability of outcome equations—might improve the efficiency of estimates obtained from matching and (4) by extending matching to a longitudinal setting. A major conclusion of our analysis is that even if $P(X)$ were known, it might be less efficient to condition on it in constructing matches rather than on the original X. (See Heckman, Ichimura and Todd (1993, revised 1997).) We also demonstrate that the ignorability conditions are overly strong for the estimation of (2), or an averaged version of it. All that is required is a weaker mean independence version formulated in terms of $P(X)$

$$E(Y_0 \mid P(X), D=1) = E(Y_0 \mid P(X), D=0). \tag{A-4}$$

(See also Heckman and Robb (1986).) (A-4) is a consequence of (A-1) and (A-2) but can be maintained as a separate and weaker assumption. The theoretical results justified in our companion paper are derived under this assumption.

It is often both conceptually and empirically fruitful to partition X into two components: $X=(T, Z)$ where the T are variables in the outcome equations and the Z are

variables in the participation equation. Thus

$$Y_1 = g_1(T) + U_1, \tag{7a}$$

$$Y_0 = g_0(T) + U_0, \tag{7b}$$

and $\Pr(D=1|Z)$ is the probability of programme participation. Since outcomes are measured after enrollment, and different factors operate on outcomes and enrollment decisions, Z and T may contain distinct variables, although they may share some variables in common. We consider several generalizations of (A-3) and (A-4) that apply to the residuals from models 1(a) and 1(b): $U_0 \perp\!\!\!\perp T|D, Z$ so

$$\Pr(U_0 \leqq u_0 | T, Z, D) = \Pr(U_0 \leqq u_0 | Z, D), \tag{A-3′}$$

and $U_0 \perp\!\!\!\perp D|P(Z)$ or

$$E(U_0|T, Z, D) = E(U_0|Z, D), \tag{A-4′}$$

and $E(U_0|P(Z), D=1) = E(U_0|P(Z), D=0)$.

Under these assumptions, it is possible to improve on the efficiency of the unrestricted matching estimator and invoke the exclusion restrictions in (A-4′). This leads to the *regression-adjusted matching estimator* formally justified in our companion paper (Heckman, Ichimura and Todd (1997)) and empirically implemented in this paper.

We extend matching to a panel or repeated cross-section context in a new nonparametric conditional difference-in-differences estimator. This estimator was first proposed in Heckman, Ichimura, Smith and Todd (1995a, revised 1996b), but has precedents in the work of Heckman and Robb (1985, 1986). Let t represent a time period after the programme start date and t' a time period before the programme. Our conditional difference-in-differences estimator compares the conditional before–after earnings of programme participants with those of non-participants. It extends the conventional difference-in-differences estimator by defining outcomes conditional on X and using semiparametric methods to construct the differences. The population moments to which the estimator converges under standard conditions are

$$D_{t,t'}(X) = E(Y_{1t} - Y_{0t'}|X, D=1) - E(Y_{0t} - Y_{0t'}|X, D=0).$$

An estimator based on sample analogues to these population moments is robust to temporally-persistent separable components of bias including those that might arise from geographical or questionnaire mismatch between participants and members of the control group.

Term $D_{t,t'}(X)$ identifies $E(\Delta|X, D=1)$ if the following assumption holds

$$E(Y_{0t} - Y_{0t'}|X, D=1) = E(Y_{0t} - Y_{0t'}|X, D=0). \tag{A-5}$$

Under additive separability this condition is equivalent to

$$E(U_{0t} - U_{0t'}|X, D=1) = E(U_{0t} - U_{0t'}|X, D=0),$$

or $B_t(X) = B_{t'}(X)$. Under index sufficiency the condition becomes

$$E(U_{0t} - U_{0t'}|P(Z), D=1) = E(U_{0t} - U_{0t'}|P(Z), D=0). \tag{A-5′}$$

If Assumptions (A-3) or (A-4) or (A-3′) or (A-4′) are satisfied at times t and t', the difference-in-difference version of the estimator will also be justified. However Assumptions (A-5) and (A-5′) are weaker than those assumptions and are consistent with the index-sufficient sample selection bias model, as noted in Heckman, Ichimura, Smith and Todd

(1996b). The difference-in-difference estimator is less demanding of the data than the sample selection estimator in that it does not require a set of X values where there is no selection bias (i.e. for which $E(U_0 | X, D=0)=0$).[12] However, it is also more demanding because it requires pre-programme data. From an economic standpoint it is an attractive estimator because, unlike conventional matching estimators, it permits selection to be based on potential programme outcomes and allows for selection on unobservables. In particular, it is consistent with a Roy model of self selection applied to a panel setting. Heckman and Robb (1985) present a parametric version of our difference-in-differences estimator, which is used in Ashenfelter and Card (1985).

5. OUR PREVIOUS EVIDENCE ON THE FUNCTIONAL FORM OF SELECTION BIAS

In a series of papers (Heckman and Roselius (1993), Heckman, Ichimura, Smith and Todd (1994, revised 1996b), Todd (1995)), we have used semiparametric methods to characterize the form of the bias that arises from using mean outcomes of comparison group members to proxy the mean outcomes that participants in a programme would have experienced if they had not participated in it. Let $B(X)$ be the bias for a particular value of X. It is defined as

$$B(X)=E(Y_0 | X, D=1)-E(Y_0 | X, D=0). \tag{8}$$

Under the conditions that justify matching, $B(X)=0$ for all X where the bias is defined.

For the additively-separable case of (1a) and (1b), the bias is

$$B(X)=E(U_0 | X, D=1)-E(U_0 | X, D=0). \tag{9}$$

Our papers present evidence for a variety of demographic groups that the bias function $B(X)$ is described by two main features. First, assuming additive separability, the bias has an *index property*. Let $P(X)=\Pr(D=1 | X)$ be the probability of participation in the programme. If the bias has the index property

$$B(X)=\tilde{B}(P(X)), \tag{P-1}$$

where $\tilde{B}$ is a function of a single index, $P(X)$, and X enters the bias function solely through the index. This representation is consistent with a broad class of widely-used index function models described in Heckman and MaCurdy (1985) and is at the heart of the conventional sample selection bias model (Heckman (1980)) and the Roy model of self selection (Heckman (1990b) and Heckman and Honoré (1990)). An index representation greatly simplifies the characterization of the problem of evaluation bias and focuses attention on the probability of selection as a central ingredient to the formulation and solution of the evaluation problem. It plays an important role in the development of semiparametric methods for solving the selection problem, which we present elsewhere. (Heckman, Ichimura, Smith and Todd (1994, 1996b).) Observe that if $\tilde{B}(P(X))$ is the same in periods t and t', then if $Z=X$, Assumption (A-5′) is justified. Thus an estimator based on the conditional difference-in-differences moment condition (A-5′) is consistent with the index-sufficient selection estimator if, for example, the bias is constant over time or if it is symmetric around the date of entry into the programme ($t=0$). Heckman, Ichimura, Smith and Todd (1996b) present graphical evidence in support of such symmetry except for low values of P for the data analysed in this paper.

12. Heckman (1990a, b) and Heckman, Ichimura, Smith and Todd (1996b) discuss the importance of this condition in using the index-sufficient self selection model.

Second, our research in this paper and other papers (Heckman, Ichimura, Smith and Todd (1996a, b)) investigates whether the bias is *balanced* both for participants and controls i.e. whether

$$B(X) = \tilde{B}(P(X)) = 0. \tag{P-2}$$

This is the consequence of Assumptions (A-1) and (A-2) that are invoked to justify matching including the weaker version introduced in Heckman and Robb (1986). The research reported in this paper decisively rejects (P-2) for four demographic groups.

6. DETERMINANTS OF PROGRAMME PARTICIPATION

The probability of participation in the programme being evaluated ($P(X)$) is a key ingredient of our empirical strategy for characterizing and solving the evaluation problem. It is a central feature of the econometric model for index-sufficient selection bias and for matching (see Heckman and Robb (1986)). Heckman and Smith (1994) present an extensive analysis of the determinants of participation in the JTPA programme. We briefly summarize their findings concerning the relative importance of background characteristics, recent labour force status and earnings histories in the participation process for eligible persons.

Their main conclusion is that for all demographic groups recent unemployment histories are important predictors of participation in training programmes. Trainees enter the JTPA programme as a form of job search. For adult women, recent marital histories are also important since recently-divorced or separated women in life cycle transitions are more likely to participate in the programme than are others. Models of the participation decision based on variables that predict job-seeking are much better able to predict participation than are models that include only demographic characteristics like education, age and race.

Heckman and Smith find a dip in the earnings of participants shortly before they apply to the programme. This earnings pattern was first noted by Ashenfelter (1978) for a predecessor to the JTPA programme. However it is unemployment dynamics and not earnings dynamics that best predicts who goes into the programme. In terms of labour force status, unemployment peaks for participants just prior to the date of enrollment but there is little change in unemployment status for members of a nonexperimental comparison group. Unemployment increases for participants both because employed persons lose their jobs and seek the assistance of training programmes and because persons previously out of the labour force enter it and use training programmes as a vehicle for labour force entry. Ashenfelter's earnings dip is a consequence of a more fundamental process of unemployment dynamics. The evidence presented by Heckman and Smith shifts the emphasis in the evaluation of job training programmes away from a focus on controlling for earnings histories, and longitudinal methods based on them (as developed in Ashenfelter (1978), Heckman (1978), Heckman and Robb (1985, 1986), and Ashenfelter and Card (1985)), and toward models based on unemployment dynamics as predictors of participation and as variables to control for bias. They further note that the main determinants of participation are not merely the consequence of eligibility rules. Among eligible persons, persons seeking work are more likely to enter job training programmes.

7. THE JTPA DATA

The National JTPA (Job Training Partnership Act) Experiment was commissioned by the U.S. Department of Labor for the purpose of evaluating the main U.S. government

training programme for disadvantaged workers. The JTPA programme provides on-the-job training, job search assistance, and classroom training to youth and adults, who qualify for the programme under title IIA of the National Job Training Partnership Act. Persons become eligible for the programme by having a family income near or below the poverty level for six months prior to application or by participating in federal, state or local welfare and foodstamp programmes.[13] The fact that eligibility is based on a relatively short earnings history makes it possible for some highly skilled or trained individuals to become eligible after a short period of unemployment.

Devine and Heckman (1996) present a detailed analysis of JTPA eligibility rules and of their implications for the eligible population. Barnow (1993) compares the eligibility rules of JTPA to those of other training programmes, including its predecessors, CETA (Comprehensive Employment and Training Act) and MDTA (Manpower Development and Training Act), and finds only minor differences. The JTPA programme we analyse is typical of a wide array of training programmes implemented in the U.S. and abroad.

Under our supervision, the JTPA experiment collected longitudinal data on a group of treatments and randomized-out controls as well as on a comparison group of eligible nonparticipants (ENPs). Two-thirds of programme applicants were assigned to treatment and one third were randomized out and denied access to JTPA services for 18 months to form a control group. Persons were assigned to the control group only after they had applied to the JTPA programme, been declared eligible, and been accepted into the programme. The samples used in this study come from four of the sixteen JTPA training sites participating in the experiment.[14]

Members of the eligible nonparticipant comparison group (ENPs) reside in the same narrowly-defined geographic regions as the programme applicants and are eligible for the programme but do not apply to it. ENP comparison group members were administered the same survey instrument as randomized-out controls, which includes detailed retrospective questions on labour force participation, job spells, earnings, marital status, training and schooling activities, transfer programme participation and other demographic characteristics. We combine data from a baseline survey and from two follow-up surveys to form a thirty-six month panel data set.[15] It is divided into the eighteen month period before

13. The specific eligibility criteria are as follows. A person is considered economically disadvantaged and therefore eligible for employment and training services provided under the Act, if he or she: (1) receives, or is a member of a family that receives, cash payments under a Federal, State or local welfare programme; (2) has, or is a member of a family which has, received a total family income for the six month period prior to application for the programme involved (inclusive of unemployment compensation, child support payments and welfare payments), which in relation to family size, was not in excess of the higher of (a) the Office of Management and Budget poverty level or (b) 70% of the lower living standard income level; (3) is receiving food stamps pursuant to the Food Stamp Act of 1977; (4) qualifies as a homeless individual under Section 103 of the Stewart B. McKinney Homeless Assistance Act; (5) is a foster child on behalf of whom State or local government payments are made; or (6) is an adult handicapped individual whose own income meets the requirements of the clause but who is a member of a family whose income does not meet such requirements. (See Job Training Partnership Act of 1982, Public Law 97-300, 29 USC 103.) In addition, JTPA training sites may admit up to 10% of those served under Title IIA for reasons other than eligibility through these criteria. See Devine and Heckman (1996).

14. Kemple, Doolittle and Wallace (1993) provide detailed descriptions of all 16 experimental sites. Eligible nonparticipant data was only collected at four of the sites: Fort Wayne, Indiana; Corpus Christi, Texas; Jersey City, New Jersey; and Providence, Rhode Island.

15. The baseline survey collected retrospective monthly data on demographic characteristics, earnings histories, labour market histories, participation in government transfer programmes and participation in schooling or training activities. A follow-up survey, administered twelve to twenty four months later, collected similar information. The response rate for this survey was around 84%. The sample used includes persons who (1) had a follow-up interview scheduled at least 18 months after random assignment, (2) responded to the survey, and (3) had usable earnings information for at least 14 of the 18 months after random assignment. Appendix E (available on request) and Smith (1994) contain additional information on the design and collection of the ENP sample.

random assignment (or before the time of eligibility determination for ENPs) and an eighteen month post-random assignment period.[16]

The ENP data we have collected to construct a non-experimental comparison group are very rich compared to the information available to previous analysts. Table 1 presents a description of the data used to construct non-experimental comparison groups in some major evaluations of job training programmes. The final column describes the JTPA data. None of the previous studies could accurately determine whether persons in the comparison groups were eligible for the programme being studied. Some lacked basic demographic information. Few had access to monthly data on earnings or on recent labour force histories. None located comparison groups members in the same labour markets as the participants, and because of the lack of geographical information, no adjustment for local labour market conditions was possible. Several studies—including the influential LaLonde study—used data that administered different questionnaires to participants and comparison group members with different definitions of earnings and reporting frames.

In light of the evidence presented below that geographical misalignment and misalignment of concepts used in questionnaires is a major source of evaluation bias, it is clear that previous analysts evaluating training programmes operated with their hands tied. In light of the evidence in Heckman and Smith (1994) that recent unemployment histories are important predictors of participation in job training programmes, it is significant that most of the major nonexperimental evaluations did not have access to this critical piece of information.

8. ESTIMATING THE PROBABILITY OF PARTICIPATION $P(X)$

We estimate logit models of programme participation for eligible persons. Predictor variables are chosen to maximize the within-sample correct prediction rates using the hit or miss method.[17] For all four demographic groups analysed in this paper, recent unemployment histories are powerful predictors of participation. Recent earnings in the six months prior to entering the programme also increase the prediction rate, and are therefore included in the model. Appendix D and Appendix E, available on request, define the regressors used in this paper and present estimated logit coefficients.[18]

Figure 1 presents a major finding noted in our previous papers.[19] For all four demographic groups, histograms of estimated probabilities of programme participation for ENPs ($D=0$) and controls ($D=1$) reveal that the estimated support common to both distributions of P is very limited. Since $B(X)=\tilde{B}(P(X))$ is only defined for values of P common to the supports of ENP and control group members, in general the bias is not defined over the full P support of either ENP or control group members. The parameter $M(S)$ estimated experimentally is not necessarily the same as $M(S)$ estimated non-experimentally if programme outcomes depend on the support. Not only are the supports different, but shapes of the distributions of P differ over the common support. We next

16. Additional detail about these surveys and about the construction of variables used in the analysis is presented in Appendix E, available on request from the authors.

17. The method classifies an observation as "1" if the estimated $P(X)$, satisfies $\hat{P}(X)>\hat{P}$, the sample proportion of eligible persons taking training and "0" otherwise. This method maximizes the overall classification rate for the sample assuming that the costs of misclassification are equal for the two groups. (See, e.g. Breiman, Friedman, Ohlsen and Stone, 1984).

18. Todd (1995) considers using semiparametric estimates of $P(X)$ and demonstrates that they have little effect on the estimates of bias presented in this paper. Therefore, for computational simplicity, we use the logit model. We use the weighting procedure first suggested by Rao (1965, 1986) and applied by Manski and Lerman (1977) to account for choice-based sampling.

19. Heckman and Roselius (1993) and Heckman, Ichimura, Smith and Todd (1996a, b).

TABLE 1

Comparison groups used in different studies

Study	Ashenfelter (1978)	Ashenfelter and Card (1985)	Dickinson–Johnson–West (1987)	Westat (1986) (Rupp and Bryant)
Programme, year, outcome variable	MDTA Classroom Trainees (First 3 months 1964), 1965–1969 annual social security record earnings	1976 CETA Trainees (1977, 1978 annual social security record earnings), CLMS data	1978 annual social security earnings. CETA trainees enrolled in 1976, CLMS data	2 cohorts of CETA Trainees 1977, 1978 annual social security record earnings, CLMS data
(1) Comparison group in the same labour market?	No	No	No	No
(2) Same questionnaire administered to comparison and treatment group	Yes	Yes	Yes	Yes
(3) Matching criteria (criteria for membership in comparison sample is also called "screening" criteria)	None specified	(a) 1975 earnings $\leqq$ \$20K Household income $\leqq$ \$30K (b) In labour force (March, 1976) Matched on age (persons $\geqq$ 21 used)	(Matching based on a metric over vectors of variables) Matched on predictors of 1978 earnings including lagged earnings (1975–1970), worked in public sector, sex, and demographics. In labour force, March, 1976	Match on 1976 earnings, change in 1976 earnings (1975–1976, 1974–1975) change in earnings, demographics, 1975 labour force status, family income (for 1976–1977 cohort one year previous for 1975–1976 cohort). Either in the labour force, 1975 or at interview March 1976. Three matching groups based on income.
(4) Eligibility for programme known for comparison group members?	No	No	No	No
Variables used in analysis				
Age, race, sex	Yes (No age restriction)	Yes (Age $\geqq$ 21 years old)	Yes (Age 21–65)	Yes (Age 14–60)
Education	No	Yes	Yes	Yes
Training history	No	No	No	No
Children	No	No	No	No
Employment histories	No	No	Yes (recent)	Yes (recent)
Hours worked	No	Yes	Yes	Yes
Unemployment histories	No	No	Yes (recent)	Yes (recent)
On welfare	No	No	Yes	No
Earning histories**	(Annual earnings) 5 years pre-programme 5 years post-programme	(Annual earnings) 2 years pre-programme 2 years post-programme	Same as Ashenfelter and Card (1985)	(Annual earnings) 4 years pre-enrollment earnings histories

** CLMS data matched Social Security Longitudinal Records to March CPS data for 1976 and 1977. The CPS data are for comparison group members. SSA data on longitudinal earnings are available for both groups. All of the personal and family information available in the CPS including short-term employment and labour-force participation histories are available but not necessarily used in the analysis. The CLMS studies all use the social security earnings data.

	NSW (supported work) data		
Study	LaLonde (1986)	Fraker and Maynard (1987) and LaLonde and Maynard (1986)	JTPA data
Programme, year, outcome variable	Annual earnings 1978 annual social security earnings and PSID earnings NSW (Supported Work) Data	1977, 1978, 1979 annual earnings for AFDC recipients and for youth NSW (Supported Work) Data	Quarterly and monthly earnings 1987–1989
(1) Comparison group in the same local labour market	No	No	Yes
(2) Same questionnaire administered to comparison and treatment group?	No	No	Yes
(3) Matching criteria (criteria for membership in comparison sample is also called "screening criteria")	PSID: (a) Men and Women who are household heads 1975–1979 CPS: Matches March 1976 CPS earnings with SSA earnings. Person with 1976 income $\leqq$ \$20K and household income $\leqq$ \$30K	Three Samples (i) Eligible in sample period: for youth: high school dropout–exclude in school youth. For AFDC: age of youngest child, receipt of welfare matching. (ii) Cell matching: based on predictors of 1979 SSA earnings of eligibles: (earnings prior to programme participation), demographics, education, family income, change in earnings. (iii) Stratified matches on imputed 1979 earnings: earnings estimated on eligible nonparticipant sample plus demographic criteria (race, sex). Same criteria for prediction as in (ii).	Persons screened to be eligible for JTPA; out of school youth, no disabled persons; Title IIA only
(4) Eligibility for programme known for comparison group members?	No	No	Yes
Variables used in analysis			
Age, race, sex	Yes: Women AFDC recipients 20–55, Males $\leqq$ 55	Same as LaLonde	Yes
Education	Yes	Same as LaLonde	Yes
Training history	No	Same as LaLonde	Yes
Children	Yes	Same as LaLonde	Yes
Employment histories	No	Same as LaLonde	Yes
Hours worked	No	Same as LaLonde	Yes
Unemployment histories	No	Same as LaLonde	Yes
Welfare receipt?	Yes	Same as LaLonde	Yes
Earnings histories	Two years post-programme Two years pre-programme	Same as LaLonde	(Five years of pre-programme earnings) monthly earnings

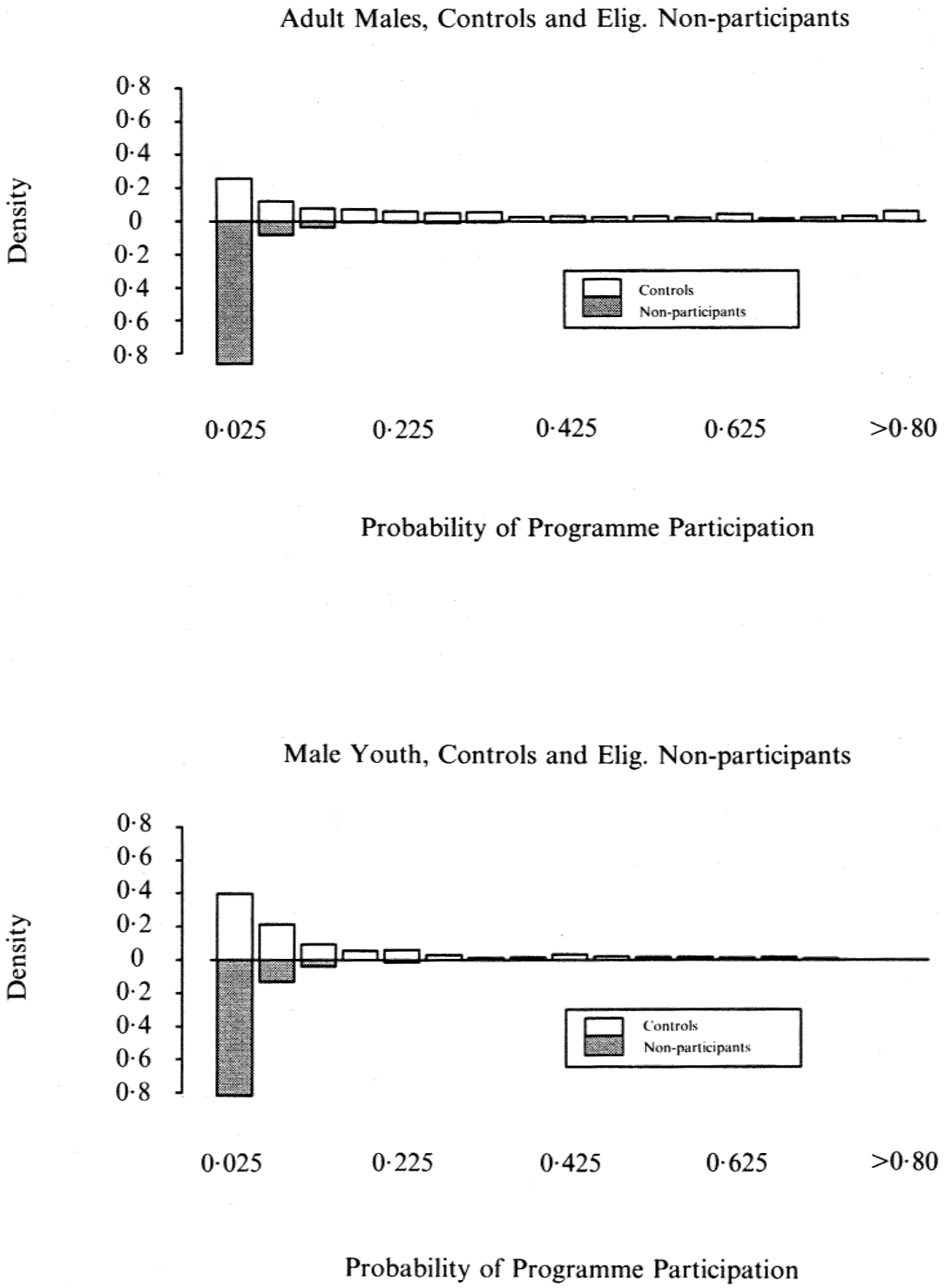
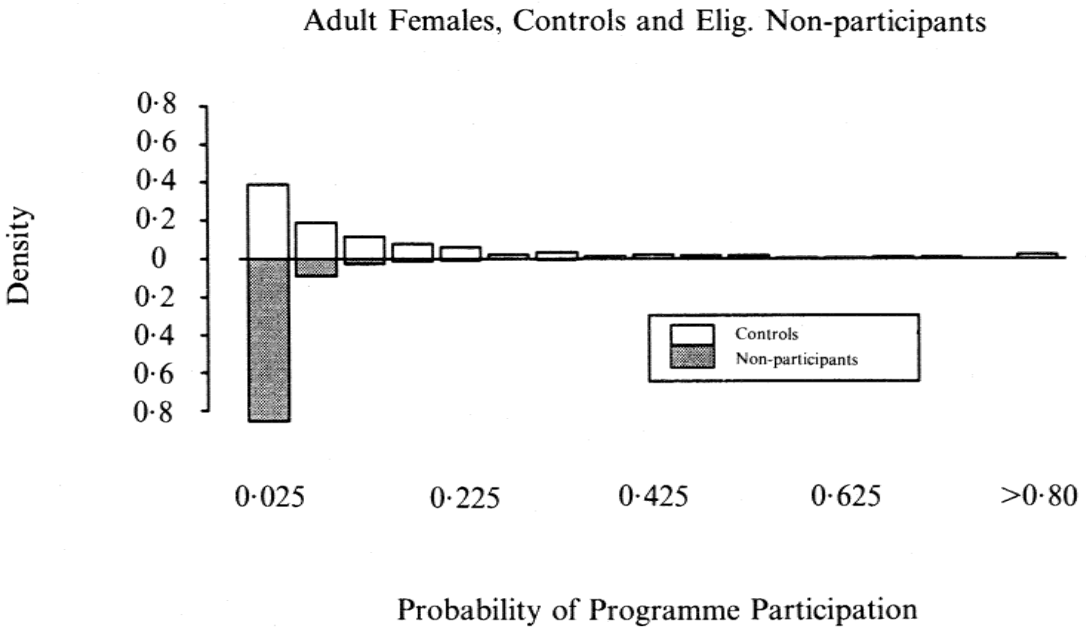


FIGURE 1. Density of Estimated Probability of Programme Participation



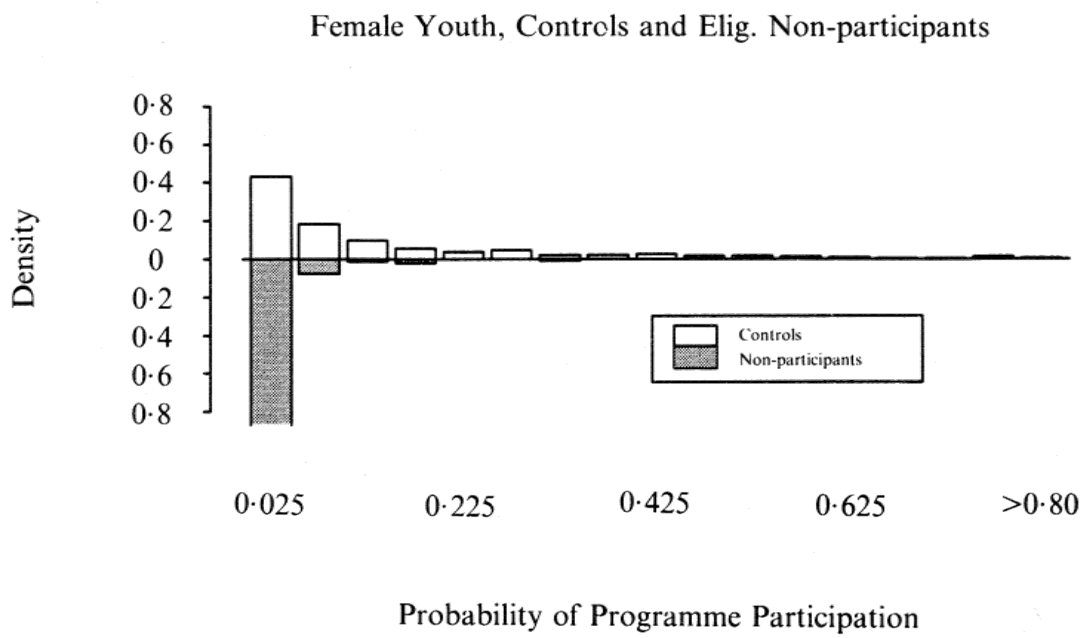


FIGURE 1. —continued.

demonstrate that different supports and differences in the density of P in comparison and control samples are major sources of evaluation bias as conventionally measured.

9. DECOMPOSING THE CONVENTIONAL MEASURE OF EVALUATION BIAS

It is instructive to decompose the conventional measure of evaluation bias into components due to selection bias, rigorously defined, components due to mismatching or misweighting of the data and an additional source of discrepancy between experimental and nonexperimental estimates of programme impact that arises from the differences in supports over which experimental and nonexperimental treatment effects are estimated. Let B be the conventional measure of bias arising from using mean comparison group ($D=0$) outcomes to proxy mean participant ($D=1$) outcomes. It is analogous to (8) but does not condition on X: $B=E(Y_0|D=1)-E(Y_0|D=0)$. To estimate this measure of bias, we use the randomized-out controls from the JTPA experimental data to construct $E(Y_0|D=1)$ combined with data from a nonexperimental comparison group to construct $E(Y_0|D=0)$. We investigate a variety of comparison groups, starting with the ENPs.

S_1 is the support of X for $D=1$, S_0 is the support of X for $D=0$, and S_{10} is the region in which the supports overlap. Denote the region contained in S_0 but not in S_{10} by $S_0 \backslash S_{10}$ and the region contained in S_1 but not in S_{10} by $S_1 \backslash S_{10}$. Bias B can be broken down into components due to differing supports of X, to differing distributions of X over the same support in the two populations and to differences in outcomes that are present even after controlling for observables. Assuming for simplicity that the X are continuously distributed,

$$B=\int_{S_1} E(Y_0|X, D=1) f(X|D=1)dX-\int_{S_0} E(Y_0|X\ D=0) f(X|D=0)dX$$
$$=B_1+B_2+B_3,$$

where

$$B_1=\left\{\int_{S_1\backslash S_{10}} E(Y_0|X, D=1) f(X|D=1)dX-\int_{S_0\backslash S_{10}} E(Y_0|X, D=0) f(X|D=0)dX\right\},$$

$$B_2=\int_{S_{10}} E(Y_0|X, D=0)\{f(X|D=1)-f(X|D=0)\}dX,$$

$$B_3=\int_{S_{10}} \{E(Y_0|X, D=1)-E(Y_0|X, D=0)\} f(X|D=1)dX.$$

The first component of bias (B_1) arises because of nonoverlapping support. For some participants there are no comparable nonparticipants and for some nonparticipants there are no comparable participants. The second component (B_2) arises from different distributions of X within the two populations. The third component (B_3) is due to differences in outcomes that remain even after conditioning on observables and making comparisons on a region of common support. This component is due to selection on unobservables and is selection bias as rigorously defined in econometrics. (See Heckman, Ichimura, Smith and Todd (1996a, b).) The selection bias that results after conditioning on X and weighting

the data comparably is

$$\bar{B}_{S_X}=\frac{\int_{S_{10}}\{E(Y_0|X,D=1)-E(Y_0|X,D=0)\}f(X|D=1)dX}{\int_{S_{10}}f(X|D=1)dX}.$$

Weighting by $f(X|D=1)$ is appropriate since we seek to investigate the bias for $M(S)$ which is defined for the group $D=1$. Observe that $B_3=(\bar{B}_{S_X})\int_{S_{10}}f(X|D=1)dX$.

Exploiting our access to data on randomized-out controls allows us to estimate the bias B by the difference between control and comparison group mean outcomes

$$\hat{B}=\frac{1}{N_1^*}\sum_{i\in I_1^\dagger}Y_{0i}^1-\frac{1}{N_0}\sum_{j\in I_0}Y_{0j}^0,$$

where Y_{0i}^1 is the outcome of randomized-out applicants ($D=1$, $R=0$) and Y_{0j}^0 is the outcome of non-applicants in the comparison group ($D=0$). N_1^* and N_0 are the sample sizes in the two groups. I_1 is the set of indices for eligible and provisionally-accepted applicants ($D=1$), I_1^* is a subset of I_1 corresponding to randomized-out accepted applicants ($D=1$, $R=0$), I_0 is a set of indices for members of the comparison group ($D=0$). Estimating the relative contribution of each of these components to the total bias as conventionally measured, we find that the first and second components contribute the most. The selection bias component B_3 is only a small component of evaluation bias as conventionally measured.

To assess the empirical importance of each of the potential sources of bias, we perform the following decompositions. The components of B are defined conditional on P rather than X. First condition on $P(X)$ to obtain

$$E(Y_0|X,D=1)=E(Y_0|P(X),D=1)+V_1,$$
$$E(Y_0|X,D=0)=E(Y_0|P(X),D=0)+V_0,$$

where V_1 and V_0 are orthogonal to $P(X)$. Denote the values of Y_0 for $D=1$ with associated value P by $Y_0^1(P)$ and the value of Y_0 for $D=0$ by $Y_0^0(P)$. Let $E(Y_0|P,D=0)$ be the local linear regression estimator of $\hat{E}(Y_0|P,D=0)$. (See Fan (1992) or Heckman, Ichimura, Smith and Todd (1996b, c) for detailed discussion of local linear regression methods.) Then the estimator for B is

$$\hat{B}=\frac{1}{N_1^*}\sum_{i\in I_1^\dagger}Y_0^1(P_i)-\frac{1}{N_0}\sum_{j\in I_0}Y_0^0(P_j).$$

The orthogonal components V_1 and V_0 are irrelevant for computing the bias since they average out to zero. Then we may decompose $\hat{B}$ in the following way, where we use the best-predicting $P(X)$ model for each demographic group to define the support

$$\begin{aligned}\hat{B}=&\left[\frac{1}{N_1^*}\sum_{i\in I_1^\dagger(S_1\setminus S_{10})}Y_0^1(P_i)-\frac{1}{N_0}\sum_{j\in I_0(S_0\setminus S_{10})}Y_0^0(P_j)\right]\\&+\left[\frac{1}{N_1^*}\sum_{i\in I_1^\dagger(S_{10})}[\hat{E}(Y_0|P_i,D=0)]-\frac{1}{N_0}\sum_{j\in I_0(S_{10})}Y_0^0(P_j)\right]\\&+\left[\frac{1}{N_1^*}\sum_{i\in I_1^\dagger(S_{10})}[Y_0^1(P_i)-\hat{E}(Y_0|P_i,D=0)]\right],\end{aligned}$$

TABLE 2

Decomposition of difference in post-programme mean earnings
Bootstrapped standard errors shown in parentheses‡
Percentage of mean difference attributable to components in square brackets
Earnings measured in average monthly dollars

	Mean difference $\hat{B}$	Non-overlap* $\hat{B}_1$	Density weighting $\hat{B}_2$	Selection bias $\hat{B}_3$	Average bias $(\hat{\bar{B}}_{S_P})$	Selection bias** $(\hat{\bar{B}}_{S_P})$ as a % of treatment impact
Experimental Controls and eligible nonparticipants (ENPs)†						
Adult males	−342	218	−584	23	38	87%
(std. err.) [%]	(47)	(38) [−64%]	(41) [170%]	(33) [−7%]	(63)	
Adult females	33	80	−78	31	38	129%
(std. err.) [%]	(26)	(13) [242%]	(18) [−235%]	(26) [94%]	(33)	
Male youth	20	142	−131	9	14	23%
(std. err.) [%]	(57)	(28) [704%]	(35) [−650%]	(42) [46%]	(64)	
Female youth	42	74	−67	35	49	7239%
(std. err.) [%]	(36)	(17) [177%]	(26) [−161%]	(28) [84%]	(42)	
Experimental Controls and SIPP Eligibles††						
	Mean difference $\hat{B}$	Non-overlap* $\hat{B}_1$	Density weighting $\hat{B}_2$	Selection bias $\hat{B}_3$	Average bias $(\hat{\bar{B}}_{S_P})$	Selection bias** $(\hat{\bar{B}}_{S_P})$ as a % of treatment impact
Adult males	−145	151	−417	121	192	440%
(std. err.) [%]	(56)	(30) [−104%]	(44) [287%]	(33) [−83%]	(57)	
Adult females	47	97	−172	122	198	676%
(std. err.) [%]	(23)	(19) [206%]	(16) [−367%]	(15) [260%]	(26)	
Male youth	−188	65	−263	9	21	36%
(std. err.) [%]	(106)	(108) [−35%]	(53) (139%]	(25) [−5%]	(90)	
Female youth	−88	83	−168	−3	−13	1969%
(std. err.) [%]	(38)	(22) [−94%]	(27) [191%]	(10) [3%]	(58)	

‡ They are based on 50 replications of the data with 100% sampling.

† The best predictor Control-ENP participation models for all the demographic groups include indicator variables for site, age, race, education, marital status, children less than 6 and labour force transitions. In addition to these variables, the adult male model also includes an indicator for vocational training history, the number of household members, earnings in the month of random assignment or eligibility determination (RA or EL) and number of jobs held in 18 months before RA or EL. The adult female model includes an indicator for recent schooling, earnings in the month of RA or EL and number of labour force transitions in the 24 months prior to RA or EL. The male youth model includes average earnings in the 6 months and 12 months prior to RA or EL and average positive earnings in the 6 months before RA or EL. The female youth model includes earnings in the 12 months before RA or EL.

†† The best predictor Control-SIPP model includes indicators for age, race, education, marital status, children age less than 6, labour force transition patterns and levels of earnings in the preceding year. The data used are SIPP eligibles and experimental JTPA controls.

* A 2% trimming rule was used for adult males and females and a 5% trimming rule for youth was used in determining the overlapping support region (see Appendix C for a description of how the support is determined). The proportion of Controls and ENPs falling in the overlap region (S_P) are: 60% and 96% of adult males, 82% and 96% of adult females, 67% and 92% of male youth and 71% and 93% of female youth. The proportion of SIPPs and Controls falling in the overlap region are: 63% and 96% of adult males, 61% and 96% of adult females, 41% and 90% of male youth and 20% and 89% of female youth. A 0·06 fixed bandwidth and a biweight kernel, defined in Appendix A, were used for the nonparametric estimates.

** The final column displays the ratio of the absolute value of $(\hat{\bar{B}}_{S_P})$ to the absolute value of experimental impact estimate.

TABLE 2—continued

Decomposition of difference in post-programme mean earnings
Bootstrapped standard errors shown in parentheses‡
Percentage of mean difference attributable to components in square brackets
Earnings measured in average monthly dollars

Experimental Controls and No-shows†						
	Mean difference $\hat{B}$	Non-overlap* $\hat{B}_1$	Density weighting $\hat{B}_2$	Selection bias $\hat{B}_{23}$	Average bias $(\hat{\bar{B}}_{S_P})$	Selection bias** $(\hat{\bar{B}}_{S_P})$ as a % of treatment impact
Adult males	29	−13	3	38	42	97%
(std. err.) [%]	(38)	(12) [−45%]	(16) [11%]	(37) [135%]	(40)	
Adult females	9	1	−9	18	20	68%
(std. err.) [%]	(23)	(6) [9%]	(10) [−99%]	(26) [190%]	(29)	
Male youth	84	14	−21	91	99	171%
(std. err.) [%]	(31)	(9) [17%]	(11) [−25%]	(31) [108%]	(34)	
Female youth	18	3	−31	46	51	7441%
(std. err.) [%]	(24)	(16) [17%]	(16) [−170%]	(24) [254%]	(29)	

‡ They are based on 50 replications of the data with 100% sampling.
† The data used are experimental JTPA controls and experimental persons assigned to treatment who enrolled in JTPA but dropped out before receiving services. The No-show predictor model for all demographic groups includes indicator variables for site, race and recommended training services. For adult males, the model also includes earnings last year, earnings squared, an indicator for preferred language Spanish and for preferred language other than Spanish or English. For adult women, the model includes indicators for last employed 0–6 months before random assignment (RA) and for 7–12 months before RA and an indicator for whether enrollment in JTPA was required.
* A 2% trimming rule was used for adult males and females and a 5% trimming rule for youth was used in determining the overlapping support region (see Appendix C for a description of how the support is determined). The proportion of controls and no-shows falling in the overlap region (S_P) are: 91% and 91% of adult males, 90% and 92% of adult females, 91% and 93% of male youth, and 90% and 89% of female youth. A 0·06 fixed bandwidth and a biweight kernel, defined in Appendix A, were used for the nonparametric estimates.
** The final column displays the ratio of the absolute value of $\hat{\bar{B}}_{S_P}$ to the absolute value of experimental impact estimate.

where $I_1^*(S_1 \setminus S_{10})$ is the set of indices in I_1^* with associated P values not in S_{10}, $I_0(S_0 \setminus S_{10})$ is the set of indices in I_0 with associated P values not in S_{10}, $I_1^*(S_{10})$ and $I_0(S_{10})$ are, respectively, the associated indices in I_1^* and I_0 with P values in S_{10}. The three terms in brackets in this decomposition are the sample analogues of the three components of B defined above. The sums are self-weighted by the $f(P|D)$ density because they average over either $D=1$ or $D=0$ values of $P(X)$.

The first panel of Table 2 reports the raw mean difference in outcomes between ENP comparison group members and controls (B), measured by monthly earnings in our study, and the contribution of each component to the overall mean bias. Y_0 corresponds to average monthly earnings over the eighteen months following random assignment or eligibility determination. For all four demographic groups, non-overlapping support and different density weighting of the P contribute most to the total bias.[20] After accounting for these two sources, selection bias B_3 is statistically insignificantly different from zero. This evidence suggests that evaluation methods based on the hypothesis of selection on observables eliminate the largest sources of bias, provided the evaluation parameter is estimated over the region of common support. The fifth column of numbers in the table gives the estimated average selection bias $\hat{\bar{B}}_{S_P}$ over the region of overlapping support. The

20. The table footnote gives the proportion of individuals in the overlap region.

last column shows that although the bias is often small in relation to the simple mean difference, it still represents a significant fraction of the experimentally-estimated treatment impact.

Matching methods eliminate two of the three sources of bias. The bias due to nonoverlapping supports is eliminated by matching only over the region of common support. The bias due to different density weighting is eliminated because matching methods effectively reweight the nonparticipant data to equate the distribution of P. Only the bias due to differences in unobservables across groups is not eliminated.

In results reported in Heckman, Ichimura, Smith and Todd (1996a, b), we decompose the bias for each quarter following the date of random assignment or eligibility determination and show that the percentage of the total difference in quarterly means that is due to selection bias component B_3 is small for adult men, adult women and male youth relative to the other components. For female youth, it is of comparable magnitude to that of other components but all components are small in absolute value. In Heckman, Ichimura, Smith and Todd (1996a, b, c), we find that even though the average bias $\bar{B}_{S_p}$ is small, the pointwise bias $B(P)$ is not small for all four demographic groups. This evidence is contrary to what would be predicted from matching conditions (A-1), (A-3) or (A-4). It is not necessarily inconsistent with difference-in-differences assumptions (A-5) or (A-5′).

10. VERIFYING THE ASSUMPTIONS THAT JUSTIFY MATCHING AND OUR EXTENSIONS OF IT

In this section, we apply methods developed in Heckman, Ichimura, Smith and Todd (1996b, c) to the JTPA data to test identifying conditions (A-3), (A-4), (A-3′), (A-4′), (A-5) and (A-5′). Assumptions (A-3) and (A-3′) are decisively rejected in our data. We find some support for the weaker assumption (A-4′) for one group—young males—but it is generally rejected for other groups. The weaker assumptions (A-5) and (A-5′) that justify our conditional difference-in-differences method are not rejected for any group.

Specifically, we test the following hypotheses, where the subscript denotes the identifying assumption being tested:

Conditional Independence:

$$\mathrm{H}_{(\text{A-3})}: F(y_0|P, D=1) = F(y_0|P, D=0);$$

$$\mathrm{H}_{(\text{A-3}')}: F(u_0|P, D=1) = F(u_0|P, D=0);$$

and

Mean Independence:

$$\mathrm{H}_{(\text{A-4})}: E(Y_0|P, D=1) = E(Y_0|P, D=0);$$

$$\mathrm{H}_{(\text{A-4}')}: E(U_0|P, D=1) = E(U_0|P, D=0);$$

and

Differences in Differences Mean Independence:

$$\mathrm{H}_{(\text{A-5})}: E(Y_{0t} - Y_{0t'}|P, D=1) = E(Y_{0t} - Y_{0t'}|P, D=0);$$

$$\mathrm{H}_{(\text{A-5}')}: E(U_{0t} - U_{0t'}|P, D=1) = E(U_{0t} - U_{0t'}|P, D=0),$$

where t is a post programme period and t' is a preprogramme period.

10.1. *Test statistics*

To test $H_{(A\text{-}3)}$ at a point $Y_0 = y_0$ conditional on $P = p$, we estimate the conditional c.d.f. $\hat{F}_d(p) = \hat{F}(y_0 | P = p, D = d)$ using local linear regression. For a given y_0 we run a nonparametric regression of the indicator variable $1(Y \leqq y_0)$ on P and evaluate the function at $P = p$. $\hat{F}_d(p) - F_d(p)$ is asymptotically normally distributed, $N(B_d, V_d)$, where B_d and V_d are defined in Appendix A. Assuming that the same kernel and bandwidths are used to estimate F_1 and F_0, under the null $H_{(A\text{-}3)}$,

$$(\hat{F}_1(p) - \hat{F}_0(p))'(\hat{V}_1 + \hat{V}_0)^{-1}(\hat{F}_1(p) - \hat{F}_0(p)) \sim \chi^2(1),$$

where $\hat{V}_1$ and $\hat{V}_0$ are consistent estimators of V_1 and V_0 that are presented in Appendix A.

To test conditional mean independence ($H_{(A\text{-}4)}$), we use local linear regression estimates of $\hat{m}_1(p) = \hat{E}(y_0 | P = p, D = 1)$ and $\hat{m}_0(p) = \hat{E}(y_0 | P = p, D = 0)$ and the fact that $(\hat{m}_d(p) - m_d(p)) \sim N(\tilde{B}_d, \tilde{V}_d)$ where $\tilde{B}_d$ and $\tilde{V}_d$ are defined in Appendix A. Assuming the same kernel and bandwidths are used to estimate m_0 and m_1, under the null

$$(\hat{m}_1(p) - \hat{m}_0(p))'(\tilde{V}_1 + \tilde{V}_0)^{-1}(\hat{m}_1(p) - \hat{m}_0(p)) \sim \chi^2_1,$$

where $\tilde{V}_1$ and $\tilde{V}_0$ are feasible covariance estimators for $\hat{m}_0$ and $\hat{m}_1$ defined in Appendix A.

An important advantage of using local linear regression methods instead of ordinary kernel regression methods in implementing these tests is that, under the two nulls, the bias terms for F_1 and F_0, (B_1, B_0), and for m_1 and m_0, $(\tilde{B}_1, \tilde{B}_0)$, cancel in the test statistics when the same bandwidth is used to estimate $(\hat{F}_1(p), \hat{F}_0(p))$ and $(\hat{m}_1(p), \hat{m}_0(p))$ respectively. With ordinary kernel methods, the bias depends on the data density $f_d(p)$, so the bias terms do not cancel and the test statistic has a noncentral chi-squared distribution. (Heckman, Ichimura, Smith and Todd (1996b, c).)

Tests of $H_{(A\text{-}3')}$ and $H_{(A\text{-}4')}$ replace Y_0 by the corresponding U_0 residuals. Appendix A defines the test statistic for the second hypothesis, $H_{(A\text{-}4')}$, gives estimators of the variances needed to perform the tests and generalizes the test statistics to apply to panel data. Tests of $H_{(A\text{-}5)}$ and $H_{(A\text{-}5')}$ are constructed on similar principles. The test for $H_{(A\text{-}5)}$ is just a version of the test for $H_{(A\text{-}4)}$ using differences in the conditional means as input to a $\chi^2(1)$ test. Residuals replace conditional means of outcomes in the test of $H_{(A\text{-}5')}$. Details on how to construct these tests are given in Appendix A.

10.2. *Empirical evidence on the validity of matching assumptions*

In tables available on request from the authors, we test the conventional identifying assumption $H_{(A\text{-}3)}$ and the corresponding assumption for population disturbances $H_{(A\text{-}3')}$. The test is performed at selected quantiles of the outcome distribution, at different values of P, and overall. Separate tests are conducted for pre-programme and post-random assignment time periods. For adult men, $H_{(A\text{-}3)}$ is decisively rejected. For adult women $H_{(A\text{-}3)}$ is rejected at some quantiles of P. For male youth, $H_{(A\text{-}3)}$ is not rejected, while for female youth there are a few rejections of the hypothesis. In tests of $H_{(A\text{-}3')}$, we reject the hypothesis for adult men and women and for female youth but not for male youth.

Table 3(a) presents tests results for the weaker hypothesis of conditional mean independence of Y_0 ($H_{(A\text{-}4)}$) and U_0 ($H_{(A\text{-}4')}$) that still justify matching to recover means. Since conditional independence implies mean independence and mean independence suffices for our parameter of interest, this is the central hypothesis for testing the validity of matching in our samples. For all four demographic groups, in the pre-programme period we reject

TABLE 3(a)

P-values from tests for conditional mean independence
$(H_{(A\text{-}4)})$: $E(Y_0 | D=1, P) = E(Y_0 | D=0, P)$
$(H_{(A\text{-}4')})$: $E(U_0 | D=1, P) = E(U_0 | D=0, P)$

	Adult men				Adult women			
	Earnings $(H_{(A\text{-}4)})$		Residuals $(H_{(A\text{-}4')})$		Earnings $(H_{(A\text{-}4)})$		Residuals $(H_{(A\text{-}4')})$	
P-points	Pre-programme quarters†	Post-programme quarters†	Pre-programme quarters†	Post-programme quarters†	Pre-programme quarters†	Post-programme quarters†	Pre-programme quarters†	Post-programme quarters†
0·0025	0·0242	0·0000	0·0293	0·0002	0·0751	0·0029	0·1472	0·0363
0·005	0·0803	0·0002	0·0815	0·0004	0·1671	0·0057	0·1847	0·0298
0·01	0·2416	0·0042	0·2586	0·0040	0·3963	0·0104	0·3035	0·0140
0·02	0·1919	0·1224	0·4078	0·2056	0·3609	0·0172	0·3836	0·0180
0·03	0·1238	0·4363	0·5680	0·6563	0·0577	0·0677	0·1048	0·1243
0·04	0·1585	0·7659	0·7177	0·9060	0·0145	0·1418	0·0790	0·2570
0·05	0·3271	0·9423	0·8064	0·9885	0·0402	0·2052	0·2718	0·4510
0·10	0·8464	0·1678	0·7456	0·2591	0·0702	0·4406	0·4752	0·8423
Overall	0·0159	0·0000	0·2515	0·0001	0·0006	0·0000	0·0375	0·0008
	Male youth				Female youth			
	Earnings $(H_{(A\text{-}4)})$		Residuals $(H_{(A\text{-}4')})$		Earnings $(H_{(A\text{-}4)})$		Residuals $(H_{(A\text{-}4')})$	
P-points	Pre-programme quarters†	Post-programme quarters†	Pre-programme quarters†	Post-programme quarters†	Pre-programme quarters†	Post-programme quarters†	Pre-programme quarters†	Post-programme quarters†
0·0025	0·2277	0·0333	0·3877	0·2142	0·1323	0·0076	0·1418	0·3047
0·005	0·2964	0·0997	0·3547	0·2911	0·0906	0·0166	0·0770	0·3648
0·01	0·3966	0·3047	0·3373	0·4894	0·0490	0·0697	0·0274	0·5512
0·02	0·5400	0·3426	0·4702	0·6166	0·0273	0·1853	0·0107	0·4803
0·03	0·7053	0·1845	0·6913	0·5214	0·0540	0·0621	0·0277	0·1570
0·04	0·7095	0·1239	0·7342	0·4567	0·1948	0·0290	0·2278	0·0575
0·05	0·4260	0·1329	0·5988	0·4315	0·4551	0·0236	0·6279	0·0245
0·10	0·0000	0·9124	0·0000	0·1944	0·1186	0·0441	0·2412	0·0000
Overall	0·0147	0·0251	0·0015	0·3561	0·0009	0·0000	0·0006	0·0001

* A fixed bandwidth of 0·06 and a biweight kernel, defined in Appendix A, are used in the test. The models for the probability of participation, *P*, are described in the footnotes to Table 2.
† Tests are performed jointly across pre-programme quarters $t=-1$ to $t=-6$ or post-programme quarters $t=1$ to $t=6$.

TABLE 3(b)

P-values from tests for difference-in-differences†
$(H_{(A\text{-}5)})$: $E(Y_{0t} - Y_{0t'} | D=1, P) = E(Y_{0t} - Y_{0t'} | D=0, P)$
$(H_{(A\text{-}5')})$: $E(U_{0t} - U_{0t'} | D=1, P) = E(U_{0t} - U_{0t'} | D=0, P)$
(Null tested jointly over $t \in \{1, 2, 3, 4, 5, 6\}$)

	Adult males		Adult females		Male youth		Female youth	
P-points	Earnings	Residuals	Earnings	Residuals	Earnings	Residuals	Earnings	Residuals
0·0025	0·0782	0·0937	0·4139	0·2340	0·7178	0·7079	0·0320	0·0707
0·005	0·0885	0·0966	0·2924	0·1454	0·7255	0·7268	0·0269	0·0538
0·01	0·1377	0·1184	0·1195	0·0482	0·6878	0·7404	0·0377	0·0596
0·02	0·3946	0·2865	0·0865	0·0454	0·6594	0·7968	0·1702	0·2420
0·03	0·7149	0·6007	0·4690	0·3661	0·6548	0·8426	0·4669	0·6340
0·04	0·9167	0·8745	0·7672	0·6089	0·6836	0·8617	0·7446	0·8577
0·05	0·9727	0·9506	0·6091	0·4626	0·7717	0·8565	0·6632	0·7594
0·10	0·9677	0·9833	0·7386	0·5098	0·8346	0·4245	0·9813	0·9786
Overall	0·5260	0·4403	0·3364	0·0506	0·9832	0·9910	0·0533	0·2119

† A fixed bandwidth equal to 0·06 and a biweight kernel, defined in Appendix A, are used. The models for the probability of participation, *P*, are those given in the footnote to Table 2. The test is for symmetric differences around $t=0$, the date of enrollment into the programme, $t=-t'$.

both nulls at conventional significance levels (5%). We also reject both hypotheses in the post-programme period for all demographic groups except for young males.[21]

Note that tests of mean independence of the residuals conducted on pre-programme earnings do not always produce the same inference as tests conducted on post-programme earnings. See especially the results for adult males or young males in Table 3(a). It is, therefore, not a safe strategy to use pre-programme tests about mean selection bias to make inferences about post-programme selection bias, as proposed by Heckman and Hotz (1989).

Table 3(b) presents test results for the weaker hypothesis of conditional mean independence in the pre- and post-programme differences, hypotheses ($H_{(A\text{-}5)}$) and the hypothesis $H_{(A\text{-}5')}$). These hypotheses are formulated in terms of symmetric differences around the date of enrollment into the programme, $t=0$, so $t=-t'$ where t' is a pre-programme period and t is a post-programme period. We do not reject these null hypotheses for any demographic group suggesting that the semiparametric conditional method of difference-in-differences proposed in this paper is consistent with the data.

In summary, we test and reject the conditional independence assumptions (A-3) maintained in the literature and the weaker mean independence assumption (A-4). The hypothesis of mean independence of U_0 is somewhat more concordant with the data than mean independence of the raw data (Y_0) but it is still rejected in most instances. Finally we test, and do not reject, the identifying assumptions for the conditional difference-in-differences estimator for symmetric differences around the date of enrollment into the programme.

11. ALTERNATIVE MATCHING ESTIMATORS

All matching estimators, including our semiparametric conditional difference-in-differences estimator, can be cast in the following framework

$$\hat{M}(S)=\sum_{i\in I_1}\omega_{N_0,N_1}(i)[Q_{1i}-\sum_{j\in I_0}W_{N_0,N_1}(i,j)Q_{0j}], \qquad \text{for } X\in S, \tag{10}$$

where Q_{1i} is a treatment outcome and Q_{0j} is a comparison group outcome. Q_{1i} and Q_{0j} sometimes represent Y_{1i} and Y_{0j}, respectively, but the notation is more general to allow us to use regression-adjusted Y_{1i} and Y_{0j}. N_1 is the number of programme participants, N_0 is the number of persons in the comparison group, $W_{N_0,N_1}(i,j)$ is a weight with $\sum_{j\in I_0}W_{N_0,N_1}(i,j)=1$, and $\omega_{N_0,N_1}(i)$ is a weight that accounts for heteroscedasticity and scale. I_1 is the set of indices for programme participants and I_0 a set of indices for comparison group members. Matches for each participant are constructed by taking weighted averages over comparison group members. Matching can be performed within various strata to recover estimates for different populations of interest.

Matching estimators differ in the weights they attach to members of the comparison group. Define a neighbourhood $C(X_i)$ for person i in the participant sample, $i\in I_1$. Neighbours for i are persons $j\in I_0$ for whom $X_j\in C(X_i)$. Persons matched to i are those people in set A_i where $A_i=\{j\in I_0 \mid X_j\in C(X_i)\}$.

A nearest-neighbour matching estimator sets $Q_{1i}=Y_{1i}$, $Q_{0j}=Y_{0j}$, $\omega_{N_0,N_1}(i)=1/N_1$ and for each i in the $D=1$ sample picks the match $C(X_i)=\min_j \|X_i-X_j\|$, $j\in I_0$ where $\|\cdot\|$ is

21. We raise one cautionary methodological note. The asymptotic theory suggests that estimation of the parameters of P and of the parameters of the outcome equation (β) is irrelevant for construction of the test statistics. This phenomenon arises because the parameters of these functions converge at a faster rate than the nonparametric functions that we use in the tests. Yet in a small scale Monte Carlo study we find that inferences from the residuals of fitted models are sensitive to parameter estimation. This problem does not affect the tests based on unadjusted earnings, only those based on residuals. See the discussion in Appendix B.

a norm. A_i is a singleton set, except for ties that are broken by a random draw. The weighting scheme for the nearest neighbour estimator assigns all the weight to the single match: $W_{N_0,N_1}(i,j)=1$ if $j\in A_i$; $W_{N_0,N_1}(i,j)=0$ otherwise. Two versions of this method are (a) X_j may be reused for other matches (sampling with replacement); (b) X_j may not be reused (sampling without replacement).

The distance between persons i and j can be substantial if $C(X_i)$ is not restricted. Caliper matching avoids the problem of a substantial gap between X_i and X_j. (Cochrane and Rubin (1973).) Matches are made to i only if

$$\|X_i-X_j\|<\varepsilon, \qquad j\in I_0,$$

where ε is a pre-specified tolerance. Otherwise no match is made and person i is left unmatched. The neighbourhood is $C(X_i)=\{X_j|\,\|X_i-X_j\|<\varepsilon\}$. If more than one person is in A_i, then the nearest neighbour under norm $\|\cdot\|$ is selected. A variant of caliper matching selects one metric to caliper match and another to pick among elements in the control sample if more than one observation qualifies under the caliper criterion. The methods can be applied with or without reuse of the observations in the comparison group. The Mahalanobis metric is commonly used in the matching literature: $\|X_i-X_j\|=(X_i-X_j)'\sum^{-1}(X_i-X_j)$ where $\sum$ is the covariance matrix formed from the $D=1$ sample.

The kernel-based matching estimators that we use and whose asymptotic distribution is derived in our companion paper (Heckman, Ichimura and Todd (1993 revised, 1997)), construct matches by forming weighted averages of the outcomes of all individuals in the $D=0$ comparison sample. If weights from a symmetric, nonnegative, unimodal kernel are used, then the average places higher weight on persons close in terms of X_i and lower weight on more distant observations. Kernel matching sets $A_i=I_0$ and defines

$$W_{N_0,N_1}(i,j)=\frac{G_{ij}}{\sum_{k\in I_0}G_{ik}},$$

where $G_{ik}=G((X_i-X_k)/a_{N_0})$ is a kernel function and a_{N_0} is a bandwidth parameter.[22] Kernel matching is a local averaging method that reuses and weights all the comparison group observations in the treatment sample.

Local linear matching uses the weight

$$W_{N_0,N_1}(i,j)=\frac{G_{ij}\sum_{k\in I_0}G_{ik}(X_k-X_i)^2-[G_{ij}(X_j-X_i)][\sum_{k\in I_0}G_{ik}(X_k-X_i)]}{\sum_{j\in I_0}G_{ij}\sum_{k\in I_0}G_{ik}(X_k-X_i)^2-\left(\sum_{k\in I_0}G_{ik}(X_k-X_i)\right)^2}. \tag{11}$$

We use local linear weights instead of more conventional kernel weights because local linear estimators converge at a faster rate at boundary points and adapt better to different data densities. A substantial fraction of our data is at boundaries. See Figure 1.[23]

The regression-adjusted local linear matching estimator developed in Heckman, Ichimura, Smith and Todd (1994, revised 1996b) and in Heckman, Ichimura and Todd (1993, revised 1997) combines local linear matching on the probability of participation $P(X)$ and regression adjustment on the X. It extends classical matching methods by utilizing information on the functional form of outcome equations and by incorporating exclusion

22. a_{N_0} satisfies $\lim_{N_0\to\infty}a_{N_0}\to 0$. Precise conditions on the rate of convergence needed for consistency and asymptotic normality of the kernel matching estimator are presented in our companion paper. (Heckman, Ichimura and Todd (1993, 1997).)

23. These properties of the local linear estimator that make it superior to the standard kernel regression estimator are discussed in Fan (1992).

restrictions across outcome and participation equations. Previously-developed matching methods do not use these sources of information.

Regression-adjusted matching is performed by the following procedure. Assume a conventional econometric model for outcomes in the no-treatment state that is additively separable in observables and unobservables: $Y_0 = X\beta_0 + U_0$. Using partially linear regression methods applied to the comparison group sample, estimate the components of $E(Y_0 | X, D=0) = X\beta_0 + E(U_0 | X, D=0)$ imposing any desired exclusion restrictions.[24] To estimate $M(S)$ remove $X\hat{\beta}_0$ from both Y_0 and Y_1. In the framework of equation (10), $Q_{1i} = (Y_{1i} - X_i\hat{\beta}_0)$, $Q_{0j} = (Y_{0j} - X_j\hat{\beta}_0)$ and local linear regression weight (11) is used. Our companion paper (Theorem 2), presents a proof of consistency and asymptotic normality of the matching estimator under Assumption (A-4′). This estimator is a compromise between a fully nonparametric approach, which is not likely to yield reliable estimates in samples at our disposal, and a more parametric model.[25]

The conditional difference-in-difference matching estimator introduced in Section 4 sets $Q_{1i} = (Y_{1it} - Y_{0it'})$ and $Q_{0j} = (Y_{0jt} - Y_{0jt'})$ in equation (10) or equivalently uses

$$\hat{D}_{t,t'}(S) = \hat{M}_t(S) - \hat{M}_{t'}(S),$$

where $M_t(S)$ and $M_{t'}(S)$ are equal to $M(S)$ defined in (3), except that Y_{1j}, Y_{0j}, I_1 and I_0 now are subscripted by t or t'. A conditional difference-in-differences matching estimator that performs semi-parametric regression-adjustment on X is defined over set S for $Q_{1i} = [(Y_{1it} - X_{it}\hat{\beta}_{0t}) - (Y_{0it'} - X_{it'}\hat{\beta}_{0t})]$ and $Q_{0j} = [(Y_{0jt} - X_{jt}\hat{\beta}_{0t}) - (Y_{0jt'} - X_{jt'}\hat{\beta}_{0t})]$ or equivalently

$$\tilde{D}_{t,t'}(S) = \tilde{M}_t(S) - \tilde{M}_{t'}(S).$$

Regression-adjusted conditional difference-in-differences matching is an effective method in reducing bias in our data. However, it is more demanding in terms of its data requirements than the cross-sectional matching estimators because it requires pre-programme data.

Finally, note that randomization is a special case of the matching estimator (10) for which the comparison group is an experimental control group, $Q_{1i} = Y_{1i}$ and $Q_{0j} = Y_{0j}$, $\omega_{N_0,N_1}(i) = 1/N_1$, $W_{N_0,N_1}(i,j) = 1/N_0$ and the entire sample of controls is used as a comparison group for each experimental observation.

12. EVALUATING THE PERFORMANCE OF DIFFERENT MATCHING ESTIMATORS

We now consider the performance of widely-used statistical matching methods for estimating parameter $M(S)$. We contrast the performance of conventional methods with those of the new methods presented here and find that our new procedures—especially the conditional differences-in-differences method—are generally more effective. Nonetheless, matching is no panacea and considerable bias relative to the size of the experimentally-determined treatment effect is found for all methods. The evidence is generally more favourable for the conditional differences-in-differences estimator.

24. Estimation of the partially linear model, when local linear regression methods are used to estimate $E(U_0 | X, D=0)$, is discussed in Heckman, Ichimura, Smith and Todd (1994, revised 1996b, c).

25. Rubin (1979) presents a regression-matching procedure in which regressions are first fit on participant and control samples and then residuals from the regression are matched. He does not formally justify his estimator and it is easy to demonstrate that it is inconsistent for $E(\Delta | X, D=1)$ unless $E(U_0 | X, D=0)$ is linear in X, and does not depend on D. We estimate parametric β and the nonparametric component $E(U_0 | X, D=0)$ jointly using the partially linear model. See Heckman, Ichimura, Smith and Todd (1996b, c).

The following matching procedures are evaluated in this paper.

1. *Simple P Nearest Neighbour Matching*: Match using the n neighbours with the closest values of P. $n=1$ defines nearest-neighbour matching which is the conventional estimator. In tables available on request, we also consider versions of nearest-neighbour estimators that average the outcomes of the nearest 5 or 10 neighbours.
2. *Local Linear P Score Matching*: Form a weighted average over the outcomes of comparison group members using local linear regression weights. A bandwidth equal to 0·06 is used in estimation of this model and all models presented in this paper.[26] The biweight kernel we use throughout this paper is defined in Appendix A.
3. *Regression-Adjusted Local Linear Matching*: Using X-adjusted outcomes for participants, $Y_{1t}-X_{1t}\hat{\beta}_{0t}$, we match to corresponding adjusted outcomes for nonparticipants, $Y_{0t}-X_{0t}\hat{\beta}_{0t}$ using local linear regression weights.
4. *Conditional Difference-in-Differences Matching*: Using either unadjusted or X-adjusted outcomes for participants, $Y_{1t}-X_{1t}\hat{\beta}_{0t}$, we match to corresponding unadjusted or adjusted outcomes for nonparticipants, $Y_{0t}-X_{0t}\hat{\beta}_{0t}$, at a post-programme time period t and at a pre-programme time period t'. Local linear regression weights are used to construct the matched outcomes.

In results available on request, we also investigate both smoothed and raw Mahalanobis metric estimators. They are comparable in performance to the other matching estimators and for the sake of brevity we delete discussion of them. To learn about the bias arising from the use of the different methods, we compare the outcomes of members of an experimental control group ($D=1, R=0$) to the matched outcomes from several candidate comparison groups ($D=0$). Since neither group receives treatment, the discrepancy between the mean control outcomes and the mean of the matched outcomes is a measure of bias for the particular estimator being investigated.

12.1. *The support problem*

Several sources of bias plague nonexperimental estimators: (a) selection on unobservables, (b) failure of a common support condition and (c) failure to weight treatment and comparison group members comparably. Matching estimators are not subject to bias (c) because they effectively reweight comparison group data to equate the distribution of observables in the $D=1$ and $D=0$ samples. Matching estimators may be subject to bias due to (a) or (b). Source (b) can be eliminated if matching is performed only over regions of common support. However, this requires that the parameter of interest $M(S)$ be redefined as the mean impact over the common support region, $M(S_{10})$. Thus, two different parameters are implicitly defined when experimental and nonexperimental analyses are used to evaluate the same programme: $M(S_E)$ and $M(S_{10})$, where S_E is the support in the experimental data and S_{10} is the region of common support or overlap between comparison groups and participant or control groups. A nonexperimental estimator of $M(S_{10})$ does not estimate the same parameter as is estimated by experimental methods unless $S_E=S_{10}$. In our study, the density and empirical support for P are very different for participants ($D=1$) and for

26. For symmetric, nonnegative, unimodel kernels, observations with closer P scores receive higher weight. Other bandwidths were also tried: 0·04, 0·08 and an "optimal" plug-in bandwidth and produced essentially the same results as the value for 0·06 for all groups. The 0·06 is the average optimal bandwidth which produces smoother fits than the "optimal" bandwidth. The "optimal" plug-in bandwidth is derived in our companion paper (Heckman, Ichimura and Todd (1993, 1997)).

TABLE 4

Estimated programme impacts
Monthly average impacts in dollars over 18 *months after random assignment**

	Experimental impact $M(S_E)$	Experimental impact for persons in overlap support region $M(S_P)$	Bias from non-overlap	% bias
Adult males	44	61	17	39
	(17)	(17)		
Adult females	29	35	6	21
	(9)	(12)		
Male youth	−58	−36	22	38
	(14)	(18)		
Female youth	−1	25	26	2500
	(11)	(18)		

* In our data, the experimental control group was administered a long-baseline survey that gathered five years of retrospective data while the experimental treatment group was not. Since information on recent labour force status and on recent earnings is missing for treatments, we are only able to obtain coarse estimates of P for the treated group. We use the coarse II model described in the notes to Table 6(a). The support region in the nonexperimental analysis is determined using the best predictor P model, so it is necessary to estimate which treatment group members would be excluded by imposing a common support to obtain impact estimates using nonexperimental methods. The impact estimates in the support region were obtained as follows. For controls and treatment, we first divide the coarse P distribution into 20 equal-size bins, then within-bin treatment estimates are estimated. The impact estimate in the overlap region is obtained as the weighted average of the within-bin estimates, with the weights given by the proportion of controls within each bin after deleting controls whose values of P lie outside the overlap region.

control group members ($D=0$). The lack of overlap and the restriction on the range of P means that the empirical counterpart to (A-2) is violated in our samples.

Table 4 presents empirical evidence on how the estimated programme effect changes when it is estimated on S_E and on S_P, where S_P is the region of overlapping support of P. The estimated experimental impact on earnings is given in the first column, while the experimental impact for the support of P common to participants and ENP comparison group members is presented in the next column. The bias is positive for all demographic groups and expressed as a fraction of the impact estimated over the full support of the programme is sizeable. (See the final column of the table.)

Nonparametric matching methods can only be meaningfully applied over regions of overlapping support. Simple nearest neighbour matching estimators could be mechanically applied over the whole region; but, as is evident from Figure 1, matches for experimental controls with high P values are likely to be poor. If nearest neighbour matching is performed with replacement, impact estimates are likely to be sensitive to the inclusion or omission of a few persons in the comparison group sample who are used repeatedly as matches for high P controls. Below we show that, in general, nearest neighbour estimators are more reliable in our application when matches are restricted to regions of overlapping support. Nonparametric kernel-based methods require that the density of the matching variables be strictly bounded away from zero. Therefore these methods require that the region of overlapping support be determined.

Appendix C discusses the details of how we impose the common support condition in constructing S_P, and an expanded version of Appendix C, available on request, presents some Monte Carlo evidence on the performance of alternative procedures for selecting S_P. In large samples, the alternative rules we investigate produce the same estimates. This is not true, however, in small samples and our results for male youth are somewhat sensitive to the choice of the procedure used to select S_P.

12.2. *Estimates of bias for alternative matching estimators*

Tables 5(a)–5(b) report estimates of the bias $\tilde{B}_{S_p}$ from a variety of matching estimators for the four demographic groups analysed in this paper. The first column in each table reports the unadjusted mean difference experimental control and nonparticipant earnings, for each quarter and averaged over the 18 month period following the date of randomization or eligibility determination ($t=1$ to $t=6$). The second row from the bottom shows the bias as a percentage of the estimated post-programme impact, obtained using experimental data over its full support. The final row of each table reports the bias as a percentage of the experimental impact estimated over the support of the nonexperimental estimator, S_P. We tried "optimal" bandwidths for the kernel estimators but found that they give less smooth estimates.[27] A bandwidth of 0·06 is used in this paper. Results are comparable for other fixed bandwidths within ±0·02 of 0·06 and for the average of the optimal bandwidths. For adult men, the average post-programme bias in the raw means is −$337 per month, which is almost nine times the estimated programme impact for them and seven times the average conditional difference-in-differences estimator of $52.

The nearest neighbour estimator that does not impose common support does surprisingly well for adult men, but this is not true for the other groups where imposing the common support restriction leads to a substantial reduction in bias. In results available on request, we show that simple averaging over the 5 or 10 closest neighbours usually improves its performance, although for adult women the bias is somewhat higher with the averaged estimator. P score matching estimators that use local linear regression weights, shown in the next two columns of the table usually perform roughly the same as the averaged nearest neighbour estimators. In results available on request from the authors, we obtain comparable results for smoothed and unsmoothed Mahalonobis metric estimators. The local linear regression estimator performs slightly better for youth groups but has no edge over the other matching estimators for the adult groups. The conditional difference-in-differences estimator performs somewhat better than the matching estimators for some groups but it is by no means the dominant estimator, despite the fact that the estimator is the only one for which the identifying assumptions are not rejected by formal statistical tests. Elsewhere (Heckman, Ichimura, Smith and Todd (1996a, b)), we document that the relatively strong performance of the matching estimators in our samples is a result of offsetting biases. The pointwise bias $B(P)$ is not zero and is substantially negative for low values of P. For different intervals of P, the bias in the matching estimators is substantial.[28] However, the pointwise bias $B(P)$ for the difference-in-differences estimator is much closer to zero.

Although the estimated biases are small relative to the simple difference in means, they are still a substantial fraction of total programme impact whether expressed over the full support of the experiment or the restricted support from the nonexperimental evaluation samples. (See the bottom two rows.) Moreover, there is substantial quarter-to-quarter variation in the estimated bias.

13. THE IMPORTANCE OF THE CONDITIONING VARIABLES

We next consider the effect of reducing the information used to predict programme participation. Previous nonexperimental evaluations of training programmes had access to much

27. The optimal plug-in bandwidth is derived in Heckman, Ichimura, Smith and Todd (1996c).

28. Persons with low P who participate in the programme have substantially lower earnings than nonparticipants.

TABLE 5(a)

Estimated bias for alternative nonparametric matching methods Experimental controls and eligible nonparticipants (ENPs)†‡*

Quarter	Difference in means ($\hat{\beta}$)	Nearest neighbour without common support	Nearest neighbour with common support	Local linear P score matching	Regression-adjusted local linear matching	Difference-in-differences from local linear P score matching	Difference-in-differences from regression-adjusted local linear matching
			Adult males				
$t=1$	−418 (38)	221 (56)	123 (67)	33 (59)	39 (60)	97 (62)	104 (63)
$t=2$	−349 (47)	−166 (151)	77 (83)	37 (61)	39 (64)	77 (89)	77 (92)
$t=3$	−337 (55)	−58 (206)	53 (96)	29 (78)	21 (80)	90 (114)	74 (114)
$t=4$	−286 (57)	161 (178)	86 (96)	80 (77)	65 (82)	112 (90)	98 (91)
$t=5$	−305 (57)	167 (196)	87 (100)	64 (77)	50 (83)	19 (95)	−5 (99)
$t=6$	−328 (63)	45 (191)	34 (113)	37 (82)	17 (90)	4 (105)	−35 (111)
Ave. 1 to 6	−337 (47)	62 (127)	77 (80)	47 (60)	38 (64)	67 (71)	52 (74)
As a % of impact**	775%	142%	177%	108%	87%	153%	120%
As a % of adjusted impact	552%	102%	126%	77%	62%	109%	85%
			Adult females				
$t=1$	−26 (24)	115 (30)	67 (36)	45 (33)	55 (36)	65 (31)	74 (30)
$t=2$	29 (25)	113 (53)	47 (46)	48 (37)	55 (39)	53 (40)	60 (39)
$t=3$	38 (26)	124 (107)	63 (59)	26 (48)	31 (52)	10 (56)	14 (59)
$t=4$	55 (30)	106 (102)	58 (52)	36 (39)	35 (45)	12 (53)	7 (56)
$t=5$	62 (34)	92 (111)	47 (51)	48 (40)	48 (45)	29 (51)	23 (53)
$t=6$	40 (36)	79 (84)	−6 (54)	23 (40)	16 (42)	−5 (51)	−18 (51)
Ave. 1 to 6	33 (26)	105 (69)	46 (43)	38 (33)	40 (38)	27 (38)	27 (39)
As a % of impact**	113%	358%	157%	130%	137%	93%	91%
As a % of adjusted impact	94%	300%	131%	109%	114%	78%	76%

* The table reports the bias for alternative matching methods. The bias in the first column is $\hat{\beta}$. The estimator in the second column does not restrict matches to a common support region. The estimators in the third through seventh columns restrict matches to a common support region and the bias estimates correspond to $\hat{\bar{B}}_{S_P}$.

† The best predictor model given in the second footnote to Table 2, is used to estimate the probability of programme participation. The conditioning variables in the regression adjusted local linear models are site, race, age, education, previous training, work experience in months, the local unemployment rate, indicator variables for marital status and for the presence of a child aged less than 6 in the household, and indicators for the quarter of the year and the year.

‡ A 2% trimming rule is used to determine the region of overlapping support (see Appendix C). A fixed bandwidth equal to 0·06 and a biweight kernel, defined in Appendix A, are used for the nonparametric estimates.

** The impacts in the table are mean monthly impacts for the six post-programme quarters, estimated using the experimental treatment and control data for the four JTPA training sites in our study. See the experimental impacts and the adjusted impacts in Table 4.

TABLE 5(b)

Estimated bias for alternative nonparametric matching methods Experimental controls and eligible nonparticipants (ENPs)†‡*

Quarter	Difference in means ($\hat{\beta}$)	Nearest neighbour without common support	Nearest neighbour with common support	Local linear P score matching	Regression-adjusted local linear matching	Difference-in-differences from local linear P score matching	Difference-in-differences from regression-adjusted local linear matching
			Male youth				
$t=1$	−51 (58)	146 (92)	49 (75)	3 (64)	8 (61)	43 (72)	80 (77)
$t=2$	2 (60)	197 (92)	98 (82)	40 (64)	28 (55)	43 (60)	61 (60)
$t=3$	5 (73)	202 (105)	83 (119)	33 (81)	−8 (77)	92 (80)	70 (86)
$t=4$	17 (69)	246 (105)	98 (94)	44 (81)	4 (71)	9 (74)	−5 (77)
$t=5$	82 (73)	283 (118)	138 (89)	84 (93)	42 (76)	18 (88)	−11 (81)
$t=6$	65 (77)	258 (145)	129 (121)	28 (93)	−31 (92)	−23 (89)	−64 (84)
Ave. 1 to 6	20 (57)	222 (88)	99 (78)	39 (66)	7 (53)	30 (49)	22 (48)
As a % of impact**	34%	382%	170%	67%	12%	52%	38%
As a % of adjusted impact	56%	617%	275%	108%	19%	84%	61%
			Female youth				
$t=1$	6 (31)	67 (54)	−7 (60)	31 (42)	−8 (46)	−7 (38)	−14 (41)
$t=2$	54 (40)	85 (57)	23 (60)	79 (53)	27 (49)	60 (49)	27 (47)
$t=3$	89 (44)	142 (62)	97 (78)	121 (60)	49 (52)	135 (59)	83 (58)
$t=4$	42 (50)	89 (56)	24 (72)	37 (59)	−28 (59)	45 (57)	4 (59)
$t=5$	64 (41)	121 (57)	51 (63)	65 (54)	8 (54)	45 (61)	−7 (63)
$t=6$	31 (46)	107 (82)	34 (70)	34 (65)	1 (62)	31 (70)	6 (69)
Ave. 1 to 6	48 (36)	102 (49)	37 (56)	61 (45)	8 (42)	52 (39)	17 (39)
As a % of impact**	7059%	15000%	5441%	8971%	1176%	7574%	2426%
As a % of adjusted impact	195%	415%	150%	248%	33%	209%	67%

* The table reports the bias for alternative matching methods. The bias in the first column is $\hat{\beta}$. The estimator in the second column does not restrict matches to a common support region. The estimators in the third through seventh columns restrict matches to a common support region and the bias estimates correspond to $\hat{\bar{B}}_{S_P}$.

† The best predictor model given in the second footnote to Table 2, is used to estimate the probability of programme participation. The conditioning variables in the regression adjusted local linear models are site, race, age, education, previous training, the local unemployment rate, indicator variables for marital status and the presence of a child aged less than 6 in the household, and indicators for the quarter of the year and the year.

‡ A 5% trimming rule is used to determine the region of overlapping support (see Appendix C), and a fixed bandwidth equal to 0·06 and a biweight kernel, defined in Appendix A, are used for nonparametric estimates.

** The impacts in the table are mean monthly impacts for the six post-programme quarters, estimated using the experimental treatment and control data for the four JTPA training sites in our study. See the experimental impacts and the adjusted impacts in Table 4.

less information about determinants of programme participation than we do. (Recall the summary in Table 1.) We seek to learn how the quality of the data used to estimate the probability of programme participation influences the effectiveness of the estimators. We use the same randomized control data and eligible nonparticipant data (ENP) as above, but vary the predictors in $P(X)$. Inclusion of additional predictors is not guaranteed to improve the effectiveness of a matching estimator. As an extreme example of this point, introducing conditioning variables that perfectly classified applicants and nonapplicants would make matching impossible, because there would be no overlap in the support of X given $D=1$ and X given $D=0$, and (A-2) would not be satisfied for any X. This example emphasizes the important point that matching requires X variables that are good enough to obtain conditional independence between Y_0 and D, but that are not "too good" i.e. that predict D perfectly.

Our evidence indicates that matching estimators perform best when variables describing recent unemployment and earnings histories, shown to be important determinants of programme participation in Heckman and Smith (1994), are used to predict participation. All of the estimates reported in Tables 5(a)–5(b) are based on participation models that use this information. Bias is substantially greater if only background demographic variables are used in estimating the probability of participation. Access to information on earnings in the previous year sometimes improves the performance of the matching estimator but the estimator performs best when data on recent labour force participation patterns are available.

We estimate the probability of programme participation, $P(X)$, using four different sets of variables that differ in the type of data used to describe recent labour force dynamics. The names ascribed to the different coarser conditioning sets are "regular", coarse I, coarse II, coarse III. The regressors included in each model and their relationship to the regular P scores are described in the footnotes to Table 6(a). Briefly, coarse I contains only limited demographic information. Coarse models II and III supplement the background data using different information about recent labour force histories and earnings prior to the date of enrollment into the programme. The "regular" models for $P(X)$ participation use information on demographics, earnings and labour force histories.[29]

Table 6(a) shows the importance of building a good model of programme participation. (The final three columns of this table are discussed in the following sections.) Results are reported for the regression-adjusted local linear regression estimator using the rich or regular model for $P(X)$ and the coarser scores. For adult males, the choice of the matching variables X matters greatly. Failure to control for earnings or employment histories (Coarse I) produces a badly biased estimator. Controlling for previous annual earnings alone results in less bias (Coarse II) but augmenting the conditioning set to include information on recent labour force transitions results in a substantial reduction in bias (Coarse III). The safest conclusion to draw from this evidence is that matching on $P(X)$ works well if it is based on a good model of programme participation. For each demographic group, the matching estimator based on our local linear regression model using the best predicting model of programme participation performs as well or better than any other model in the table. For adult women, the use of the richer models for $P(X)$ leads to a slight increase in the bias. For male youth, the bias is substantially higher under the coarse

29. In figures available upon request, we plot the P scores for comparison and control group participants for the different demographic groups. The nonoverlapping support problem is present for all the coarse scores. Use of coarse I exacerbates the support problem, shrinking the overlapping support region, but for other coarse scores there is no particular pattern across the groups.

TABLE 6(a)

Bias from local linear regression matching estimator†
Under alternative predictor models for the probability of programme participation

Quarter	Regular††	Coarse I‡	Coarse II‡	Coarse III‡	SIPP§	Site mismatch§§	No-show§§§
				Adult males			
$t=1$	39 (60)	−390 (51)	−228 (67)	−84 (77)	249 (77)	−184 (110)	58 (38)
$t=2$	39 (64)	−312 (58)	−193 (61)	−39 (88)	123 (79)	−154 (120)	37 (39)
$t=3$	21 (80)	−286 (62)	−153 (57)	−36 (96)	76 (81)	−147 (127)	27 (42)
$t=4$	65 (82)	−231 (63)	−104 (66)	−9 (92)	13 (93)	−164 (132)	−6 (48)
$t=5$	50 (83)	−244 (73)	−146 (70)	20 (96)	* (*)	−211 (132)	1 (48)
$t=6$	17 (90)	−286 (84)	−172 (79)	−3 (111)	* (*)	−189 (112)	−21 (48)
Ave. 1 to 6	38 (64)	−291 (54)	−166 (56)	−25 (83)	115 (78)	−175 (108)	16 (37)
				Adult females			
$t=1$	55 (36)	−69 (33)	−73 (29)	40 (30)	167 (35)	−84 (56)	26 (28)
$t=2$	55 (39)	−9 (33)	−15 (29)	63 (34)	122 (40)	−57 (69)	9 (36)
$t=3$	31 (52)	5 (34)	−6 (31)	42 (40)	98 (40)	−62 (70)	−13 (37)
$t=4$	35 (45)	5 (34)	−10 (34)	21 (47)	87 (43)	−42 (60)	2 (35)
$t=5$	48 (45)	14 (38)	−6 (37)	26 (48)	* (*)	−38 (63)	1 (31)
$t=6$	16 (42)	−10 (37)	−24 (37)	−2 (44)	* (*)	−35 (58)	−2 (34)
Ave. 1 to 6	40 (38)	−11 (31)	−22 (29)	32 (35)	119 (39)	−53 (57)	4 (30)
				Male youth			
$t=1$	8 (61)	−41 (56)	−40 (53)	37 (61)	302 (120)	−29 (106)	104 (44)
$t=2$	28 (55)	10 (62)	9 (63)	45 (65)	275 (140)	12 (110)	36 (43)
$t=3$	−8 (77)	−29 (74)	−24 (76)	10 (83)	217 (153)	38 (136)	70 (48)
$t=4$	4 (71)	2 (69)	8 (70)	30 (81)	157 (176)	110 (162)	116 (45)
$t=5$	42 (76)	63 (72)	73 (71)	46 (99)	* (*)	132 (182)	95 (48)
$t=6$	−31 (92)	9 (76)	21 (75)	−68 (131)	* (*)	−63 (210)	108 (53)
Ave. 1 to 6	7 (53)	2 (52)	8 (52)	17 (70)	238 (144)	33 (128)	88 (38)
				Female youth			
$t=1$	−8 (46)	3 (34)	17 (32)	60 (45)	−11 (72)	74 (76)	55 (32)
$t=2$	27 (49)	46 (39)	54 (39)	81 (46)	−31 (79)	91 (77)	52 (32)
$t=3$	49 (52)	64 (42)	72 (41)	101 (51)	−37 (82)	84 (90)	74 (34)
$t=4$	−28 (59)	18 (48)	18 (47)	48 (57)	−55 (85)	−33 (119)	21 (36)
$t=5$	8 (54)	46 (43)	48 (41)	46 (56)	* (*)	21 (131)	37 (39)
$t=6$	1 (62)	37 (50)	40 (48)	38 (62)	* (*)	−3 (114)	57 (36)
Ave. 1 to 6	8 (42)	36 (36)	41 (35)	62 (42)	−34 (78)	39 (83)	49 (26)

† A 2% trimming rule is used for adult males and females to determine the overlapping support region (see Appendix C) and a 5% trimming rule is used for male and female youth. A fixed bandwidth of 0·06 and a biweight kernel, described in Appendix A, are used to compute the estimates for all four groups. Bootstrapped standard errors are shown in parentheses. They are based on 50 replications with 100% sampling.

* Data not available to compute for this quarter. Averages reported over available quarters.

†† The regular predictor model is the model for the probability of programme participation that maximizes the percent correctly classified. The regressors in the model for each demographic group are given in the footnote to Table 2.

‡ Coarse I predictor model includes indicator variables for site, race, age, education, marital status, and for the presence of children aged less than 6 in the household. Coarse II predictor model augments Coarse I with earnings from the year preceding enrollment into the programme. Coarse III predictor model augments Coarse I with indicators for labour force transition patterns.

§ SIPP predictor model includes indicators for age, race, education, marital status, children aged less than 6, labour force transition patterns and levels of earnings in the preceding year. The data used are SIPP JTPA eligibles matched with Experimental JTPA Controls.

§§ Site Mismatch predictor model is the same as the regular predictor model. The data used are Controls from Providence and Jersey City matched with ENPs from Corpus Christi and Fort Wayne.

§§§ The data used are Experimental JTPA Controls matched with experimental JTPA Persons assigned to Treatment who enrolled in JTPA but dropped out before receiving services. No-show predictor model includes indicator variables for site and recommended training services. For adult males, the model also includes earnings last year, earnings squared, indicators for preferred language Spanish and for preferred language other than Spanish or English, and an indicator for whether enrollment in JTPA was required. For adult women, the model also includes indicators for last employed 0–6 months ago and for 7–12 months ago and an indicator for whether enrollment in JTPA was required. For male youth, the model also includes indicator variables for race and for the quarter of the year. For female youth, the model also includes an indicator for race. (Models were chosen to maximize the percent correctly classified using available variables—see Appendix E available on request from the authors).

TABLE 6(b)

Bias from difference-in-differences local linear regression estimator†
Under alternative predictor models for the probability of programme participation

Quarter	Regular‡	Coarse I‡	Coarse II‡	Coarse III†	SIPP‡	Site mismatch‡
			Adult males			
$t=1$	104 (63)	167 (67)	31 (57)	67 (68)	−97 (38)	−135 (126)
$t=2$	77 (92)	143 (82)	−80 (62)	103 (107)	−230 (51)	−72 (130)
$t=3$	74 (114)	62 (95)	−158 (71)	105 (134)	−277 (52)	−9 (141)
$t=4$	98 (91)	33 (93)	−150 (82)	47 (109)	−338 (72)	19 (151)
$t=5$	−5 (99)	−73 (104)	−254 (86)	−29 (122)	* (*)	−136 (167)
$t=6$	−35 (111)	−143 (106)	−255 (96)	−36 (129)	* (*)	−82 (165)
Ave. 1 to 6	52 (74)	32 (78)	−144 (61)	43 (95)	−236 (45)	−69 (123)
			Adult females			
$t=1$	74 (30)	80 (24)	71 (23)	86 (25)	−43 (10)	38 (42)
$t=2$	60 (39)	69 (35)	54 (33)	85 (35)	−86 (25)	66 (56)
$t=3$	14 (59)	20 (39)	−1 (37)	25 (49)	−99 (27)	43 (66)
$t=4$	7 (56)	−16 (42)	−34 (40)	−15 (59)	−116 (36)	67 (62)
$t=5$	23 (53)	−9 (42)	−26 (42)	−5 (55)	* (*)	80 (68)
$t=6$	−18 (51)	−44 (42)	−52 (43)	−40 (53)	* (*)	48 (72)
Ave. 1 to 6	27 (39)	17 (31)	2 (30)	23 (39)	−86 (21)	57 (50)
			Male youth			
$t=1$	80 (77)	123 (56)	111 (56)	194 (85)	22 (53)	−92 (98)
$t=2$	61 (60)	102 (73)	81 (72)	58 (72)	12 (78)	1 (118)
$t=3$	70 (86)	−9 (88)	−23 (87)	38 (100)	−60 (102)	33 (152)
$t=4$	−5 (77)	−45 (81)	−54 (80)	−6 (85)	−85 (135)	32 (157)
$t=5$	−11 (81)	34 (85)	28 (82)	−3 (108)	* (*)	25 (188)
$t=6$	−64 (84)	18 (83)	19 (80)	−74 (126)	* (*)	−117 (211)
Ave. 1 to 6	22 (48)	37 (56)	27 (54)	34 (59)	−28 (81)	−20 (122)
			Female youth			
$t=1$	−14 (41)	59 (39)	62 (41)	14 (36)	−14 (31)	5 (66)
$t=2$	27 (47)	82 (42)	75 (42)	48 (47)	−67 (33)	29 (88)
$t=3$	83 (58)	116 (51)	106 (52)	91 (62)	−90 (46)	89 (111)
$t=4$	4 (59)	53 (48)	36 (48)	30 (56)	−96 (60)	−21 (139)
$t=5$	−7 (63)	−1 (43)	−8 (43)	−16 (60)	* (*)	44 (154)
$t=6$	6 (69)	2 (53)	5 (53)	−3 (71)	* (*)	2 (134)
Ave. 1 to 6	17 (39)	52 (35)	46 (35)	28 (39)	−67 (35)	25 (87)

† A 2% trimming rule is used to determine the overlapping support region for adult groups and a 5% trimming rule is used for the youth groups (see Appendix C). A fixed bandwidth equal to 0·06 and a biweight kernel, described in Appendix A, are used to compute the nonparametric estimates.
* Data not available to compute for these periods. Averages are reported over available quarters.
‡ The alternative predictor models for the probability of programme participation are described in the footnote to Table 6(a).

III model for $P(X)$ and of comparable magnitude for the other models and for female youth the richest model for $P(X)$ has the lowest estimated bias.

Table 6(b) presents analogous bias estimates for the conditional difference-in-differences estimator. For the richest conditioning information ("regular" $P(X)$ models) the difference-in-differences estimator sometimes exhibits lower bias, but for other conditioning sets the bias varies. For this estimator, the choice of X in $P(X)$ makes less of a difference to the effectiveness of the method in our samples and is an attractive feature of it.

We have performed a comparable sensitivity analysis for variations in the regressors included in the outcome equations (T), including no regressors. These results are available on request from the authors. We find little variation in estimated biases across alternative specifications of the outcome model.

14. THE IMPORTANCE OF GEOGRAPHIC MISMATCH AND NONUNIFORMITY OF THE SURVEY INSTRUMENT

We now examine the empirical importance of two additional sources of bias that plague nonexperimental evaluations: mismatch of the geographic location of programme participants and comparison group members and mismatch of the survey instrument used to collect participant and comparison group data. Participants, experimental controls and nonexperimental comparison group members were administered the same questionnaire and reside in the same labour markets. Therefore they are not subject to these sources of bias. To investigate this problem we consider alternative comparison group samples drawn from SIPP (Survey of Income and Programme Participation).

The SIPP data are far richer than those used to construct nonexperimental comparison groups in previous evaluations of job training programmes. It is longitudinal with monthly observations on both earnings and employment and the samples are large. It also contains enough information to determine eligibility status for JTPA, which requires very detailed information about family structure, six-months earnings histories and welfare participation (see Devine and Heckman (1996)). As noted in Table 1, few major evaluations of job training programmes have had rich enough data to determine the eligibility of comparison group members. We restrict our SIPP comparison group samples to persons who meet the eligibility requirements of the JTPA programme. Like the ENP comparison group samples, the SIPP samples consist of eligible nonparticipating persons.

14.1. *The SIPP data*

We use the 1988 SIPP panel data set with observations that span from October, 1987 to December, 1989. The 1988 panel covers the time period of the JTPA experiment where random assignment took place from November, 1987 to September, 1989. The data contain earnings and labour force histories and information on participation in JTPA, although few JTPA participants are found in the data in any month. Appendix E, available on request, describes the SIPP data in detail.

A drawback in using SIPP as a comparison group for the JTPA data is that it is a broad, nationally-representative sample while the JTPA experiment was conducted in a few cities of moderate size. To protect confidentiality, information about the exact location of respondents was suppressed so it is not possible to find SIPP comparison group members in the same labour market as JTPA programme participants. Smith (1995) compares the earnings of the SIPP and ENP comparison groups. He finds that mean differences in earnings levels remain even after a variety of geographic and local labour market matching schemes are tried and that these differences cannot be explained solely by differences in local labour market variables like local unemployment rates.

We compare the earnings of JTPA experimental controls with those of the JTPA-eligible SIPPs. They are substantially different even after matching on region of residence or local unemployment rates in region of residence. However employment rates and labour force participation rates are similar between the two groups, suggesting that the observed earnings differences are due to cost of living differences or to differences in the survey questions on earnings. As shown in the middle panel of Table 2, with the exception of adult males, the bias in the raw data (B) is greater using the SIPP samples compared to using the ENP samples. For adults, the component of bias B attributable to selection B_3, is both absolutely and proportionately larger in the SIPP-control data than in the ENP-control data. The uncontrolled heterogeneity in questionnaires and locations across ENP

and control groups substantially exacerbates the selection problem in conducting evaluations.

14.2. *Matching results for the SIPP samples*

The column labelled "SIPP" in Table 6(a) presents the estimated bias from the local linear regression matching estimator based on the best-predicting model for $P(X)$.[30] A new model for $P(X)$ is fit using the SIPP data pooled with the data from the experimental control group. The data are coarser since we lack the detailed geographical location available in the ENP data. Otherwise, the model for $P(X)$ is very similar in specification to that used in the ENP-control analysis. The contribution of questionnaire and labour market mismatch to evaluation bias B is of the same order of magnitude as the contribution of poor predictors. LaLonde's influential study used comparison groups that are mismatched both geographically and in terms of the questionnaires. Even closely regionally-aligned SIPP samples produce estimates that are substantially biased. Applying the difference-in-differences version of the estimator often attenuates the bias (see Table 6(b)) as would be expected if discrepancies in questionnaires and local labour markets are temporally-stable sources of the participant group—comparison group bias.

14.3. *An internal check of the importance of geographical mismatch*

In an effort to disentangle survey effects from local labour market effects, we split the JTPA ENP-experimental control data into two geographically-mismatched samples. Since the questionnaires are the same for both groups, the only new source of bias is local labour market or site effects. The sample splitting reduces the sample sizes, and the youth estimates can no longer be considered very reliable because the samples are so small. The induced bias is substantial (see Table 6(a) column labelled "Site Mismatch"), and comparable in absolute value to the bias that results from using SIPP. Note that the differences in estimated geographical effects across demographic groups rules out cost of living differences as the major reason for local labour markets effects. Presumably those effects would be the same across demographic groups. Taking conditional difference-in-differences greatly attenuates the bias, as would be expected if local labour market characteristics are persistent over time. (See Table 6(b).)

Smith (1995) investigates the separate roles of survey instruments and geographical mismatch in accounting for the discrepancy between ENP and SIPP earnings. He concludes that about two-thirds of the discrepancy between the two data sources is due to differences in the questionnaires and the rest is attributable to geographical mismatch of participants and control group members.[31]

15. USING NO-SHOWS AS A COMPARISON GROUP

For most social programmes, there is a group of persons who apply to the programme and are accepted into it but for some reason do not enroll in it. These people are distinct

30. A complete set of SIPP results in the format of Table 5 is available on request from the authors. The numbers in Table 6(a) and the numbers in Table 6(b) are based on the local linear regression estimator used in the third to last column and the last column of Tables 5, respectively. The bias estimates for the other estimators is of the same order of magnitude as the SIPP estimators recorded in Table 6(a) and 6(b).

31. His findings are of general interest because the JTPA data were collected in the format of the widely-used NLSY (National Longitudinal Survey of Youth) data.

from dropouts among enrollees. It may be that in learning about the programme through applying for it, they change their mind or that better opportunities arise.[32] No-shows are in many ways similar to treated persons, and are good candidates for a comparison group. They are located in the same labour market as participants and are administered the same questionnaire as participants. If not enrolling in a programme were random with respect to outcomes, then non-enrollees would be like an experimental control group. In reality, enrollment is probably not random, but we can attempt to control for the differences between non-enrollees and treated persons using the same methods as used to compare programme participants with ENP comparison group members.

We assess the performance of matching methods for eliminating differences between no-shows and a randomized-out control group. As before, $D=1$ refers to controls but now $D=0$ refers to no-shows. In this context, $P(X)$ represents the conditional probability of being an enrollee (i.e. of not being a no-show) among persons who at one stage expressed their intention to enroll in the programme.[33] Figure 2 plots the empirical densities of estimated $P(X)$ values which look very similar across the control and non-enrollee groups. The bottom section of Table 2 reports the mean difference in outcomes between controls and no-shows, B, and its components for all four demographic groups. For three of four groups, the bias B is smaller than it is for the ENP—control comparison. However, the composition of B is weighted more heavily toward selection bias component B_3. B is lower for adults than it is in the ENP-control comparison but the proportion of B due to selection bias is larger. For male youths, B is substantial and the contribution of selection bias B_3 is large both absolutely and proportionately.

Table 7 presents the estimated quarterly bias estimates associated with the local linear matching and regression-adjusted local linear matching estimators. The results from the other matching methods are comparable and are available on request from the authors. The first column of each table indicates that for the adult demographic groups, there is very little bias, as conventionally measured, (B), that arises from using a simple mean difference estimator. For the adult groups, the raw means for no-shows are very similar to those for the randomized-out control group without any adjustment for group characteristics. When matching methods are used, and a more theoretically appropriate measure of bias is used, the local linear regression-adjusted matching estimator emerges as an effective method but in general the matching estimators yield biases that are slightly higher than the difference in raw means. These results are not surprising. The components of bias B discussed in Section 9, may coincidentally offset each other. Small values of B do not imply a small bias for $M(S)$. Data limitations prevent us from applying the conditional difference-in-differences estimator to the no-show data.[34]

In the adult samples, it appears that the no-shows are already well-matched to the control group as indicated by the densities of $P(X)$, so matching would not be expected to lead to a significant improvement over the raw means. However, the discrepancy between $P(X)$ for $D=1$ and $D=0$ is greater for youth and the performance of matching is worse for the youth samples.[35]

32. Still another possibility considered in the literature on performance standards is that programme administrators have an incentive to give persons trial treatments before enrolling them in the programme. Through such interventions, they can learn about the probable success of the person in the programme and then encourage or discourage that individual from enrolling. Individuals who fail to enroll are usually not counted against the JTPA site. There is some evidence that this type of behaviour, called cream-skimming, took place at JTPA sites. (See Heckman, Smith and Taber (1996).)

33. Estimated coefficients are reported in Appendix E, which is available on request from the authors.

34. Pre-programme earnings data are not available for JTPA no-shows.

35. We observe lower prediction rates for the no-show model than for the participation model, which indicates that we are better able to predict the programme participation decision than the no-show decision. This is due in part to the way we collected the data on programme participation which focused on the application-acceptance decision (ENP-Control).

TABLE 7

Estimated bias for alternative nonparametric matching methods Experimental controls and no-shows†*

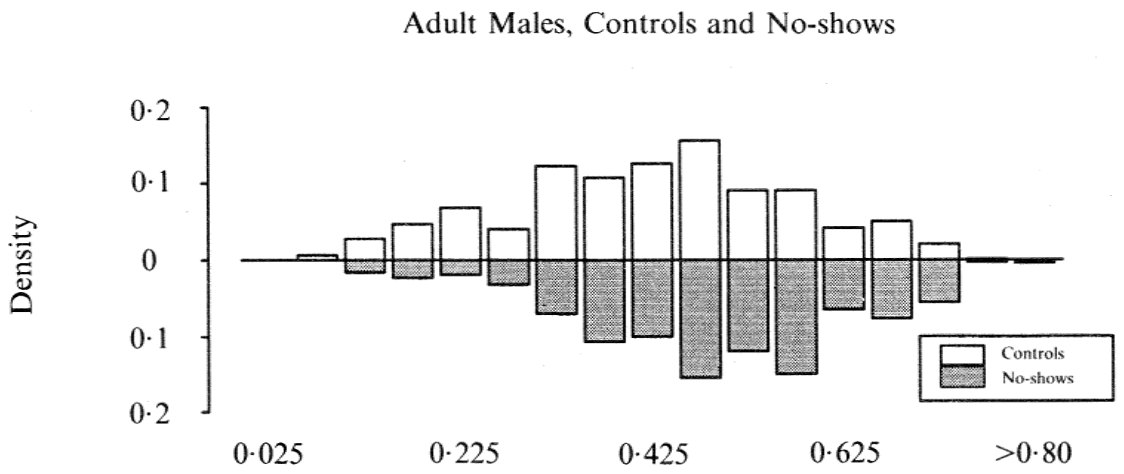
Quarter	Difference in means	Local linear P score matching	Regression-adjusted local linear matching	Difference in means	Local linear P score matching	Regression-adjusted local linear matching
	Adult males			Adult females		
$t=1$	64 (35)	66 (39)	58 (38)	17 (24)	28 (30)	26 (28)
$t=2$	32 (37)	45 (40)	37 (39)	7 (30)	11 (38)	9 (36)
$t=3$	26 (41)	36 (42)	27 (42)	−5 (30)	−9 (38)	−13 (37)
$t=4$	19 (46)	3 (49)	−6 (48)	12 (28)	5 (36)	2 (35)
$t=5$	22 (49)	10 (51)	1 (48)	14 (24)	4 (32)	1 (31)
$t=6$	7 (50)	−12 (52)	−21 (48)	11 (27)	1 (36)	−2 (34)
Ave. 1 to 6	29 (37)	25 (39)	16 (37)	9 (23)	7 (31)	4 (30)
As a % of impact**	66%	57%	37%	32%	23%	14%
As a % of adjusted impact	47%	41%	26%	27%	20%	11%
As a % of Control-ENP	8%	53%	42%	29%	18%	10%
	Male youth			Female youth		
$t=1$	92 (37)	116 (41)	104 (44)	12 (30)	53 (33)	55 (32)
$t=2$	33 (38)	49 (43)	36 (43)	17 (27)	52 (37)	52 (32)
$t=3$	56 (42)	80 (47)	70 (48)	39 (31)	80 (35)	74 (34)
$t=4$	111 (36)	133 (42)	116 (45)	−7 (32)	23 (39)	21 (36)
$t=5$	100 (39)	108 (46)	95 (48)	17 (34)	39 (46)	37 (39)
$t=6$	111 (43)	111 (47)	108 (53)	30 (31)	58 (39)	57 (36)
Ave. 1 to 6	84 (31)	99 (34)	88 (38)	18 (23)	51 (30)	49 (26)
As a % of impact**	144%	171%	152%	2639%	7474%	7273%
As a % of adjusted impact	233%	276%	245%	73%	207%	201%
As a % of Control-ENP	419%	255%	1258%	37%	83%	618%

* The table reports the bias for alternative matching methods. The bias in the first column is $\hat{\beta}$. The estimators in the second and third columns for each group restrict matches to the common support region and the bias estimates correspond to $\hat{B}_{S_P}$.

† The predictor model given in the second footnote to Table 2 is used to estimate the probability of programme participation. The conditioning variables in the regression adjusted local linear models are site, age, education, race, the local unemployment rate, indicator variables for marital status and for the presence of children aged less than 6 in the household, and indicators for the quarter of the year and the year.

‡ A 2% trimming rule is used to determine the region of overlapping support for adult groups and a 5% trimming rule is used for youth groups (see Appendix C). A fixed bandwidth equal to 0·06 and a biweight kernel, defined in Appendix A, are used for nonparametric estimates.

** The impacts in the table are mean monthly impacts for the six post-programme quarters, estimated using the experimental treatment and control data for the four JTPA training sites in our study. See the experimental impacts and the adjusted impacts in Table 4.



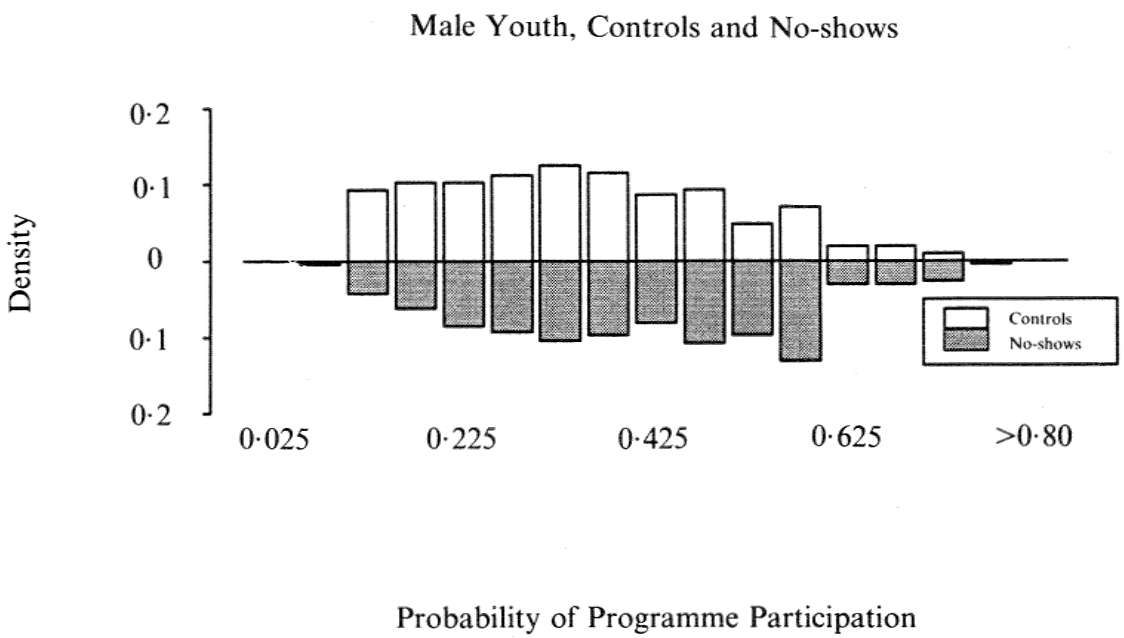
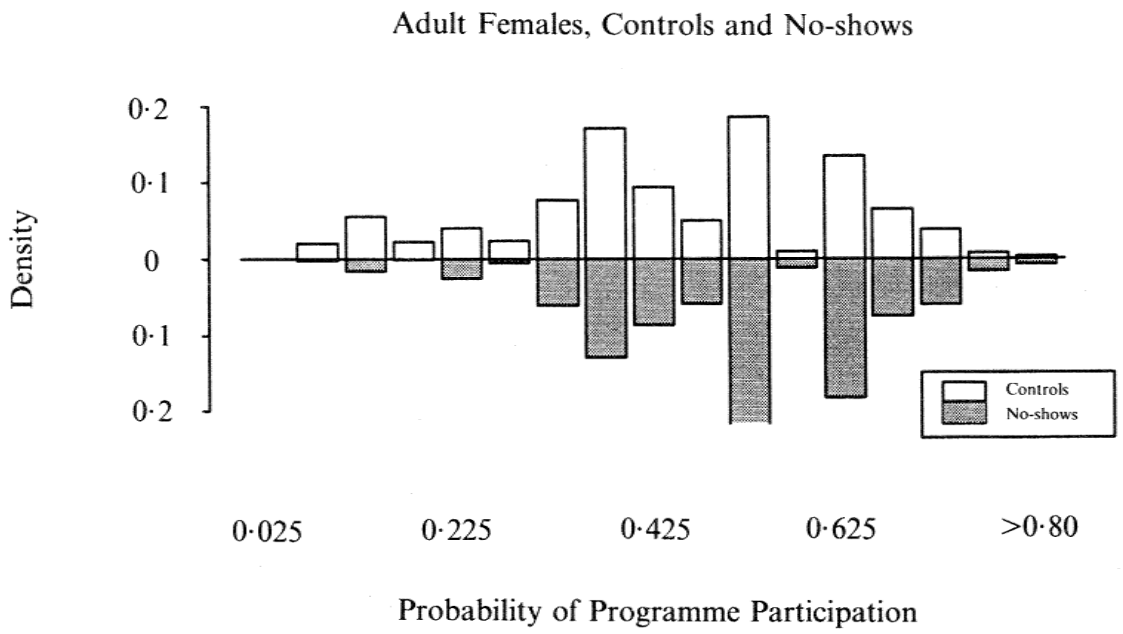


FIGURE 2. Density of Estimated Probability of Programme Participation for No-shows



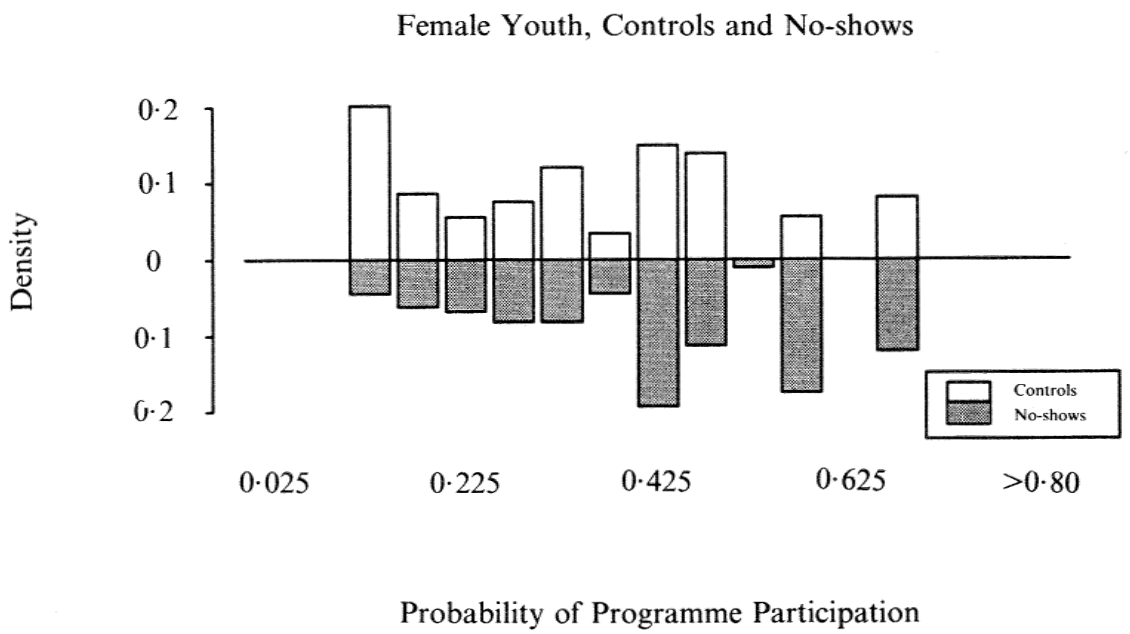


FIGURE 2. continued.

Little of a general nature can be concluded from our evidence about the use of no-shows. As programmes are improved, the proportion of no-shows would be expected to decline. No-shows would then become a more select sample. Nonetheless, the lower biases estimated for adult no-show comparison groups confirm the conclusions reached in the analysis of ENP-control samples. A major component of what is commonly regarded as evaluation bias is due to mismatch on observables. Picking comparison groups of persons in the same labour market, administering them the same questionnaire, and weighting the comparisons comparably goes a long way toward replicating the estimates produced from a social experiment.

16. CONCLUSION AND DISCUSSION OF RELATED WORK

This paper examines various matching methods and extensions of matching methods for evaluating job training programmes using several different sources of nonexperimental data combined with experimental data. Our approach to evaluation has two main components: (a) estimating a model that successfully predicts who participates in the programme; and (b) using the estimated probability of participation in matching and extensions of the matching method. The estimators that we propose in this paper and rigorously justify in Heckman, Ichimura Todd (1997) perform well among the estimators we examine here. Their application to the evaluation of a major job training programme produces impact estimates and inferences "fairly close" to those produced from a randomized evaluation of the programme, although estimated selection bias is still a substantial fraction of the experimentally-estimated programme impacts.

We determine that a regression-adjusted semiparametric conditional difference-in-differences matching estimator often performs the best among the class of estimators we examine, especially when omitted time-invariant characteristics are an important source of bias. Placing nonparticipants in the same labour market as participants, administering both the same questionnaire, and weighting their observed characteristics in the same way as that of participants, produces estimates of programme impacts that are fairly close to those produced from an experimental evaluation. We find that for adult males, selection bias due to differences in unobservables between participants and non-participants is a relatively small component of total evaluation bias as conventionally measured (B). For other groups, selection bias is a larger component of bias as conventionally measured. If a few simple empirical principles are followed, one can dramatically improve on the poor performance of nonexperimental estimators emphasized in LaLonde's (1986) influential paper. Our evidence also indicates that a substantial fraction of the bias reported by LaLonde was due to questionnaire mismatch and geographic mismatch and not self-selection bias generated from a common distribution of unobservables across comparison and treatment groups. It is the distributions of the characteristics that appear to be different across groups.

A major issue that arises in the application of matching in ordinary observational analyses is determining if the assumptions that justify matching estimators are satisfied. In Heckman, Ichimura and Todd (1996b), we confront this issue and propose a test of the validity of matching under the assumption that there is an exclusion restriction. Letting $X=(T, Z)$, we assume that there is some matching variable in Z not in T. Then, if Assumption (A-4′) is true, $E(Y_1 \mid T, P(Z), D=1)-E(Y_0 \mid T, P(Z), D=0)$ should not be a function of the Z variables excluded from T provided that Z does not predict from the programme for participants. Using kernel estimates of the two conditional expectations, we develop a formal test of this hypothesis and related hypotheses.

Matching and our extensions of it differ from the widely-used instrumental variables estimator. In our context, the conventional instrumental variables estimator is based on the assumption that $E(U_0 | X, P, D=1)=0$ and $E(U_0 | X, P, D=0)=0$.[36] Elsewhere we have shown that using our estimated P as an ordinary instrument produces substantially biased estimators of programme impact. (Heckman (1995).) Our evidence does not justify application of the instrumental variables estimator to our data. The matching estimator assumes that P balances the bias, $E(U_0 | P, D=1)=E(U_0 | P, D=0)$, but does not assume that either term equals zero. Similarly, the conditional difference-in-differences estimator introduced in this paper also assumes bias is balanced in the differences, either in levels or residuals, but does not equate conditional means to zero.

A major finding of this paper is that comparing the incomparable—i.e. violating the common support condition for the matching variables—is a major source of evaluation bias as conventionally measured. A major limitation of the non-experimental method is that the support in comparison samples may be very different from the support in control samples. Restricting the application of nonexperimental methods to regions of common support may change the parameter being estimated in a non-experimental evaluation from what is estimated by an experiment. ($M(S_{10}) \neq M(S_E)$ where S_{10} is the common support in a nonexperimental evaluation and S_E is the support in the benchmark experimental evaluation). The evidence reported in Table 4 demonstrates that this source of bias is substantial for certain demographic groups. In our data, restricting S to the common subset to balance the bias produces estimates of mean programme impacts that are uniformly higher than the experimentally-determined estimator for the full support of programme participants.

As is true of any empirical study, our findings may not generalize beyond our data. However, it is important to note that the proposed two stage evaluation strategy performs comparably in its application to four distinct demographic groups. Moreover, the main features of the JTPA programme are common to many other job training programmes both in the U.S. and elsewhere. Participants in the JTPA programme have the same pre-programme dip in earnings as found in numerous job training programmes and it is likely that similar factors operate in generating selection decisions in all of these programmes. The range of services offered is comparable to other programmes. (Heckman (1995).) Thus it is likely that the insights gained from our study of the JTPA programme on the effectiveness of different estimators also apply in evaluating other training programmes targeted toward disadvantaged workers.

APPENDIX

A. *Test statistics and variance estimators*

In this appendix, we discuss the details of the tests for conditional independence, conditional mean independence, given by hypotheses $H_{(A\text{-}3)}$, $H_{(A\text{-}3')}$, $H_{(A\text{-}4)}$, and $H_{(A\text{-}4')}$ in the text as well as the hypotheses concerning the validity of the conditional differences-in-differences method, $H_{(A\text{-}5)}$ and $H_{(A\text{-}5')}$. We first give test statistics for use in a single cross-section and then present joint tests across multiple time periods, allowing for the possibility of unbalanced panel data.[37]

Testing for conditional independence ($H_{(A\text{-}3)}$ *and* $H_{(A\text{-}3')}$):

36. Heckman and Smith (1997) compare matching and IV methods.

37. The test statistics for conditional mean independence were proposed in Heckman, Ichimura, Smith and Todd (1996b).

To test for $H_{(A\text{-}3)}$ at a point $Y_0=y_0$ conditional on $P=p$, we estimate the conditional c.d.f. $\hat{F}_d(p)=\hat{F}_d(y_0|P=p, D=d)$, for $D=0$ and $D=1$ groups using local linear regression smoothing methods.[38] The estimator $\hat{F}_d(p)$ asymptotically satisfies

$$(\hat{F}_d(p)-F_d(p))\sim N(B_d, V_d),$$

where

$$B_d=\tfrac{1}{2}F_d''(p)\frac{C_2}{C_1}a_{N_d}^2 \quad \text{and} \quad V_d=(N_d a_{N_d})^{-1}\frac{\text{Var}(\varepsilon_{id}|D=d, P=p)}{f_d(p)}\frac{C_3}{C_1}.$$

B_d is a bias term. $F_d''(p)$ is the second derivative of the conditional c.d.f. with respect to p evaluated at p, $f_d(p)$ is the density of P evaluated at p, $\varepsilon_{id}=1(Y_{0i}\leqq y_0)-F_d(P_i)$, N_d is the number of observations, and a_{N_d} is a bandwidth that converges to zero as N_d gets large and satisfies $N_d a_{N_d}\to\infty$, $N_d a_{N_d}^5<\infty$.[39] The terms C_1, C_2, and C_3 are constants that depend on the kernel function $G(\cdot)$ used in the local linear regression:

$$C_1=\int s^2G(s)ds\int G(s)ds-\left[\int sG(s)ds\right]^2,$$

$$C_2=\left[\int s^2G(s)ds\right]^2-\int s^3G(s)ds\int sG(s)ds,$$

$$C_3=\int\left[\int s^2G(s)ds-u\int sG(s)ds\right]^2G^2(u)du,$$

where the integrals are evaluated over all $s\in(-\infty, \infty)$.[40] The kernel function used to obtain all estimates reported in this paper is the biweight kernel

$$G(s)=15/16(s^2-1)^2 \quad |s|<1;$$
$$=0 \quad |s|\geqq 1.$$

For testing $H_{(A\text{-}3)}$, the test statistic is

$$(\hat{F}_1(p)-\hat{F}_0(p))'(\hat{V}_0+\hat{V}_1)^{-1}(\hat{F}_1(p)-\hat{F}_0(p))\sim\chi_1^2.$$

Under the null, if the same bandwidth is used to estimate $\hat{F}_1(p)$ and $\hat{F}_0(p)$ then the bias expressions cancel so only estimators for the variance are needed.[41] Consistent estimators $\hat{V}_1$ and $\hat{V}_2$ of V_1 and V_2 are presented below. Test of conditional independence of residuals ($H_{(A\text{-}3')}$) are analogous, except that they replace Y_{0i} by $Y_{0i}-X_i\hat{\beta}=\hat{U}_i$.

Testing the conditional mean independence assumption ($H_{(A\text{-}4)}$, $H_{(A\text{-}4')}$)

The test of hypothesis $H_{(A\text{-}4)}$ is similar to that of $H_{(A\text{-}3)}$. First estimate $\hat{m}_1(p)=\hat{E}(Y_0|P=p, D=1)$ and $\hat{m}_0(p)=\hat{E}(Y_0|P=p, D=0)$ by a local linear regression of Y_{0i} on P_i evaluated at p. $\hat{m}_d(p)$ is distributed as

$$(\hat{m}_d(p)-m_d(p))\sim N(\tilde{B}_d, \tilde{V}_d),$$

where

$$\tilde{B}_d=\tfrac{1}{2}m_d''(p)\frac{C_2}{C_1}a_{N_d}^2 \quad \text{and} \quad \tilde{V}_d=(N_d a_{N_d})^{-1}\frac{\text{Var}(\varepsilon_{id}|D=d, P=p)}{f_d(p)}\frac{C_3}{C_1}.$$

38. For a given y_0 we perform a nonparametric regression of the indicator variable $1(Y_i\leqq y_0)$ on P_i for $D=0$ and $D=1$ groups and evaluate the function at p. A fixed bandwidth equal to 0·06 is used in the test.

39. See Heckman, Ichimura, Smith and Todd (1996c), Theorem 5, for why these restrictions are needed on the bandwidth.

40. Here we assume p is an interior point of the support.

41. If kernel regression is used instead of local linear regression, the bias terms would not cancel and the test statistic would have a noncentrality parameter. Thus, using local linear regression, with a common bandwidth for estimating each of the nonparametric functions, simplifies the test because it yields a test statistic that is centrally distributed.

Here, $\varepsilon_{id} = Y_{0i} - m_d(p_i)$. The constant terms C_1, C_2 and C_3 are those previously defined. The test statistic for hypothesis $H_{(A\text{-}4)}$ is

$$(\hat{m}_1(p) - \hat{m}_0(p))'(\tilde{V}_0 + \tilde{V}_1)^{-1}(\hat{m}_1(p) - \hat{m}_0(p)) \sim \chi_1^2.$$

Estimators for $\tilde{V}_1$ and $\tilde{V}_0$ are given below.

To test conditional independence and conditional mean independence of the residuals ($H_{(A\text{-}3')}$ and $H_{(A\text{-}4')}$ respectively), the test statistics are analogous except that Y_{0i} are replaced by estimated residuals $Y_{0i} - X_i\hat{\beta} = \hat{U}_i$. Tests of hypotheses $H_{(A\text{-}5)}$ and $H_{(A\text{-}5')}$, which test the identifying assumptions of conditional difference-in-difference matching estimators, are discussed below in the section on testing with panel data.

Variance estimators

First, we discuss the variance estimator for $\tilde{V}_d$ in the test for tests of the conditional mean independence of earnings. One could estimate each of the components of the variance expression separately and construct a plug-in estimator, but instead we use an estimator proposed in Heckman *et al.* (1996b) that is more accurate at boundary points:

$$\hat{V}_d(p) = \sum_{i \in I_d} \tilde{\varepsilon}_{id}^2 \hat{W}_{id}^2(p),$$

where $\tilde{\varepsilon}_{id} = Y_{0i} - \hat{m}_d(p_i)$ and $\hat{W}_{id}(p)$ are weights equal to

$$\hat{W}_{id}(p) = \frac{G\left(\frac{P_i - p}{a_{N_d}}\right)\sum_{k \in I_d} G\left(\frac{P_k - p}{a_{N_d}}\right)(P_k - p)^2 - G\left(\frac{P_i - p}{a_{N_d}}\right)(P_i - p)\sum_{k \in I_d} G\left(\frac{P_k - p}{a_{N_d}}\right)(P_k - p)}{\sum_{i \in I_d} G\left(\frac{P_i - p}{a_{N_d}}\right)\sum_{k \in I_d} G\left(\frac{P_k - p}{a_{N_d}}\right)(P_k - p)^2 - \left[\sum_{k \in I_d} G\left(\frac{P_k - p}{a_{N_d}}\right)(P_k - p)\right]^2}.$$

The weights $\hat{W}_{id}(p)$ are the same as the local linear regression weights given in the text by equation (11), except that X is replaced by p. For the test of conditional independence at a point y_0 conditional on $P = p$, the estimators for the variance are analogous except that $\tilde{\varepsilon}_{id}$ is replaced by $\tilde{\varepsilon}_{id} = 1(Y_{0i} < y_0) - \hat{F}_d(p_i)$.

For tests of conditional independence of residuals, $\tilde{\varepsilon}_{id}$ is replaced by $\hat{U}_{0i} - \hat{E}(U_{0i} | P = p, D = d)$. For tests of conditional mean independence of residuals $\hat{\varepsilon}_{id} = 1(\hat{U}_{0i} < u_0) - \hat{E}(1(\hat{U}_{0i} < u_0) | P = p, D = d)$. Each of the conditional means is estimated by local linear regression.

Generalization of the test statistics to panel data and tests for conditional difference-in-difference matching ($H_{(A\text{-}5)}$, $H_{(A\text{-}5')}$)

In our empirical work, we use panel data and we conduct tests of conditional independence and conditional mean independence jointly across time periods and across p-points. Let i denote the individual and $t \in \{1, \ldots, T\}$ the time periods in the panel. The generalization of the test for conditional mean independence of earnings ($H_{(A\text{-}4)}$) to one that is joint across time periods is:

$$[\tilde{m}(p)]'(\tilde{V}_0^T + \tilde{V}_1^T)^{-1}[\tilde{m}(p)] \sim \chi_T^2,$$

where $\tilde{m}(p) = [\{\hat{m}_{11}(p) - \hat{m}_{01}(p)\}, \ldots, \{\hat{m}_{1T}(p) - \hat{m}_{0T}(p)\}]'$, with the first subscript denoting the group and the second subscript the time period. $\tilde{V}_d^T$ is the T by T estimators of the variance-covariance matrix for group d. (Estimators of the covariance matrices are proposed below.) If the p-points are at least one bandwidth apart, the chi-squared statistics can be combined across p-points for an overall joint test across time periods and across quarters.

We next discuss tests of the identifying assumptions of conditional difference-in-difference methods (hypotheses $H_{(A\text{-}5)}$ and $H_{(A\text{-}5')}$. Let L denote the restriction matrix that takes differences $\hat{m}_{1t}(p) - \hat{m}_{0t}(p) - [\hat{m}_{1t'}(p) - \hat{m}_{0t'}(p)]$, where t' and t refer to a time period before and after the time of random assignment/eligibility determination, respectively, which are chosen to be symmetric around $t = 0$. For example, to test $H_{(A\text{-}5)}$ over time periods $t \in \{4, 5, 6\}$, define

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \end{bmatrix},$$

and the test statistic is

$$(L \cdot (\tilde{m}(p)))'(L \cdot (\tilde{V}_0^T + \tilde{V}_1^T) \cdot L')^{-1}(L \cdot \tilde{m}(p)) \sim \chi_k^2,$$

where k is the number of restrictions imposed by L. For testing hypothesis $H_{(A\text{-}5')}$, the test statistic is analogous except that Y_i are replaced by estimated residuals $Y_i - X_i\hat{\beta} = \hat{U}_i$.

Variance estimators

For the test of conditional mean independence jointly over all time periods at a point $P=p$, a natural estimator of the variance-covariance matrix is

$$\hat{V}_d^*(p) = \sum_{i \in I_d} \tilde{\varepsilon}_i \tilde{\varepsilon}_i' \hat{W}_{id}^2(p),$$

where $\tilde{\varepsilon}_i = [\tilde{\varepsilon}_{i1}, \ldots, \tilde{\varepsilon}_{iT}]$ and $\hat{W}_{id}(p)$ is the scalar weight defined above. The weights do not vary across time because P_i is constant across time for each individual. However, as discussed in Heckman, Ichimura, Smith and Todd (1996b), this estimator is infeasible when the panel is not balanced, as is the case in our data. If we construct consistsent estimates of each component in the variance-covariance matrix, we cannot guarantee that the resulting estimated covariance matrix will be positive definite. The alternative estimator proposed in Heckman, Ichimura, Smith and Todd (1996b) that we use, which is always guaranteed to be positive semidefinite, is:

$$\hat{V}_d^T(p) = \sum_{i \in I_d} \hat{V}_{id} \hat{V}_{id}',$$

where

$$\hat{V}_{id} = [\tilde{\varepsilon}_{i1} q(i, 1) \hat{W}_{id}(p), \ldots, \tilde{\varepsilon}_{iT} q(i, T) \hat{W}_{id}(p)]',$$

and where $q(i, t) = 1$ if observation i has usable data in period t and $q(i, t) = 0$ otherwise.

B. *Accounting for parameter estimation*

Conditional mean independence tests based on estimated residuals

When tests for conditional mean independence are conducted on estimated residuals, there is an additional source of estimation error. The fitted residuals come from estimating the following model for outcomes in the no treatment state:

$$Y_{0it} = X_{it}\beta + D_t E(U_{0it} | X_{it}, D_i = 1) + (1 - D_i) E(U_{0it} | X_{it}, D_i = 0) + \{U_{0it} - E(U_{0it} | X_{it}, D_i)\},$$

where the term in braces is a mean zero disturbance term by construction. The variables bear an *it* subscript because of the panel nature of the data. β is assumed to be the same across time. The estimation method for this type of partially linear model, which has both a parametric and a nonparametric component, is described in Heckman, Ichimura, Smith and Todd (1996b, c). In estimating this model, we draw on research reported in Heckman, Ichimura, Smith and Todd (1996b) who establish that in our data index sufficiency characterizes the conditional mean function, i.e. that $E(U_{0it} | X_{it}, D_t)$ has an index representation $E(U_{0it} | P(Z_i), D_i)$.

Estimates of the residuals for nonparticipants ($D_i = 0$) and for randomized-out controls ($D_i = 1$) are $\hat{U}_{0it} = Y_{0it} - X_{it}\hat{\beta}$. For the test of conditional mean independence of the residuals ($H_{(A\text{-}4')}$ in the text), the estimation of β should not affect the asymptotic distribution of the text statistic, because $\hat{\beta}$ converges at a faster rate than $\hat{E}(U_{0it} | P = p, D = d)$ (i.e. at rate $\sqrt{N}$ vs. $\sqrt{Na_N}$).

Table B-1 presents evidence on the importance of adjusting for parameter estimation error in constructing test statistics for testing $H_{(A\text{-}4')}$. The p values in the adjacent pairs of columns corresponding to unadjusted and adjusted test statistics differ dramatically. Both are asymptotically equivalent. In our view the truth is somewhere in between. Adjusting for parameter estimation greatly increases the sampling variances and makes it easy to accept any null. The less-than-perfect performance of matching based on Assumption (A − 4′) that is documented in the text makes us wary of relying uncritically on p-values in selecting a model.

C. *Operational definition of common support and a Monte Carlo study of the relationship between sample size, trimming sensitivity and bandwidth sensitivity*

In implementing nonparametric matching estimators, it is necessary to define a region of overlap. Using the notation defined in the text, the region S_{10} over which the supports of P overlap for the $D = 1$ and $D = 0$ groups is the region where $f(P | D = 1) > 0$ and $f(P | D = 0) > 0$. To operationally determine S_{10}, we first estimate the densities at all the sample P values using a kernel density estimator and determine $\hat{S}_{10}$ by forming

$$\hat{S}_{10} = \{P \in \hat{S}_1 \cap \hat{S}_0 : \hat{f}(P | D = 1) > 0 \quad \text{and} \quad \hat{f}(P | D = 0) > 0\},$$

TABLE B-1

*P-values from tests for conditional mean independence of residuals**
(Asymptotic standard errors not adjusted and adjusted for estimation of β)
$(H_{(A\text{-}4')})$: $E(U_0 \mid D=1, P) = E(U_0 \mid D=0, P)$

	Residuals not adjusted		Residuals adjusted		Residuals not adjusted		Residuals adjusted	
P-points	Pre-programme quarters†	Post-programme quarters‡	Pre-programme quarters†	Post-programme quarters‡	Pre-programme quarters†	Post-programme quarters‡	Pre-programme quarters†	Post-programme quarters‡
	Adult males				Adult females			
0.0025	0.0293	0.0002	0.6421	0.0955	0.1472	0.0363	0.8947	0.4683
0.005	0.0815	0.0004	0.7346	0.1129	0.1847	0.0298	0.8850	0.3761
0.01	0.2586	0.0040	0.7773	0.2170	0.3035	0.0140	0.8441	0.1377
0.02	0.4078	0.2056	0.6738	0.5925	0.3836	0.0180	0.6531	0.0532
0.03	0.5680	0.6563	0.7808	0.8289	0.1048	0.1243	0.4789	0.4907
0.04	0.7177	0.9060	0.8726	0.9563	0.0790	0.2570	0.4857	0.9276
0.05	0.8064	0.9885	0.9042	0.9939	0.2718	0.4510	0.6226	0.9325
0.10	0.7456	0.2591	0.9253	0.5790	0.4752	0.8423	0.7448	0.8878
Overall	0.2515	0.0001	1.0000	1.0000	0.0375	0.0008	1.0000	1.0000
	Male youth				Female youth			
0.0025	0.3877	0.2142	0.8999	0.4387	0.1418	0.3047	0.2058	0.4833
0.005	0.3547	0.2911	0.9403	0.5956	0.0770	0.3648	0.1323	0.5715
0.01	0.3373	0.4894	0.9697	0.8440	0.0274	0.5512	0.0668	0.7702
0.02	0.4702	0.6166	0.9616	0.7575	0.0107	0.4803	0.0561	0.7852
0.03	0.6913	0.5214	0.9378	0.5572	0.0277	0.1570	0.1654	0.4659
0.04	0.7342	0.4567	0.8816	0.4970	0.2278	0.0575	0.5479	0.2882
0.05	0.5988	0.4315	0.7717	0.4340	0.6279	0.0245	0.8679	0.2014
0.10	0.0000	0.1944	0.0021	0.6329	0.2412	0.0000	0.6887	0.0024
Overall	0.0015	0.3561	1.0000	1.0000	0.0006	0.0001	1.0000	1.0000

* A fixed bandwidth of 0.06 and biweight kernel, defined in Appendix A, are used in the test. The models for the probability of participation, *P*, are those described in the footnote on Table 2.
† Tests are performed jointly across pre-programme quarters $t=-1$ to $t=-6$.
‡ Tests are performed jointly across post-programme quarters $t=1$ to $t=6$.

where $\hat{S}_1$ and $\hat{S}_0$ are the estimated smoothed supports.[42] The formal theory presented in our companion paper requires that the densities be strictly greater than 0, and in practice matches made at points P where the comparison group density is extremely small are likely to be inaccurate. We therefore require that the points that fall within the smoothed support $\hat{S}_{10}$ have a positive density that exceeds zero by a certain amount, determined by a "trimming level" $\hat{q}$. The set of points potentially eligible for matches are

$$S_q = \{P \in \hat{S}_{10} : \hat{f}(P|D=1) > c_q \quad \text{and} \quad \hat{f}(P|D=0) > c_q\},$$

where c_q satisfies

$$\sup_{c_q} \frac{1}{2J} \sum_{\{i \in \bar{I}_1\}} \{1(\hat{f}(P_i|D=1) < c_q) + 1(\hat{f}(P_i|D=0) < c_q)\} \leqq q,$$

where $\bar{I}_1$ is the set of observed values of P that lie in $\hat{S}_{10}$ and J is the cardinality of $\bar{I}_1$. Actual matches are made at the control P points (i.e. points for which $D=1$) in the original unsmoothed control point set, $\tilde{S}_1$, that lie in S_q, i.e. $\tilde{S}_1 \cap S_q$.

In conducting our empirical analysis we have explored the sensitivity of our estimates to choices of q. Monte Carlo evidence demonstrates the importance of trimming when sample sizes are small as they are for male youth. Smaller samples require higher trimming levels to reduce bias. There is no sensitivity to wide ranges of trimming rules for estimates obtained for adult samples or for female youth. (The empirical results are available on request from the authors).

Acknowledgements. The work reported here is a distant outgrowth of numerous papers and conversations with Ricardo Barros, Bo Honoré, and Richard Robb. We thank Manuel Arellano, three referees and Ed Vytlacil for helpful comments. We thank Annie Zhang and Jingjing Hsee for careful programming assistance. An earlier version of this paper circulated under the title "Evaluating The Impact of Training on the Earnings and Labor Force Status of Young Women: Better Data Help A Lot" (with Rebecca Roselius, 1993) and was presented at the Review of Economic Studies conference on evaluation research in Madrid in September, 1993. This paper was also presented by Heckman in his Harris Lectures at Harvard, November, 1995, at the Latin American Econometric Society meeting in Caracas, Venezuela, August, 1994; at the Rand–UCLA workshop in September, 1994; at UC Riverside, September, 1994; at Princeton, October, 1994; and at Texas, Austin, March, 1996.

REFERENCES

ASHENFELTER, O. (1978), "Estimating the Effect of Training Programs on Earnings", *Review of Economics and Statistics*, **60**, 47–57.

ASHENFELTER, O. and CARD, D. (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs", *Review of Economics and Statistics*, **67**, 648–660.

BARNOW, B. (1993), "Thirty Years of Changing Federal, State, and Local Relationships in Employment and Training Programs", *Publius: The Journal of Federalism*, **23**, 75–94.

BLANCHFLOWER, D. and OSWALD, A. (1994) *The Wage Curve* (Cambridge: MIT Press).

BREIMAN, L., FRIEDMAN, J. H., OHLSEN, R. and STONE, C. J. (1984) *Classification and Regression Trees* (Belmont: Wadsworth International Group).

BRYANT, E. and RUPP, K. (1987), "Evaluating the Impact of CETA on Participant Earnings", *Evaluation Review*, **11**, 473–492.

BURTLESS, G. (1995), "The Case For Randomized Field Trials in Economic and Policy Research", *Journal of Economic Perspectives*, **9**, 63–84.

COCHRANE, W. and RUBIN, D. (1973), "Controlling Bias In Observational Studies", *Sankyha*, **35**, 417–446.

DICKENSON, K., JOHNSON, T. and WEST, R. (1987), "An Analysis of the Sensitivity of Quasi-experimental Net Impact Estimates of CETA Programs", *Evaluation Review*, **11**, 452–472.

DEVINE, T. and HECKMAN, J. (1996), "Consequences of Eligibility Rules for a Social Program: A Study of the Job Training Partnership Act (JTPA)", *Research in Labor Economics*, **15**, 111–170.

FAN, J. (1992), "Design Adaptive Nonparametric Regression", *Journal of the American Statistical Association*, **87**, 998–1004.

FISHER, R. (1951) *The Design of Experiments*, 6th edition (London: Oliver and Boyd).

HECKMAN, J. (1978), "Longitudinal Analysis of Labor Market Data" (Unpublished manuscript presented at SSRC Conference in Mt. Kisco, New York, September, 1978).

HECKMAN, J. (1980), "Addendum To Sample Selection Bias As A Specification Error", *Evaluation Studies Review Annual*, **5**, 69–74.

HECKMAN, J. (1990a), "Alternative Approaches to Evaluating Social Programs", (Barcelona Lecture, Fifth World Meeting of Econometric Society, Barcelona, Spain).

42. We use Silverman's "rule of thumb" bandwidth in estimating the densities. See equation 3.31 in Silverman (1986).

HECKMAN, J. (1990b), "Varieties of Selection Bias", *American Economic Review*, **80**, 313–318.
HECKMAN, J. (1992), "Randomization and Social Program", in Manski, C. and Garfinkel, I. (eds.), *Evaluating Welfare and Training Programs* (Cambridge: Harvard University Press).
HECKMAN, J. (1995), "Evaluating Social Programs: The Harris Lectures" (Unpublished manuscript).
HECKMAN, J. (1996), "Randomization As An Instrumental Variable Estimator", *Review of Economics and Statistics*, **56**, 336–341.
HECKMAN, J. (1997), "Instrumental Variables: A Study of the Implicit Assumptions Underlying One Widely Used Estimator For Program Evaluations", *Journal of Human Resources*, **32**, 441–461.
HECKMAN, J. and HONORÉ, B. (1990), "The Empirical Content of the Roy Model", *Econometrica*, **58**, 1121–1149.
HECKMAN, J. and HOTZ, V. J. (1989), "Choosing Among Alternatives Nonexperimental Methods For Estimating The Impact of Social Programs", *Journal of The American Statistical Association*, **84**, 862–874.
HECKMAN, J., ICHIMURA, H., SMITH, J. and TODD, P. (1996a), "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching As A Program Evaluation Method, *Proceedings of the National Academy of Sciences*, **93**, 13416–13420.
HECKMAN, J., ICHIMURA, H., SMITH, J. and TODD, P. (1994, revised 1996b), "Characterizing Selection Bias Using Experimental Data," *Econometrica* (forthcoming).
HECKMAN, J., ICHIMURA, H., SMITH, J. and TODD, P. (1994, revised 1996c), "Nonparametric Characterization of Selection Bias Using Experimental Data: A Study of Adult Males in JTPA. Part II, Theory and Methods and Monte Carlo Evidence" (Unpublished manuscript, University of Chicago).
HECKMAN, J., ICHIMURA, H. and TODD, P. (1993, revised 1997), "Matching As An Econometric Evaluation Estimator", *Review of Economic Studies* (forthcoming).
HECKMAN, J., ICHIMURA, H. and TODD, P. (1996a), "Characterizing Program Outcomes" (Unpublished manuscript, University of Chicago).
HECKMAN, J., ICHIMURA, H. and TODD, P. (1996b), "Nonparametric Tests for Selection and Matching" (Unpublished manuscript, University of Chicago).
HECKMAN, J. and MACURDY, T. (1985), "Labor Econometrics", in Griliches, Z. and Intriligator, M. (eds.), *Handbook of Econometrics*, (Amsterdam: North Holland).
HECKMAN, J. and ROBB, R. (1985), "Alternative Methods For Evaluating The Impact of Interventions", in Heckman, J. and Singer, B. (eds.), *Longitudinal Analysis of Labor Market Data* (New York: Wiley).
HECKMAN, J. and ROBB, R. (1986), "Alternative Methods For Solving The Problem of Selection Bias in Evaluating The Impact of Treatments on Outcomes", in Wainer, H. (ed.), *Drawing Inferences from Self-Selected Samples* (Berlin: Springer-Verlag).
HECKMAN, J. and ROSELIUS, R. (1993), "Evaluating The Impact of Training on the Earnings and Labor Force Status of Women: Better Data Help A Lot" (Unpublished manuscript, University of Chicago).
HECKMAN, J. and SMITH, J. (1994), "Ashenfelter's Dip and the Determinants of Program Participation" (Mimeo, University of Chicago).
HECKMAN, J. and SMITH, J. (1997), "Evaluating The Welfare State", in Strom, S. (ed.), *Frisch Centenary* (Cambridge: Cambridge University Press).
HECKMAN, J., SMITH, J. and CLEMENTS, N. (1997), "Making The Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts", *Review of Economic Studies*, **64**, 487–535.
HECKMAN, J., SMITH, J. and TABER, C. (1996), "What Do Bureaucrats Do? The Effects of Performance Standards and Bureaucratic Preferences on Acceptance Into the JTPA Program", in Liecap, G. (ed.), *Studies in Bureaucratic Behavior*, (Greenwich: JAI Press), 191–218.
HECKMAN, J., SMITH, J. and TABER, C. (1998), "Accounting for Dropouts in the Evaluation of Social Experiments", *Review of Economics and Statistics* (forthcoming).
KEMPLE, J., DOOLITTLE, F. and WALLACE, J. (1993) *The National JTPA Study: Site Characteristics in Participation Patterns* (New York: Manpower Demonstration Research Corporation).
LALONDE, R. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data", *American Economic Review*, **76**, 604–620.
MANSKI, C. and LERMAN, S. (1977), "The Estimation of Choice Probabilities from Choice-Based Samples", *Econometrica*, **45**, 1977–1988.
ORR, L., BLOOM, H., BELL, S., LIN, W., CAVE, G. and DOOLITTLE, F. (1994) *The National JTPA Study: Impacts, Benefits and Costs of Title II-A* Bethesda: Abt Associates)
QUANDT, R. (1972), "A New Approach to Estimating Switching Regressions", *Journal of the American Statistical Association*, **67**, 306–310.
RAO, C. R. (1965), "On Discrete Distributions Arising Out of Methods of Ascertainment", in Patil, G. P. (ed.), *Classical and Contagious Discrete Distributions* (Calcutta: Statistical Publication Society), 320 333.
RAO, C. R. (1986), "Weighted Distributions", in Feinberg, S. (ed.), *A Celebration of Statistics* (Berlin: Springer-Verlag), 543–569.
ROSENBAUM, P. and RUBIN, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, **70**, 41–55.
ROSENBAUM, P. and RUBIN, D. B. (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score", *American Statistician*, **39**, 38 39.

ROY, A. D. (1951), "Some Thoughts on The Distribution of Earnings", *Oxford Economics Papers*, **3**, 135–146.
RUBIN, D. B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization", *Annals of Statistics*, **7**, 34–58
RUBIN, D. B. (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias In Observational Studies", *Journal of the American Statistical Association*, **74**, 318–329.
SILVERMAN, R. W. (1986) *Density Estimation For Statistics and Data Analysis* (London: Chapman and Hall).
SMITH, J. (1994), "Sampling Frame for the Eligible Non-Participant Sample" (Mimeo, University of Chicago).
SMITH, J. (1995), "A Comparison of the Earnings Patterns of Two Samples of JTPA Eligibles" (Mimeo, University of Chicago).
TODD, P. (1995), "Matching and Local Linear Regression Approaches to Solving the Evaluation Problem with a Semiparametric Propensity Score" (Mimeo, University of Chicago).