Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Journal of Econometrics 125 (2005) 305–353

JOURNAL OF Econometrics

www.elsevier.com/locate/econbase

Does matching overcome LaLonde's critique of nonexperimental estimators?

Jeffrey A. Smith[a,*,1], Petra E. Todd[b,1]

[a] Department of Economics, University of Maryland, 3105 Tydings Hall, College Park, MD 20742-7211, USA

[b] University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, USA

Available online 24 July 2004

Abstract

This paper applies cross-sectional and longitudinal propensity score matching estimators to data from the National Supported Work (NSW) Demonstration that have been previously analyzed by LaLonde (1986) and Dehejia and Wahba (1999, 2002). We find that estimates of the impact of NSW based on propensity score matching are highly sensitive to both the set of variables included in the scores and the particular analysis sample used in the estimation. Among the estimators we study, the difference-in-differences matching estimator performs the best. We attribute its performance to the fact that it eliminates potential sources of temporally invariant bias present in the NSW data, such as geographic mismatch between participants and nonparticipants and the use of a dependent variable measured in different ways for the two groups. Our analysis demonstrates that while propensity score matching is a potentially useful econometric tool, it does not represent a general solution to the evaluation problem.

1. Introduction

There is a long-standing debate in the literature over whether social programs can be reliably evaluated without a randomized experiment. Randomization has a key advantage over nonexperimental methods in generating a control group that has the same distributions of both observed and unobserved characteristics as the treatment group. At the same time, social experimentation also has some drawbacks, such as

*Corresponding author. Tel.: +1-301-405-3532; fax: +1-301-405-3542.

E-mail addresses: smith@econ.umd.edu (J. A. Smith), petra@athena.sas.upenn.edu (P. E. Todd).

[1] Affiliated with the National Bureau of Economic Research (NBER) and the IZA.

0304-4076/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2004.04.011

high cost, the potential to distort the operation of an ongoing program, the common problem of program sites refusing to participate in the experiment and the problem of randomized-out controls seeking alternative forms of treatment.[2] In contrast, evaluation methods that use nonexperimental data tend to be less costly and less intrusive. Also, for some questions of interest, they are the only alternative.[3]

The major obstacle in implementing a nonexperimental evaluation strategy is choosing among the wide variety of estimation methods available in the literature. This choice is important given the accumulated evidence that impact estimates are often highly sensitive to the estimator chosen. A literature has arisen, starting with LaLonde (1986), that evaluates the performance of nonexperimental estimators using experimental data as a benchmark. Much of this literature implicitly frames the question as one of searching for "the" nonexperimental estimator that will always solve the selection bias problem inherent in nonexperimental evaluations. Two recent contributions to this literature by Dehejia and Wahba (DW) (1999, 2002) have drawn attention to a class of estimators called *propensity score matching estimators*. They apply these matching estimators to the same experimental data from the National Supported Work (NSW) Demonstration, and the same nonexperimental data from the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID), analyzed by LaLonde (1986) and find very low biases. Their findings have made propensity score matching the estimator *du jour* in the evaluation literature.

Dehejia and Wahba's (1999, 2002) finding of low bias from applying propensity score matching to LaLonde's (1986) data is surprising in light of the lessons learned from the analyses of Heckman, Ichimura and Todd and Heckman, Ichimura, Smith and Todd (Heckman et al. (1997a), Heckman et al. (1996, 1998a)) (henceforth HIT and HIST) using the experimental data from the U.S. National Job Training Partnership Act (JTPA) Study. They conclude that in order for matching estimators to have low bias, it is important that the data include a rich set of variables related to program participation and labor market outcomes, that the nonexperimental comparison group be drawn from the same local labor markets as the participants, and that the dependent variable (typically earnings) be measured in the same way for participants and nonparticipants. All three of these conditions fail to hold in the NSW data analyzed by LaLonde (1986) and Dehejia and Wahba (1999, 2002).

In this paper, we reanalyze these data, applying both cross-sectional and longitudinal variants of propensity score matching. We find that the low bias estimates obtained by DW (1999, 2002) using various cross-sectional matching estimators are highly sensitive to their choice of a particular subsample of LaLonde's (1986) data for their analysis. We also find the changing the set of variables used to

[2] On these points see, e.g., Burtless and Orr (1986), Heckman (1992), Burtless (1995), Heckman and Smith (1995), Heckman et al. (1999) and Heckman et al. (2000).

[3] For example, Eberwein et al. (1997) analyze the effects of a job training program on employment probabilities and on the lengths of employment spells. Experimental data do not solve the selection problem that arises when comparing spells for program participants and nonparticipants at points in time after leaving the program. Solving this selection problem requires application of nonexperimental evaluation methods.

estimate the propensity scores strongly affects the estimated bias in LaLonde's original sample. At the same time, we find that difference-in-differences (DID) matching estimators exhibit better performance than the cross-sectional estimators. This is consistent with the evidence from the JTPA data in HIT (1997a) and HIST (1998a) on the importance of avoiding geographic mismatch and of measuring the dependent variable in the same way in the treatment and comparison groups. Both these sources of bias are likely to be relatively stable over time, and so should difference out. More generally, our findings make it clear that propensity score matching does not represent a "magic bullet" that solves the selection problem in every context. The implicit search for such an estimator in the literature cannot succeed. Instead, the optimal nonexperimental evaluation strategy in a given context depends critically on the available data and on the institutions governing selection into the program.

The plan of the paper is as follows. Section 2 reviews some key papers in the previous literature on the choice among alternative nonexperimental estimators. Section 3.1 lays out the evaluation problem and Section 3.2 briefly describes commonly used nonexperimental estimators. Section 3.3 describes the cross-sectional and difference-in-differences matching estimators that we focus on in our study. Sections 3.4 and 3.5 briefly address the issues of choice-based sampling and the bias that arises from incomplete matching, respectively. Section 3.6 explains how we use the experimental data to benchmark the performance of nonexperimental estimators. Section 4 describes the NSW program. Section 5 describes our analysis samples from the NSW data and the two comparison groups. Section 6 presents our estimated propensity scores and Section 7 discusses the "balancing tests" used in some recent studies to aid in selecting a propensity score specification. Sections 8 and 9 give the bias estimates obtained using matching and regression-based estimators, respectively. Section 10 displays evidence on the use of specification tests applied to our cross-sectional matching estimators and Section 11 concludes.

2. Previous research

Several previous papers use data from the National Supported Work Demonstration experiment to study the performance of econometric estimators. LaLonde (1986) was the first and the data we use come from his study. He arranged the NSW data into two samples: one of AFDC women and one of disadvantaged men. The comparison group subsamples were constructed from two national survey datasets: the CPS and the PSID. LaLonde (1986) applies a number of standard evaluation estimators, including simple regression adjustment, difference-in-differences, and the two-step version of the bivariate normal selection model in Heckman (1979). His findings show that alternative estimators produce very different estimates, most of which deviate substantially from the experimental benchmark impacts. This is not necessarily surprising, given that the different estimators depend on different assumptions about the nature of the outcome and program participation processes. Unless there is no selection problem, at most one set of assumptions will be satisfied in the data. Using a limited set of specification tests, LaLonde (1986) concludes that no good way exists to sort among the competing estimators and, hence, that nonexperimental methods do not provide an effective means of evaluating programs. His paper played an important role in the late 1980s movement towards using experiments to evaluate social programs (see, e.g., Burtless and Orr, 1986; Burtless, 1995). Fraker and Maynard (1987) perform a similar analysis that focuses more on comparison group selection than LaLonde (1986) and reach similar conclusions.

Heckman and Hotz (1989) respond to the LaLonde (1986) study by applying a broader range of specification tests to guide the choice among nonexperimental estimators.[4] The primary test they consider is based on pre-program data, so its validity depends on the assumption that the outcome and participation processes are similar in pre-program and post-program time periods. Heckman and Hotz (1989) find that the tests they apply to the NSW data exclude the estimators that would imply a substantially different qualitative conclusion (impact sign and statistical significance) than the experiment.[5]

In the more recent evaluation literature, researchers have focused on matching estimators, which were not considered by LaLonde (1986) or Heckman and Hotz (1989). Unlike some of the early studies evaluating the Comprehensive Employment and Training Act (JTPA's predecessor) surveyed in Barnow (1987), which used variants of matching, the recent literature focuses on matching on the probability of participating in the program. This technique, introduced in Rosenbaum and Rubin (1983), is called propensity score matching. Traditional propensity score matching methods pair each program participant with a single nonparticipant, where pairs are chosen based on the degree of similarity in the estimated probabilities of participating in the program (the propensity scores). The mean impact of the program is estimated by the mean difference in the outcomes of the matched pairs.

HIT (1997a, 1998b) and HIST (1998a) extend traditional pairwise matching methods in several ways. First, they describe kernel and local linear matching estimators that use multiple nonparticipants in constructing the estimated counterfactual outcome. The main advantage of these estimators vis-a-vis pairwise matching is a reduction in the asymptotic mean squared error. Second, HIT (1997a) and HIST (1998a) propose modified versions of matching estimators that can be implemented when longitudinal or repeated cross-section data are available. These estimators

[4] Heckman and Hotz (1989) make use of somewhat different data from the NSW experiment than LaLonde does. Their two samples consist of female AFDC recipients, as in LaLonde, and young high school dropouts, most but not all of whom are men. They do not make use of the ex-convict and ex-addict samples. In addition, they use grouped earnings data from Social Security earnings records for both the NSW samples and the comparison groups, while LaLonde uses individual level Social Security earnings records for the CPS comparison group and survey-based earnings measures for the NSW sample and for the PSID comparison group. Because their administrative data do not suffer from attrition problems, the sample of AFDC women used in Heckman and Hotz (1989) is substantially larger than that used in LaLonde (1986).

[5] These tests have also been applied in an evaluation context by, among others, Ashenfelter (1978), Bassi (1984), LaLonde (1986), Friedlander and Robins (1995), Regnér (2002) and Raaum and Torp (2002).

eliminate time-invariant differences in outcomes between participants and nonparticipants that cross-sectional matching fails to eliminate.

HIT (1997a) and HIST (1998a) evaluate the performance of both the traditional pairwise matching estimators and cross-sectional and longitudinal versions of their kernel and local linear matching estimators using experimental data from the U.S. National JTPA Study combined with comparison group samples drawn from three sources. They show that data quality is a crucial ingredient to any reliable estimation strategy. Specifically, the estimators examined are only found to perform well in replicating the results of the experiment when they are applied to comparison group data satisfying the following criteria: (i) the same data sources (i.e., the same surveys or the same type of administrative data or both) are used for participants and nonparticipants, so that earnings and other characteristics are measured in an analogous way, (ii) participants and nonparticipants reside in the same local labor markets, and (iii) the data contain a rich set of variables that affect both program participation and labor market outcomes. If the comparison group data fail to satisfy these criteria, the performance of the estimators diminishes greatly. Based on this evidence, HIT (1997a) and HIST (1998a) hypothesize that data quality probably accounts for much of the poor performance of the estimators in LaLonde's (1986) study, where participant and nonparticipant samples were located in different local labor markets, the data were collected using a combination of different survey instruments and administrative data sources and the data contain only very limited information on observable characteristics.

More recently, DW (1999, 2002) use the NSW data to evaluate the performance of propensity score matching methods, including pairwise matching and caliper matching (see Section 3.3 for detailed descriptions). They find that these simple matching estimators succeed in closely replicating the experimental NSW results, even through the comparison group data do not satisfy any of the criteria found to be important in HIT (1997a) and HIST (1998a). DW (1999, 2002) are now widely cited in the empirical literature as showing that propensity score matching solves the selection problem.

In this paper, we use the same NSW data employed by DW (1999, 2002) to evaluate the performance of both traditional, pairwise matching methods and of the newer methods developed in HIT (1997a, 1998b) and HIST (1998a). We find that a major difference between the DW (1999, 2002) studies and the LaLonde (1986) study is that DW exclude about 40 percent of the observations used in LaLonde's (1986) study in order to incorporate one additional variable into their propensity score model. As we show below, this restriction makes a tremendous difference to their results, as it has the effect of eliminating many of the higher earners from the sample. Eliminating participants with high pre-program earnings attenuates the pre-program "dip" and thereby makes the selection problem easier to solve. In fact, almost any conventional evaluation estimator applied to the smaller DW samples exhibits lower bias than when applied to the full LaLonde (1986) sample. When we apply cross-sectional matching estimators to either the full LaLonde (1986) sample or an alternative subsample of persons randomly assigned early in the experiment, we find large biases. Similarly, changing to an alternative propensity score specification also increases the estimated bias. Consistent with the likely sources of bias in the NSW data, we find that difference-in-differences matching estimators developed in HIT (1997a) and HIST (1998a) perform better than cross-sectional matching estimators for both comparison groups.

3. Methodology

3.1. The evaluation problem

Assessing the impact of any intervention requires making an inference about the outcomes that would have been observed for program participants had they not participated. Denote by Y_1 the outcome conditional on participation and by Y_0 the outcome conditional on non-participation, so that the impact of participating in the program is

$$\Delta = Y_1 - Y_0.$$

For each person, only Y_1 or Y_0 is observed, so Δ is not observed for anyone. This missing data problem lies at the heart of the evaluation problem.

Let $D = 1$ for the group of individuals who applied and got accepted into the program for whom Y_1 is observed. Let $D = 0$ for persons who do not enter the program for whom Y_0 is observed. Let X denote a vector of observed individual characteristics used as conditioning variables. The most common evaluation parameter of interest is the *mean impact of treatment on the treated*,[6]

$$\begin{aligned} TT &= \mathrm{E}(\Delta | X, D = 1) = \mathrm{E}(Y_1 - Y_0 | X, D = 1) \\ &= \mathrm{E}(Y_1 | X, D = 1) - \mathrm{E}(Y_0 | X, D = 1), \end{aligned} \tag{1}$$

which estimates the average impact of the program among those participating in it. It is the parameter on which LaLonde (1986) and DW (1999, 2002) focus.[7] When Y represents earnings, a comparison of the mean impact of treatment on the treated with the average per-participant cost of the program indicates whether or not the program's benefits outweigh its costs, which is a key question of interest in many evaluations.

Most experiments are designed to provide evidence on the treatment-on-the-treated parameter. Data on program participants identifies the mean outcome in the treated state, $\mathrm{E}(Y_1 | X, D = 1)$, and the randomized-out control group provides a direct estimate of $\mathrm{E}(Y_0 | X, D = 1)$. In nonexperimental (or observational) studies, no direct estimate of this counterfactual mean is available. Instead, the econometrically adjusted outcomes of the nonparticipants proxy for the missing counterfactual. Selection bias, or evaluation bias, consists of the difference between the adjusted

[6] Following the literature, we use "treatment" and "participation" interchangeably throughout.

[7] However, many other parameters may be of interest in an evaluation. See, e.g., Eberwein et al. (1997), Heckman et al. (1997b), Heckman et al. (1999), Heckman (2001), Heckman et al. (2001) and Heckman and Vytlacil (2001) for discussions of other parameters of interest.

outcomes of the nonparticipants and the desired counterfactual mean. In the next section, we discuss common approaches for estimating the missing counterfactual mean. We apply these approaches to the NSW data in Section 9.

3.2. Three commonly used nonexperimental estimators

Nonexperimental estimators use two types of data to impute counterfactual outcomes for program participants: data on participants prior to entering the program and data on nonparticipants. Three common evaluation estimators are the *before–after*, *cross-section* and *difference-in-differences* estimators. We next describe the estimators and their assumptions.

Assume that outcome measures Y_{1it} and Y_{0it}, where i denotes the individual and t the time period, can be represented by

$$\begin{aligned} Y_{1it} &= \varphi_1(X_{it}) + U_{1it}, \\ Y_{0it} &= \varphi_0(X_{it}) + U_{0it}, \end{aligned} \tag{2}$$

where U_{1it} and U_{0it} are distributed independently across persons and satisfy $\mathrm{E}(U_{1it}) = 0$ and $\mathrm{E}(U_{0it}) = 0$. The observed outcome is $Y_{it} = D_i Y_{1it} + (1 - D_i) Y_{0it}$, which can be written as

$$Y_{it} = \varphi_0(X_{it}) + D_i \alpha^*(X_{it}) + U_{0it}, \tag{3}$$

where $\alpha^*(X_{it}) = \varphi_1(X_{it}) - \varphi_0(X_{it}) + U_{1it} - U_{0it}$ is the treatment impact. This is a random coefficient model because the impact of treatment varies across persons even conditional on X_{it}. Assuming that $U_{0it} = U_{1it} = U_{it}$, so that the unobservable is the same in the treated and untreated states, and assuming that $\varphi_1(X_{it}) - \varphi_0(X_{it})$ is constant with respect to X_{it}, yields the fixed coefficient or "common effect" version of the model often used in empirical work.

Before–after estimators: A before–after estimator uses pre-program data to impute counterfactual outcomes for program participants. To simplify notation, assume that the treatment impact α^* is constant across individuals. Let t' and t denote time periods before and after the program start date. The before–after estimator of the program impact is the least-squares solution ($\hat{\alpha}_{\mathrm{BA}}$) to α^* in

$$Y_{it} - Y_{it'} = \varphi(X_{it}) - \varphi(X_{it'}) + \alpha^* + U_{it} - U_{it'}.$$

For $\hat{\alpha}_{\mathrm{BA}}$ to be a consistent estimator, we require that $\mathrm{E}(U_{it} - U_{it'}) = 0$ and $\mathrm{E}((U_{it} - U_{it'})(\varphi(X_{it}) - \varphi(X_{it'}))) = 0$. A special case where this assumption would be satisfied occurs when $U_{it} = f_i + v_{it}$, where f_i depends on i but does not vary over time and v_{it} is a random error term (i.e., U_{it} satisfies a fixed effect assumption).

A drawback of a before–after estimation strategy is that identification of α^* breaks down in the presence of time-specific intercepts.[8] Before–after estimates can also be sensitive to the choice of base time period due to "Ashenfelter's dip", the commonly observed pattern that the mean earnings of program participants decline during the

[8] Suppose $\varphi(X_{it}) = X_{it}\beta + \gamma_t$, where γ_t is a time-specific intercept common across individuals. Such a common time effect may arise, for example, from life-cycle wage growth or from the business cycle. In this example, α^* is confounded with $\gamma_t - \gamma_{t'}$.

period just prior to participation. See the discussions in Ashenfelter (1978), Heckman and Smith (1999) and Heckman et al. (1999).

Cross-section estimators: A cross-section estimator uses data on $D = 0$ persons in a single time period to impute the counterfactual outcomes for $D = 1$ persons in the same time period. Define $\hat{\alpha}_{\mathrm{CS}}$ as the ordinary least-squares solution to α^* in

$$Y_{it} = \varphi(X_{it}) + D_i\alpha^* + U_{it}. \tag{4}$$

Bias for α^* arises if $\mathrm{E}(U_{it}D_i) \neq 0$ or $\mathrm{E}(U_{it}\varphi(X_{it})) \neq 0$.

Difference-in-differences estimators: A difference-in-differences (DID) estimator measures the impact of the program by the difference between participants and nonparticipants in the before–after difference in outcomes. It uses both pre- and post-program data (t and t' data) on $D = 1$ and $D = 0$ observations. The difference-in-differences estimator $\hat{\alpha}_{\mathrm{D}}$ corresponds to the least-squares solution for α^* in

$$Y_{it} - Y_{it'} = \varphi(X_{it}) - \varphi(X_{it'}) + D_i\alpha^* + \{U_{it} - U_{it'}\}. \tag{5}$$

This estimator addresses one shortcoming of the before–after estimator in that it allows for time-specific intercepts that are common across groups.[9] The estimator requires that $\mathrm{E}(U_{it} - U_{it'}) = 0$, $\mathrm{E}((U_{it} - U_{it'})D_i) = 0$ and $\mathrm{E}((U_{it} - U_{it'})\{\varphi(X_{it}) - \varphi(X_{it'})\}) = 0$. LaLonde (1986) implements both the standard estimator just described and an "unrestricted" version that includes $Y_{it'}$ as a right-hand-side variable. The latter estimator relaxes the implicit restriction in the standard DID estimator that the coefficient associated with lagged $Y_{it'}$ equals 1.

3.3. Matching methods

Traditional matching estimators pair each program participant with an observably similar nonparticipant and interpret the difference in their outcomes as the effect of the program (see, e.g., Rosenbaum and Rubin, 1983). Matching estimators are justified by the assumption that outcomes are independent of program participation conditional on a set of observable characteristics. That is, matching assumes that there exists a set of observable conditioning variables Z (which may be a subset or a superset of X) for which the nonparticipation outcome Y_0 is independent of participation status D conditional on Z,[10]

$$Y_0 \perp\!\!\!\perp D \mid Z. \tag{6}$$

It is also assumed that for all Z there is a positive probability of either participating ($D = 1$) or not participating ($D = 0$), i.e.

$$0 < \Pr(D = 1 \mid Z) < 1. \tag{7}$$

[9] To see this, suppose that $Y_{it} = \gamma_t + D_i\alpha^* + U_{it}$ and that $Y_{it'} = \gamma_{t'} + U_{it'}$. Then $Y_{it} - Y_{it'} = (\gamma_t - \gamma_{t'}) + D_i\alpha^* + \{U_{it} - U_{it'}\}$, where the difference in the time-specific intercepts, $(\gamma_t - \gamma_{t'})$, becomes the intercept in the difference equation. In contrast to the before–after estimator, in this case $(\gamma_t - \gamma_{t'})$ and α^* are separately identified, because the $D = 0$ observations, which are not used in the before–after estimator, identify $(\gamma_t - \gamma_{t'})$.

[10] In the terminology of Rosenbaum and Rubin (1983) treatment assignment is "strictly ignorable" given Z.

This assumption implies that a match can be found for all $D = 1$ persons. If assumptions (6) and (7) are satisfied, then, after conditioning on Z, the Y_0 distribution observed for the matched nonparticipant group can be substituted for the missing Y_0 distribution for participants.

Assumption (6) is overly strong if the parameter of interest is the mean impact of treatment on the treated (TT), in which case conditional mean independence suffices:

$$\mathrm{E}(Y_0|Z, D = 1) = \mathrm{E}(Y_0|Z, D = 0) = \mathrm{E}(Y_0|Z). \tag{8}$$

Furthermore, when TT is the parameter of interest, the condition $0<\Pr(D = 1|Z)$ is also not required, because that condition only guarantees the possibility of a participant analogue for each nonparticipant. The TT parameter requires only the possibility of a nonparticipant analogue for each participant. For completeness, the required condition is

$$\Pr(D = 1|Z)<1. \tag{9}$$

Under these assumptions—either (6) and (7) or (8) and (9)—the mean impact of treatment on the treated can be written as

$$\begin{aligned} TT &= \mathrm{E}(Y_1 - Y_0|D = 1) \\ &= \mathrm{E}(Y_1|D = 1) - \mathrm{E}_{Z|D=1}\{\mathrm{E}_Y(Y_0|D = 1, Z)\} \\ &= \mathrm{E}(Y_1|D = 1) - \mathrm{E}_{Z|D=1}\{\mathrm{E}_Y(Y_0|D = 0, Z)\}, \end{aligned}$$

where the first term can be estimated from the treatment group and the second term from the mean outcomes of the matched (on Z) comparison group.

In a social experiment, (6) and (7) are satisfied by virtue of random assignment of treatment. For nonexperimental data, there may or may not exist a set of observed conditioning variables for which the conditions hold. A finding of HIT (1997a) and HIST (1996, 1998a) in their application of matching methods to the JTPA data and of DW (1999, 2002) in their application to the NSW data is that (9) was not satisfied, meaning that for a fraction of program participants no match could be found. If there are regions where the support of Z does not overlap for the $D = 1$ and 0 groups, then matching is only justified when performed over the *common support region*.[11] The estimated treatment effect must then be redefined as the treatment impact for program participants whose propensity scores lie within the common support region.

3.3.1. Reducing the dimensionality of the conditioning problem

Matching may be difficult to implement when the set of conditioning variables Z is large.[12] Rosenbaum and Rubin (1983) prove a result that is useful in reducing the dimension of the conditioning problem in implementing matching methods. They

[11] One advantage of experiments noted by Heckman (1997), as well as HIT (1997a) and HIST (1998a), is that they guarantee that the treated and untreated individuals have the same support. This allows estimation of the mean impact of the treatment over the entire support.

[12] If Z is discrete, small (or empty) cell problems may arise. If Z is continuous and the conditional mean $\mathrm{E}(Y_1|D = 0, Z)$ is estimated nonparametrically, then convergence rates will be slow due to the "curse of dimensionality" problem.

show that for random variables Y and Z and a discrete random variable D

$$\mathrm{E}(D|Y, \Pr(D=1|Z)) = \mathrm{E}(\mathrm{E}(D|Y,Z)|Y, \Pr(D=1|Z)),$$

so that $\mathrm{E}(D|Y,Z) = \mathrm{E}(D|Z) = \Pr(D=1|Z)$ implies $\mathrm{E}(D|Y, \Pr(D=1|Z)) = \mathrm{E}(D|\Pr(D=1|Z))$. This implies that when Y_0 outcomes are independent of program participation conditional on Z, they are also independent of participation conditional on the propensity score, $\Pr(D=1|Z)$. Provided that the conditional participation probability can be estimated using a parametric method, such as a logit or probit model, or semi-parametrically using a method that converges faster than the nonparametric rate, the dimensionality of the matching problem is reduced by matching on the univariate propensity score. If the propensity score must be estimated nonparametrically, then the curse of dimensionality reappears in the estimation of the propensity score. This potential for reducing the dimensionality of the problem has led much of the recent evaluation literature on matching to focus on propensity score matching methods.[13]

Propensity score matching combines groups with different values of Z but the same values of $\Pr(D=1|Z)$. To see why this works, consider two groups, one with $Z=Z_1$ and the other with $Z=Z_2$, but where $\Pr(D=1|Z=Z_1) = \Pr(D=1|Z=Z_2)$. Combining these groups in the matching works because they will have the same relative proportions in the $D=0$ and 1 populations precisely because they have the same probability of participation. As a result, any difference between $E(Y_0|Z=Z_1)$ and $E(Y_0|Z=Z_2)$ in the two groups differences out when calculating $\mathrm{E}(Y_1|D=1, P(Z)) - \mathrm{E}(Y_0|D=0, P(Z))$.[14]

3.3.2. Matching estimators

For notational simplicity, let $P = \Pr(D=1|Z)$. A typical matching estimator takes the form

$$\hat{\alpha}_{\mathrm{M}} = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} [Y_{1i} - \hat{\mathrm{E}}(Y_{0i}|D_i = 1, P_i)], \tag{10}$$

where

$$\hat{\mathrm{E}}(Y_{0i}|D_i = 1, P_i) = \sum_{j \in I_0} W(i,j) Y_{0j},$$

and where I_1 denotes the set of program participants, I_0 the set of nonparticipants, S_P the region of common support (see below for ways of constructing this set), and n_1 the number of persons in the set $I_1 \cap S_P$. The match for each participant $i \in I_1 \cap S_P$ is constructed as a weighted average over the outcomes of nonparticipants, where the weights $W(i,j)$ depend on the distance between P_i and P_j.

[13] HIT (1998b), Hahn (1998) and Angrist and Hahn (1999) consider whether it is better in terms of efficiency to match on $P(Z)$ or on Z directly. For the TT parameter, neither is necessarily more efficient than the other. If the treatment effect is constant, then it is more efficient to condition on the propensity score.

[14] See, e.g., Zhao (2004) for a discussion of dimension reduction methods other than propensity score matching, with an application to the NSW data.

Define a neighborhood $C(P_i)$ for each i in the participant sample. Neighbors for i are nonparticipants $j \in I_0$ for whom $P_j \in C(P_i)$. The persons matched to i are those people in set A_i where $A_i = \{j \in I_0 \mid P_j \in C(P_i)\}$. Alternative matching estimators (discussed below) differ in how the neighborhood is defined and in how the weights $W(i,j)$ are constructed.

Nearest-neighbor matching: Traditional, pairwise matching, also called *single nearest-neighbor matching without replacement*, sets

$$C(P_i) = \min_j \|P_i - P_j\|, \quad j \in I_0.$$

That is, the nonparticipant with the value of P_j that is closest to P_i is selected as the match and A_i is a singleton set. This estimator is often used in practice due to its ease of implementation. Traditional applications of this estimator typically did not impose any common support condition and matched without replacement, so that each $D = 0$ observation could serve as the match for at most one $D = 1$ observation. In our empirical work we implement this method with both a single nearest neighbor and with the ten nearest neighbors. Each nearest neighbor receives equal weight in constructing the counterfactual mean when using multiple nearest neighbors. The latter form of the estimator trades reduced variance (resulting from using more information to construct the counterfactual for each participant) for increased bias (resulting from using, on average, poorer matches). We also match with replacement, which allows a given nonparticipant to get matched to more than one participant. Matching with replacement also involves a tradeoff between bias and variance. Allowing replacement increases the average quality of the matches (assuming some re-use occurs), but reduces the number of distinct nonparticipant observations used to construct the counterfactual mean, thereby increasing the variance of the estimator. DW (2002) show very clearly that matching without replacement in contexts such as the NSW data, where there are many participants with high values of P_i and few nonparticipants with such values, results in many bad matches, in the sense that many participants get matched to nonparticipants with very different propensity scores. More generally, nearest neighbor matching without replacement has the additional defect that the estimate depends on the order in which the observations get matched.

Caliper matching: Caliper matching (Cochran and Rubin, 1973) is a variant of nearest neighbor matching that attempts to avoid “bad” matches (those for which P_j is far from P_i) by imposing a tolerance on the maximum distance $\|P_i - P_j\|$ allowed. That is, a match for person i is selected only if $\|P_i - P_j\| < \varepsilon, j \in I_0$, where ε is a pre-specified tolerance. For caliper matching, the neighborhood is $C(P_i) = \{P_j \mid \|P_i - P_j\| < \varepsilon\}$. Treated persons for whom no matches can be found within the caliper are excluded from the analysis. Thus, caliper matching is one way of imposing a common support condition. A drawback of caliper matching is that it is difficult to know a priori what choice for the tolerance level is reasonable. DW (2002) employ a variant of caliper matching called “radius matching.” In their variant, the counterfactual consists of the mean outcome of all the comparison group members within the caliper, rather than just the nearest neighbor.[15]

Stratification or interval matching: In this variant of matching, the common support of P is partitioned into a set of intervals. Within each interval, a separate impact is calculated by taking the mean difference in outcomes between the $D=1$ and $D=0$ observations within the interval. A weighted average of the interval impact estimates, using the fraction of the $D=1$ population in each interval for the weights, provides an overall impact estimate. DW (1999) implement interval matching using intervals that are selected such that the mean values of the estimated P_i's and P_j's are not statistically different within each interval.

Kernel and local linear matching: Recently developed nonparametric matching estimators construct a match for each program participant using a kernel-weighted average over multiple persons in the comparison group. Consider, for example, the *kernel matching estimator* described in HIT (1997a, 1998b) and HIST (1998a), given by

$$\hat{\alpha}_{\mathrm{KM}} = \frac{1}{n_1} \sum_{i \in I_1} \left\{ Y_{1i} - \frac{\sum_{j \in I_0} Y_{0j} G(\frac{P_j - P_i}{a_n})}{\sum_{k \in I_0} G(\frac{P_k - P_i}{a_n})} \right\}.$$

where $G(\cdot)$ is a kernel function and a_n is a bandwidth parameter. In terms of Eq. (10), the weighting function, $W(i,j)$, equals $G((P_j - P_i)/a_n)/\sum_{k \in I_0} G((P_k - P_i)/a_n)$. The neighborhood $C(P_i)$ depends on the specific kernel function chosen for the analysis. For example, for a kernel function that takes on non-zero values only on interval $(-1,1)$, the neighborhood is $C(P_i) = \{|\frac{P_i - P_j}{a_n}| \leqslant 1\}$, $j \in I_0$. Under standard conditions on the bandwidth and kernel, $\frac{\sum_{j \in I_0} Y_{0j} G(\frac{P_j - P_i}{a_n})}{\sum_{k \in I_0} G(\frac{P_k - P_i}{a_n})}$ is a consistent estimator of $\mathrm{E}(Y_0 | D=1, P_i)$.[16]

In this paper, we implement a generalized version of kernel matching, called local linear matching. Research by Fan (1992a, b) demonstrates several advantages of local linear estimation over more standard kernel estimation methods.[17] The local linear weighting function is given by

$$W(i,j) = \frac{G_{ij} \sum_{k \in I_0} G_{ik}(P_k - P_i)^2 - [G_{ij}(P_j - P_i)][\sum_{k \in I_0} G_{ik}(P_k - P_i)]}{\sum_{j \in I_0} G_{ij} \sum_{k \in I_0} G_{ij}(P_k - P_i)^2 - (\sum_{k \in I_0} G_{ik}(P_k - P_i))^2}, \tag{11}$$

where $G_{ij} = G((P_j - P_i)/a_n)$.

Kernel matching can be thought of as a weighted regression of Y_{0j} on an intercept with weights given by the kernel weights, $W(i.j)$, that vary with the point of

[15] In addition, if there are no comparison group members within the caliper, they use the single nearest neighbor outside the caliper as the match rather than dropping the corresponding participant observation from the analysis.

[16] We assume that $G(\cdot)$ has a mean of zero and integrates to one and that $a_n \to 0$ as $n \to \infty$ and $na_n \to \infty$. In estimation, we use the quartic kernel function, $G(s) = \frac{15}{16}(s^2-1)^2$ for $|s| \leqslant 1$, else $G(s) = 0$.

[17] These advantages include a faster rate of convergence near boundary points and greater robustness to different data design densities. See Fan (1992a, b).

evaluation. The weights depend on the distance between each comparison group observation and the participant observation for which the counterfactual is being constructed. The estimated intercept provides the estimate of the counterfactual mean. Local linear matching differs from kernel matching in that it includes in addition to the intercept a linear term in P_i. Inclusion of the linear term is helpful whenever comparison group observations are distributed asymmetrically around the participant observations, as would be the case at a boundary point of P or at any point where there are gaps in the distribution of P.[18]

Trimming to determine the support region: To implement the matching estimator given by Eq. (10), the region of common support S_P needs to be determined. By definition, the region of common support includes only those values of P that have positive density within both the $D=1$ and 0 distributions. The common support region can be determined by

$$\hat{S}_P = \{P : \hat{f}(P|D=1) > 0 \text{ and } \hat{f}(P|D=0) > 0\},$$

where $\hat{f}(P|D=d)$, $d \in \{0,1\}$ are nonparametric density estimators given by $\hat{f}(P|D=d) = \sum_{k \in I_d} G((P_k - P)/a_n)$.[19] To ensure that the densities are strictly greater than zero, we require that the densities be strictly positive and exceed zero by a threshold amount determined by a "trimming level" q. After excluding any P points for which the estimated density is exactly zero, we exclude an additional q percent of the remaining P points for which the estimated density is positive but very low. The set of eligible matches are therefore given by

$$\hat{S}_q = \{P \in I_1 \cap \hat{S}_P : \hat{f}(P|D=1) > c_q \text{ and } \hat{f}(P|D=0) > c_q\},$$

where c_q is the density cut-off trimming level.[20]

HIST (1998a) and HIT (1997a) also implement a variant of local linear matching which they call "regression-adjusted matching". In this variant, the residual from a regression of Y_{0j} on a vector of exogenous covariates replaces Y_{0j} as the dependent variable in the matching. Regression adjustment can, in principal, be applied in combination with any of the other matching estimators; we apply it in combination with the local linear estimator in Section 8 below.

Difference-in-differences matching: The estimators described above assume that after conditioning on a set of observable characteristics, mean outcomes are conditionally mean independent of program participation. However, there may be

[18] See Fan and Gijbels (1996) for detailed discussion of the distinction between standard kernel regression and local linear regression methods and Frölich (2004) for a Monte Carlo analysis of alternative matching methods.

[19] In implementation, we select a fixed, global bandwidth parameter using Silverman's (1986) rule-of-thumb method.

[20] The qth quantile, c_q, is determined by solving for

$$\sup_{c_q} \frac{1}{2J} \sum_{\{i \in I_1 \cap \hat{S}_P\}} \{1(\hat{f}(P|D=1) < c_q) + 1(\hat{f}(P|D=0) < c_q)\} \leqslant q,$$

where J is the number of observed values of P that lie in $I_1 \cap \hat{S}_P$. Matches are constructed only for the program participants whose propensity scores lie in $\hat{S}_q$. In our empirical work, we set the trimming level at 2 percent.

systematic differences between participant and nonparticipant outcomes even after conditioning on observables. Such differences may arise, for example, (i) because of selection into the program based on unmeasured characteristics, (ii) because of differences in earnings levels among the labor markets in which the participants and nonparticipants reside, or (iii) because earnings outcomes for participants and nonparticipants are measured in different ways (as when data are collected using different survey instruments). Such differences violate the identification conditions required for matching.

A difference-in-differences (DID) matching strategy, as defined in HIT (1997a) and HIST (1998a), allows for temporally invariant differences in outcomes between participants and nonparticipants. This type of estimator is analogous to the standard DID regression estimator defined in Section 3.2, but it does not impose the linear functional form restriction in estimating the conditional expectation of the outcome variable and it reweights the observations according to the weighting functions used by the matching estimators. The DID propensity score matching estimator requires that

$$\mathrm{E}(Y_{0t} - Y_{0t'}|P, D = 1) = \mathrm{E}(Y_{0t} - Y_{0t'}|P, D = 0),$$

where t and t' are time periods after and before the program start date, respectively. This estimator also requires the support condition given in (7) or (9), which must hold in both periods t and t' (a nontrivial assumption given the attrition present in many panel data sets). The difference-in-differences matching estimator is given by

$$\hat{\alpha}_{\mathrm{DDM}} = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} \left\{ (Y_{1ti} - Y_{0t'i}) - \sum_{j \in I_0 \cap S_P} W(i,j)(Y_{0tj} - Y_{0t'j}) \right\},$$

where the weights depend on the particular cross-sectional matching estimator employed. If repeated cross-section data are available, instead of longitudinal data, the estimator can be implemented as

$$\hat{\alpha}_{\mathrm{DDM}} = \frac{1}{n_{1t}} \sum_{i \in I_{1t} \cap S_P} \left\{ (Y_{1ti} - \sum_{j \in I_{0t} \cap S_P} W(i,j) Y_{0tj} \right\} - \frac{1}{n_{1t'}} \sum_{i \in I_{1t'} \cap S_P} \left\{ (Y_{0t'i} - \sum_{j \in I_{0t'} \cap S_p} W(i,j) Y_{0t'j} \right\},$$

where $I_{1t}, I_{1t'}, I_{0t}, I_{0t'}$ denote the treatment and comparison group datasets in each time period. We implement the panel data version of the estimator in the empirical work reported below and find it to be more robust than the cross-sectional matching estimators.[21]

[21] When using repeated cross section data, the identity of future participants and nonparticipants may not be known in the pre-program period. A variant of the difference-in-differences matching estimator presented here for that context appears in Blundell and Costa-Dias (2000).

3.4. Choice-based sampled data

The samples used in evaluating the impacts of programs are often choice-based, with program participants oversampled relative to their frequency in the population of persons eligible for the program. Under choice-based sampling, weights are required to consistently estimate the probabilities of program participation.[22] When the weights are unknown, Heckman and Todd (1995) show that with a slight modification, matching methods can still be applied, because the odds ratio estimated using the incorrect weights (i.e., ignoring the fact that samples are choice-based) is a scalar multiple of the true odds ratio, which is itself a monotonic transformation of the propensity scores. Therefore, matching can proceed on the (misweighted) estimate of the odds ratio (or of the log odds ratio). In our empirical work, the data are choice-base sampled and the sampling weights are unknown, so we match on the odds ratio, $P/(1-P)$.[23]

3.5. When does bias arise in matching?

The success of a matching estimator clearly depends on the availability of observable data to construct the conditioning set Z, such that (6) and (7), or (8) and (9), are satisfied. Suppose only a subset $Z_0 \subset Z$ of the variables required for matching is observed. The propensity score matching estimator based on Z_0 then converges to

$$\alpha'_{\mathrm{M}} = E_{P(Z_0)|D=1}(\mathrm{E}(Y_1|P(Z_0), D=1) - \mathrm{E}(Y_0|P(Z_0), D=0)). \tag{12}$$

The bias for the parameter of interest, $\mathrm{E}(Y_1 - Y_0|D=1)$, is

$$bias_{\mathrm{M}} = \mathrm{E}(Y_0|D=1) - E_{P(Z_0)|D=1}\{\mathrm{E}(Y_0|P(Z_0), D=0)\}. \tag{13}$$

HIST (1998a) show that what variables are included in the propensity score matters in practice for the estimated bias. They find that the lowest bias values arise when Z includes a rich set of variables that affect both program participation and labor market outcomes. They obtain higher bias values using cruder sets of Z variables. Similar findings regarding the sensitivity of matching estimates to the set of matching variables appear in Lechner (2002) and in Section 8 of this paper.

3.6. Using data on randomized-out controls and nonparticipants to estimate evaluation bias

With only nonexperimental data, it is impossible to disentangle the treatment effect from the evaluation bias associated with any particular estimator. However,

[22] See, e.g., Manski and Lerman (1977) for a discussion of weighting for logistic regressions.

[23] With single nearest neighbor matching, it does not matter whether matching is performed on the odds ratio or on the propensity scores (estimated using the wrong weights), because the ranking of the observations is the same and the same neighbors will be selected. Thus, failure to account for choice-based sampling should not affect the nearest-neighbor point estimates in the DW (1999, 2002) studies. However, for methods that take account of the absolute distance between observations, such as kernel matching or local linear matching, it does matter.

data on a randomized-out control group makes it possible to separate out the bias. First, subject to some caveats discussed in Heckman and Smith (1995) and Heckman et al. (1999), randomization ensures that the control group is identical to the treatment group in terms of the pattern of self-selection. Second, the randomized-out control group does not participate in the program, so the impact of the program on them is known to be zero. Thus, a nonexperimental estimator applied to the control group data combined with nonexperimental comparison group data should, if consistent, produce an estimated impact equal to zero. Deviations from zero are properly interpretable as evaluation bias.[24] Therefore, the performance of alternative nonexperimental estimators can be evaluated by applying the estimators to data from the randomized-out control group and from the nonexperimental comparison group and then checking whether the resulting estimates yield are statistically distinguishable from zero.

4. The national supported work demonstration

The National Supported Work (NSW) Demonstration[25] was a transitional, subsidized work experience program that operated for 4 years at 15 locations throughout the United States. It served four target groups: female long-term AFDC recipients, ex-drug addicts, ex-offenders, and young school dropouts. The program first provided trainees with work in a sheltered training environment and then assisted them in finding regular jobs. About 10,000 persons experienced 12–18 months of employment through the program, which cost around $13,850 per person in 1997 dollars.

To participate in NSW, potential participants had to satisfy a set of eligibility criteria that were intended to identify individuals with significant barriers to employment. The main criteria were: (1) the person must have been currently unemployed (defined as having worked no more than 40 hours in the 4 weeks preceding the time of selection into the program), and (2) the person must have spent no more than 3 months on one regular job of at least 20 hours per week during the preceding 6 months. As a result of these criteria as well as of self-selection into the program, persons who participated in NSW differ in many ways from the general U.S. population.

[24] A different way of isolating evaluation bias would be to compare the program impact estimated experimentally (using the treatment and randomized-out control samples) to that estimated nonexperimentally (using the treatment and comparison group samples). This approach is taken in LaLonde (1986) and in DW (1999, 2002). The procedure we use, which compares the randomized-out controls to nonparticipants, is equivalent and a more direct way of estimating the bias. It is also more efficient in our application as the control group is larger than the treatment group. The latter approach is also taken in HIT (1997a) and HIST (1998a).

[25] See Hollister et al. (1984) for a detailed description of the NSW demonstration and Couch (1992) for long-term experimental impact estimates.

From April 1975 to August 1977[26] the NSW program in 10 locations operated as a randomized experiment with some program applicants being randomly assigned to a control group that was not allowed to participate in the program.[27] The experimental sample includes 6616 treatment and control observations for which data were gathered through a retrospective baseline interview and four follow-up interviews. These interviews cover the two years prior to random assignment and up to 36 months thereafter. The data provide information on demographic characteristics, employment history, job search, mobility, household income, housing and drug use.[28] As noted in Heckman et al. (1999), the NSW is an ideal experiment in the sense that, unlike many other social experiments, almost everyone in the experimental treatment group participates in the program and no one in the experimental control group receives a similar treatment from other sources (though a small fraction receive much less intensive employment and training services under the CETA program).

5. Samples

In this study, we consider three experimental samples and two nonexperimental comparison groups. All of the samples are based on the male samples from LaLonde (1986).[29] The experimental sample includes male respondents in the NSWs ex-addict, ex-offender and high school dropout target groups who had valid pre- and post-program earnings data.

The first experimental sample is the same as that employed by LaLonde (1986). The sample consists of 297 treatment group observations and 425 control group observations. Descriptive statistics for the LaLonde experimental sample appear in the first column of Table 1. These statistics show that solid majorities of male NSW participants were minorities (mostly black), high school dropouts and unmarried. As was its aim, the NSW program served a highly economically disadvantaged population.

The earnings variables for the NSW samples are all based on self-reported earnings measures from surveys.[30] Following LaLonde (1986), all of the earnings variables (for all of the samples) are expressed in 1982 dollars. The variable denoted "Real Earnings in 1974" consists of real earnings in months 13–24 prior to the

[26] Our sample does not include persons randomly assigned in all of these months due to the sample restrictions imposed by LaLonde (1986).

[27] Then ten locations where random assignment took place are Atlanta, Chicago, Hartford, Jersey City, Newark, New York, Oakland, Philadelphia, San Francisco, and Fond du Lac and Winnebago counties in Wisconsin.

[28] In addition, persons in the AFDC target group were also asked about children in school and welfare participation and non-AFDC target groups were asked about illegal activities.

[29] We do not examine LaLonde's (1986) sample of AFDC women as it is no longer available due to data storage problems.

[30] As noted in Section 2, grouped social security earnings data are also available for the NSW experimental sample, and were employed by Heckman and Hotz (1989) in their analysis. We do not use them here in order to maintain comparability with LaLonde (1986) and DW (1999, 2002).

Table 1
Descriptive statistics for male experimental and comparison group samples

Variable	NSW experimental samples			Comparison groups	
	LaLonde	Dehejia-Wahba	Early random assignment	CPS sample	PSID sample
Age	24.52	25.37	25.74	33.23	34.85
	(6.63)	(7.10)	(6.75)	(11.05)	(10.44)
Education	10.27	10.2	10.37	12.03	12.12
	(1.70)	(1.79)	(1.60)	(2.87)	(3.08)
Black	0.80	0.84	0.82	0.07	0.25
	(0.40)	(0.37)	(0.38)	(0.26)	(0.43)
Hispanic	0.11	0.09	0.10	0.07	0.03
	(0.31)	(0.28)	(0.30)	(0.26)	(0.18)
Married	0.16	0.17	0.20	0.71	0.87
	(0.37)	(0.37)	(0.40)	(0.45)	(0.34)
No H.S. degree	0.78	0.78	0.76	0.30	0.31
	(0.41)	(0.41)	(0.43)	(0.46)	(0.46)
"Real earnings in 1974"	3631	2102	3742	14017	19429
	(6221)	(5364)	(6718)	(9570)	(13407)
Real earnings in 1975	3043	1377	2415	13651	19063
	(5066)	(3151)	(3894)	(9270)	(13597)
Real earnings in 1978	5455	5301	5796	14847	21554
	(6253)	(6632)	(7582)	(9647)	(15555)
Real earnings in 1979	...	...	...	14730	...
				(11028)	
"Zero earnings in 1974"	0.45	0.73	0.52	0.12	0.09
	(0.50)	(0.44)	(0.50)	(0.32)	(0.28)
Zero earnings in 1975	0.40	0.65	0.41	0.11	0.10
	(0.49)	(0.48)	(0.49)	(0.31)	(0.30)
Experimental impact	886	1794	2748	...	...
(1978 earnings)	(488)	(670)	(1005)		
Sample size	297 treatments 425 controls	185 treatments 260 controls	108 treatments 142 controls	15992	2490

Notes: Estimated standard deviations in parentheses. Robust standard errors are reported for experimental impact estimates.

month of random assignment. For persons randomly assigned early in the experiment, these months largely overlap with calendar year 1974. For persons randomly assigned later in the experiment, these months largely overlap with 1975. This is the variable denoted "Re74" in DW (1999, 2002). The variable "Zero Earnings in 1974" is an indicator variable equal to one when the "Real Earnings in 1974" variable equals zero.[31] The Real Earnings in 1975 variable corresponds to earnings in calendar year 1975; the indicator variable for Zero Earnings in 1975 is coded to one if Real Earnings in 1975 equal zero. Mean earnings in the male NSW

[31] This is the variable denoted "U74" in DW (1999, 2002); note that it corresponds to nonemployment rather than unemployment.

sample prior to random assignment were quite low. They also fall from 1974 to 1975, another example of the common pattern denoted ''Ashenfelter's dip'' in the literature (see, e.g., Heckman and Smith, 1999). The simple mean-difference experimental impact estimate for this group is $886, which is statistically significant at the 10 percent level.

The second experimental sample we use is that used in DW (1999, 2002). In order to include two years of pre-program earnings in their model of program participation, DW omit the (approximately) 40 percent of LaLonde's (1986) original sample for which that information is missing.[32] While DW (1999, 2002) provide general descriptions of the sample selection criteria they used to generate their analysis samples, we required the exact criteria to replicate their results and to examine alternative propensity scores using their sample.[33] Table 2 illustrates the sample inclusion criteria that we found (partly through trial and error) which correctly account for all but one observation in their sample.[34] The table is a cross-tabulation of LaLonde's (1986) sample with month of random assignment as rows and zero earnings in months 13–24 as columns. Corresponding to the rows and columns of Table 2, their rule has two parts. First, include everyone randomly assigned in January through April of 1976. This group corresponds to the eight shaded cells in the bottom four rows of Table 2. Second, of those who were randomly assigned after April of 1976, only include persons with zero earnings in months 13–24 before random assignment. This group corresponds to the six shaded cells at the top of the left column of Table 2. Left out of the sample are those members of LaLonde's (1986) sample who were randomly assigned after April 1976 and had positive earnings in months 13–24 before random assignment. This rule corresponds fairly closely to the verbal statement in DW (1999). We do not believe that the second rule is appropriate. DW state that they want to use ''earnings in 1974'' as an additional control variable. However, as already noted, earnings in months 13–24 before random assignment either do not overlap calendar year 1974 or do so only for a few months for those included under the second part of the rule.

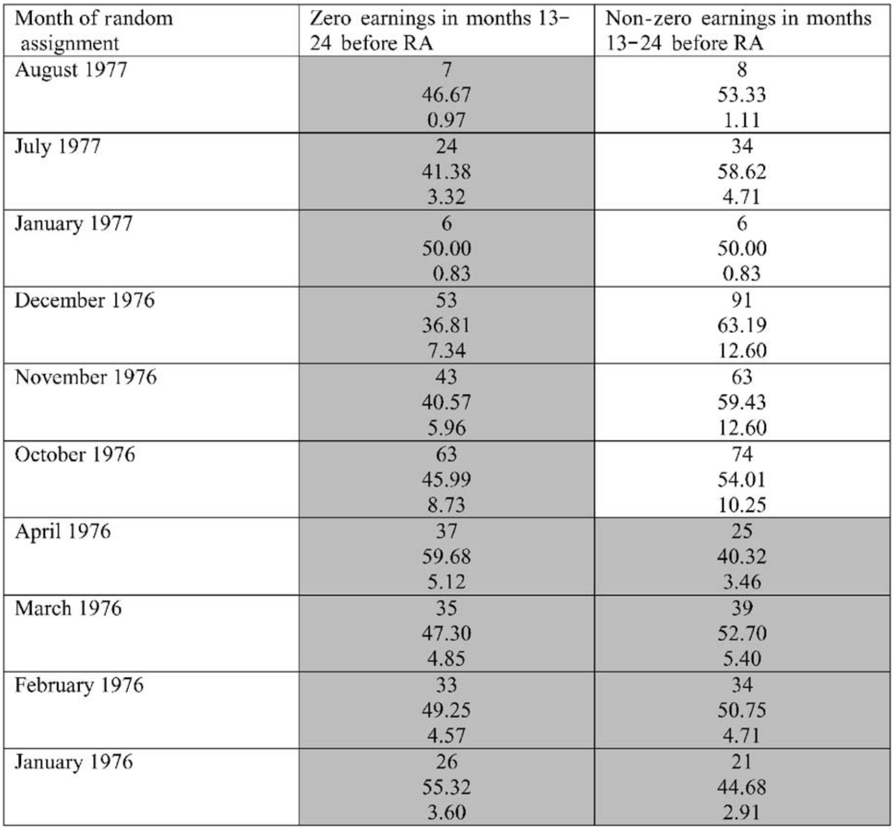
The second column of Table 1 displays the descriptive statistics for the DW sample. Along most dimensions, the DW sample is similar to the full LaLonde sample. One key difference results from the second part of the rule, which differentially includes persons with zero earnings in parts of 1974 and 1975. As a

[32] The inclusion of the additional variable was motivated by findings in the earlier literature. Heckman and Smith (1999) show that variables based on labor force status in the months leading up to the participation decision perform better at predicting program participation in the National JTPA Study data than do annual or quarterly earnings. See also related discussions in Ashenfelter (1978), Ashenfelter and Card (1985), Card and Sullivan (1988) and Angrist (1990,1998) on this point.

[33] See the bottom of the first column of page 1054 of DW (1999) for their descriptions.

[34] Dehejia provided us with both their version of the LaLonde sample and a version of the DW sample in separate files. Both files are available on Dehejia's web page at http://www.columbia.edu/˜rd247/. However, neither file includes identification numbers, so there is no simple way to link them to determine the exact sample restrictions used. By trying different combinations of sample inclusion criteria, we determined the rules for generating the subsample. One control group observation is included by the rules stated here but excluded from their sample. Our estimates below using the ''DW'' sample do not include this extra observation.

Table 2
Dehejia and Wahba (1999, 2002) sample composition



Month of random assignment	Zero earnings in months 13–24 before RA	Non-zero earnings in months 13–24 before RA
August 1977	7 46.67 0.97	8 53.33 1.11
July 1977	24 41.38 3.32	34 58.62 4.71
January 1977	6 50.00 0.83	6 50.00 0.83
December 1976	53 36.81 7.34	91 63.19 12.60
November 1976	43 40.57 5.96	63 59.43 12.60
October 1976	63 45.99 8.73	74 54.01 10.25
April 1976	37 59.68 5.12	25 40.32 3.46
March 1976	35 47.30 4.85	39 52.70 5.40
February 1976	33 49.25 4.57	34 50.75 4.71
January 1976	26 55.32 3.60	21 44.68 2.91

The values for each cell indicate the number of observations in the cell, the row percentage and the overall percentage. The shaded area indicates the DW sample.

result, mean earnings in both years are lower for the DW sample than for the larger LaLonde sample. The other key difference is in the experimental impact estimate. At $1794 it is more than twice as large as that for the LaLonde sample.

The third experimental sample we examine is not used in either LaLonde (1986) or DW (1999, 2002). It is a proper subset of the DW sample that excludes persons randomized after April of 1976. We examine this sample because we find their decision to include persons randomized after April of 1976 only if they had zero earnings in months 13–24 problematic. Our “Early RA” sample consists of persons randomly assigned during January through April of 1976; put differently, this sample consists of the observations in the bottom four rows of Table 2. This sample includes 108 treatment group members and 142 control group members. Descriptive statistics for this sample appear in the third column of Table 1. Ashenfelter’s dip is stronger for this sample (a drop of about $1200 rather than one of about $700) than for the DW sample, as is to be expected given that it drops the large contingent of persons with zero earnings in months 13–24 prior to random assignment. The $2748 experimental impact for the Early RA sample is the largest among the three experimental samples.

The comparison group samples we use are the same ones used by LaLonde (1986) and DW (1999, 2002). Both are based on representative national samples drawn from throughout the United States. This implies that the vast majority of comparison group members, even those with observed characteristics similar to the experimental sample members, are drawn from different local labor markets. In addition, earnings are measured differently in both comparison group samples than they are in the NSW data.

The first comparison group sample is based on Westat's matched Current Population Survey—Social Security Administration file. This file contains male respondents from the March 1976 Current Population Survey (CPS) with matched Social Security earnings data. The sample excludes persons with nominal own incomes greater than $20,000 and nominal family incomes greater than $30,000 in 1975. Men over age 55 are also excluded. Descriptive statistics for the CPS comparison group appear in the fourth column of Table 1. Examination of the descriptive statistics reveals that the CPS comparison group is much older, much better educated (70 percent completed high school), much less likely to be black or Hispanic and much more likely to be married than any of the NSW experimental samples.

The earnings measures for the CPS sample are individual-level administrative annual earnings totals from the U.S. Social Security system. The CPS comparison group sample had, on average, much higher earnings than the NSW experimental sample in every year. (The "Real Earnings in 1974" variable for the CPS comparison group corresponds to calendar year 1974.) There is a slight dip in the mean earnings of the CPS comparison group from 1974 to 1975; this dip is consistent with the imposition of maximum individual and family income criteria in 1975 for inclusion in the sample along with some level of mean-reversion in earnings (see related discussion in Devine and Heckman, 1996). The very substantial differences between this comparison group and the NSW experimental group pose a tough problem for any nonexperimental estimator to solve.

The second comparison group sample is drawn from the Panel Study of Income Dynamics (PSID). It consists of all male household heads from the PSID who were continuously present in the sample from 1975 to 1978, who were less than 55 years old and who did not classify themselves as retired in 1975.[35] Descriptive statistics for the PSID comparison group sample appear in the fifth column of Table 1. The PSID comparison group strongly resembles the CPS comparison group in its observable characteristics. Mean earnings levels in the PSID sample are higher than those in the CPS sample and the fraction with zero earnings in 1974 and 1975 lower, most likely due to the maximum income criteria imposed in selecting the CPS sample. The over-representation of blacks in the PSID comparison group sample relative to the U.S.

[35] Following DW (1999, 2002), we drop the three persons from LaLonde's sample who are missing data on education.

population appears to result from the use of both the representative sample component of the PSID and the component based on the Survey of Economic Opportunity sample, which consists of low income urban residents in the North and low income rural residents in the South.[36]

LaLonde (1986) also considers four other comparison groups consisting of various subsets of the CPS and PSID comparison groups just described. As defined in the notes to his Table 3, these subsamples condition on various combinations of employment, labor force status and income in 1975 or early 1976. We do not examine these subsamples here for two main reasons. First, taking these subsamples and then applying matching essentially represents doing "matching" in two stages—first crudely based on a small number of characteristics and then more carefully using the propensity score.[37] As discussed in Heckman et al. (1999), such estimators (like estimators consisting of crude matching followed by some other nonexperimental estimator) do not always have clear economic or econometric justifications. One case where they do is when the first stage amounts to imposing the program's eligibility rules, thereby dropping from the sample individuals whose probability of participation is known to equal zero; unfortunately, the CPS and PSID data sets lack the information required to apply the NSW eligibility rules with any degree of precision. Another case where such sample restrictions would have some justification consists of excluding non-participants in local labor markets with no participants from the analysis. Doing this with the CPS or PSID data in the NSW context would leave only tiny comparison samples. Second, Table 3 of DW (1999) shows that, in the context of propensity score matching, the first round of crude matching performed by LaLonde (1986) has little effect on the resulting estimates. The propensity score matching estimator for the full sample assigns little or no weight to those sample members who get excluded by the crude matching used to create the subsamples.[38]

6. Propensity scores

We present matching estimates based on two alternative specifications of the propensity score, $\Pr(D = 1|Z)$. The first specification is that employed in DW (1999, 2002); the second specification is based on LaLonde (1986). Although LaLonde does not consider matching estimators, he estimates a probability of participation in the course of implementing the classical selection estimator of Heckman (1979). In both cases, we use the logit model to estimate the scores.

[36] See the detailed information about the PSID at http://www.isr.umich.edu/src/psid/overview.html.

[37] Even the full CPS comparison group sample we use has this feature due to the conditioning on individual and family income in 1975 performed by Westat in creating the sample.

[38] Because the NSW operated in only ten locations and served only a small number of individuals in total, the probability that even one of our comparison group members participated in NSW is very low. Hence, the problem of "contamination bias" (the nonexperimental analogue of substitution bias) defined by Heckman and Robb (1985) does not arise in our analysis.

Table 3
Dehejia and Wahba (2002) propensity score model coefficient estimates (estimated standard errors in parentheses)

Variable	LaLonde experimental sample		DW experimental sample		Early RA experimental sample	
	CPS	PSID	CPS	PSID	CPS	PSID
Age	2.6119	0.1739	2.7441	0.2386	3.0783	0.2292
	(0.2146)	(0.0739)	(0.2681)	(0.0932)	(0.3288)	(0.1095)
Age squared	−0.7560	−0.0042	−0.0779	−0.0051	−0.0879	−0.0059
	(0.0068)	(0.0011)	(0.0085)	(0.0014)	(0.0104)	(0.0017)
Age cubed/1000.0	0.6678	N.A.	0.6769	N.A.	0.7723	N.A.
	(0.0678)		(0.0837)		(0.1029)	
Years of schooling	1.2755	1.0247	1.2274	0.9748	1.7877	1.6650
	(0.1909)	(0.2433)	(0.2249)	(0.3028)	(0.3739)	(0.4639)
Years of schooling squared	−0.0700	−0.0539	−0.0692	−0.0525	−0.0938	−0.0850
	(0.0099)	(0.0124)	(0.0120)	(0.0160)	(0.0193)	(0.0246)
High school dropout	1.4282	0.9112	1.3515	0.7490	1.3823	0.7184
	(0.1929)	(0.2564)	(0.2588)	(0.3481)	(0.3003)	(0.3877)
Married	−1.8725	−2.2825	−1.7307	−2.0301	−1.6805	−1.9142
	(0.1471)	(0.1747)	(0.1932)	(0.2416)	(0.2149)	(0.2545)
Black	3.8540	2.0369	3.9988	2.6277	3.9600	2.2967
	(0.1445)	(0.2004)	(0.2000)	(0.2998)	(0.2451)	(0.3211)
Hispanic	2.1957	2.6524	2.2457	3.3643	2.3164	3.0703
	(0.1879)	(0.3687)	(0.2637)	(0.5426)	(0.3188)	(0.5441)
"Real earnings in 1974"	−0.00011	−0.00005	−0.00007	−0.00002	−0.00002	−0.00003
	(0.00005)	(0.00027)	(0.00007)	(0.00003)	(0.00008)	(0.00004)
Real earnings in 1974 squared	N.A.	1.54e-09	N.A.	1.64e-09	N.A.	1.86e-09
		(5.0e-10)		(6.87e-10)		(6.32e-10)
Real earnings in 1975	−0.00011	−0.00013	−0.00020	−0.00025	−0.00022	−0.00024
	(0.00002)	(0.00003)	(0.00003)	(0.00004)	(0.00003)	(0.00004)
Real earnings in 1975 squared	N.A.	2.97e-11	N.A.	5.28e-10	N.A.	4.10e-10
		(3.9e-10)		(5.68e-10)		(5.30e-10)
Zero earnings in 1974	0.7660	2.2754	1.9368	3.2583	1.3592	2.4476
	(0.1693)	(0.3788)	(0.2209)	(0.4340)	(0.2398)	(0.4360)
Zero earnings in 1975	−0.0320	−1.0192	0.2513	−1.0396	−0.5564	−1.3899
	(0.1703)	(0.3547)	(0.1994)	(0.3871)	(0.2329)	(0.3932)
Schooling * "Real earnings in 1974"	9.92e-06	N.A.	0.00001	N.A.	6.25e-06	N.A.
	(4.4e-06)		(6.14e-06)		(7.15e-06)	
"Zero earnings in 1974" * Hispanic	N.A.	−1.0683	N.A.	−1.4627	N.A.	−0.7382
		(0.7193)		(0.7882)		(0.8670)
Intercept	−36.9901	−6.6368	−39.8326	−8.5683	−46.1939	−12.7065
	(2.4165)	(1.6405)	(3.0398)	(2.0629)	(3.9116)	(2.7713)

The estimated coefficients and associated estimated standard errors for the propensity scores based on the DW (2002) specification appear in Table 3.[39] We estimate six sets of scores, one for each pair of experimental and comparison group samples. In each case, the dependent variable is an indicator for being in the

[39] DW (1999) use slightly different specifications for both the CPS and PSID comparison groups. Compare the notes to Tables 2 and 3 in DW (2002) with the notes to Table 3 in DW (1999).

experimental sample. We follow DW in including slightly different sets of higher order and interaction terms in the specifications for the CPS and PSID comparison groups. These terms were selected using their propensity score specification selection algorithm, discussed in the next section. Our estimated scores for the DW specification with the DW sample differ slightly from theirs for two reasons. First, for efficiency reasons we use both the experimental treatment and experimental control group in estimating the scores, whereas DW (1999, 2002) appear to use only the treatment group.[40] Second, although DW (2002) did not include a constant term in the logistic model, we do.

Most of the coefficient estimates for the DW model are in the expected direction given the differences observed in Table 1. For example, high school dropouts are more likely to participate in NSW, as are blacks and hispanics, while marriage has a strong negative effect on the probability of participation. In the CPS sample, participation probabilities decrease with earnings in both "1974" and 1975. In the PSID sample, the relationship is quadratic. The estimated probability of participation is also nonlinear in age and education in both samples, with a maximum at around 23.4 years of age for the DW experimental sample and the PSID comparison group. The qualitative, and also the quantitative, pattern of the coefficients is extremely similar across experimental samples with the same comparison group. There are, though, a few differences across comparison groups for the same experimental sample, perhaps because of the somewhat different specifications.

With the CPS comparison group, the correlations between scores estimated on different experimental samples are around 0.93. With the PSID, they are a bit higher at around 0.97. Neither figure suggests that estimating the score on a particular experimental sample matters much. Using the prediction rate metric as one tool to assess the quality of the propensity scores shows that the specification does a good job of separating out the participants and the nonparticipants.[41] We use the fraction of the combined sample that consists of experimentals as the cutoff for predicting someone to be a participant. For the DW scores applied to the DW sample, 94.1 percent of the CPS comparison group members are correctly predicted to be nonparticipants and 94.6 percent of the experimental sample is correctly predicted to participate. For the DW scores applied to the LaLonde and early RA samples, the corresponding correct prediction rates are (95.6,85.3) and (91.2,94.8). The prediction rates are similar, but a bit lower in some cases, with the PSID comparison group.

Fig. 1 presents histograms of the estimated log-odds ratios for the DW propensity score model applied to each of the three experimental samples with each of the two comparison groups. The columns in these figures allow a graphical assessment of the

[40] We experimented a bit with generating estimates based on scores estimated using just the treatment group, just the control group and both the treatment and control groups. The samples are small enough that this choice can move the resulting impact estimates around by two or three hundred dollars.

[41] This metric is discussed in Heckman and Smith (1999) and HIST (1998a). For caveats, see Heckman and Navarro-Lazano (2004).

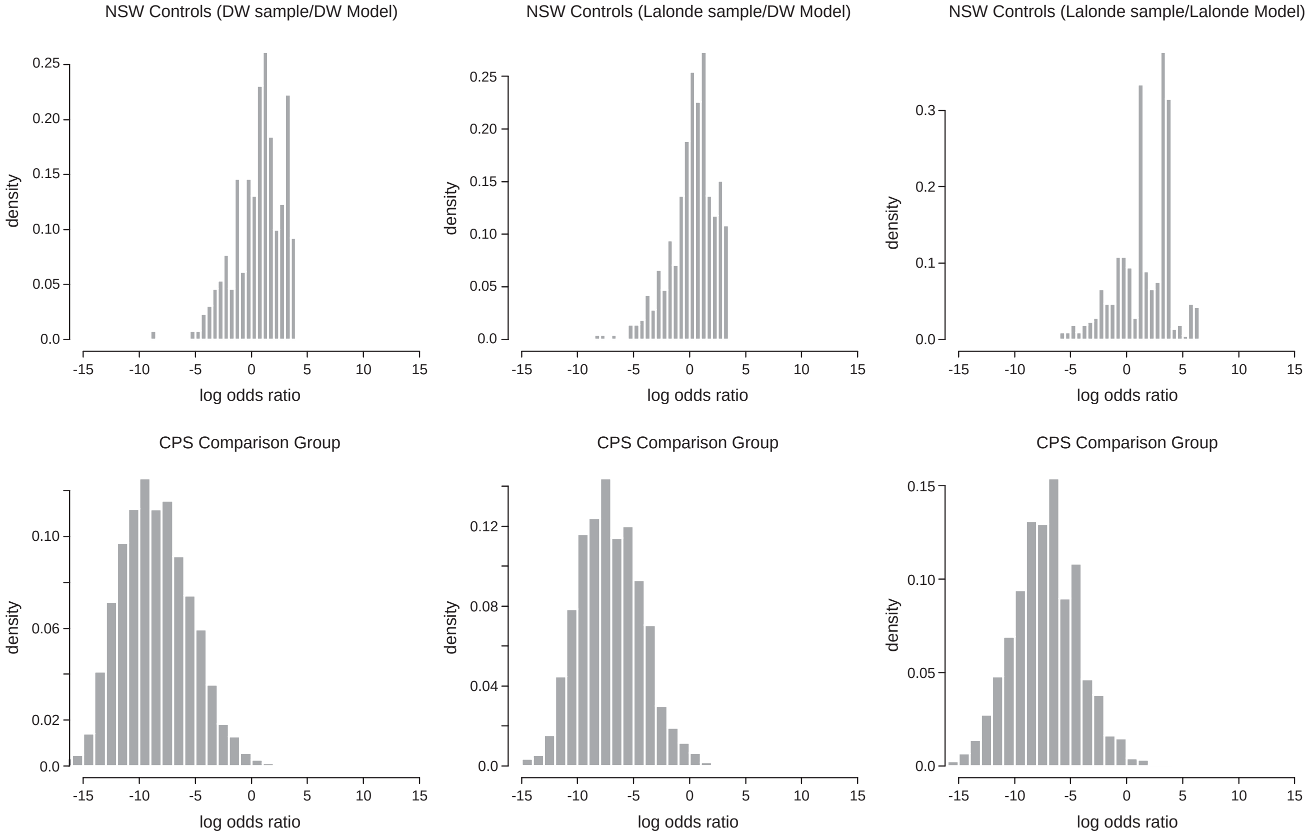


Fig. 1. (a,b) Distribution of estimated log odds ratios.

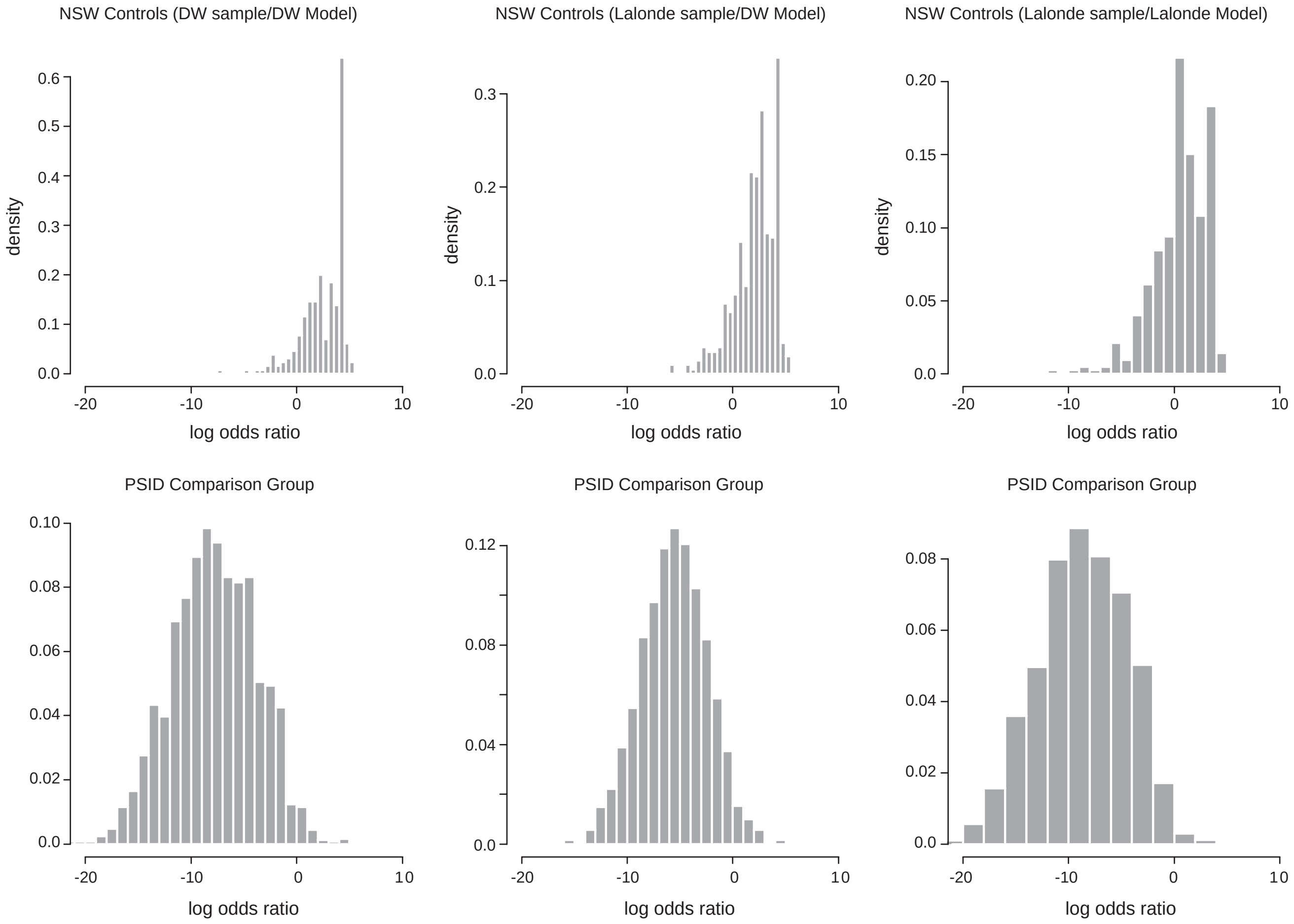


Fig. 1 (*continued*).

extent of any support problems in the NSW data. The figures make readily apparent that the distributions of scores among the experimental samples differ strongly from those of both of the comparison groups. For every combination of experimental sample and comparison group, the density for the comparison group lies well to the left of that of the experimentals. This indicates that many comparison group members have very low predicted probabilities of participation in the NSW program. This finding comports with the strong differences in observable characteristics reported in Table 1. However, the support problem here is not as strong as in the JTPA data examined in HIST (1996, 1998a), where there were large intervals of P with no comparison group observations at all. For the two comparison groups employed here, even at high probabilities, such as those above 0.9, there are at least a handful of comparison group observations.

Table 4 presents the coefficient estimates from the participation model in LaLonde (1986).[42] The patterns are quite similar to those for the DW scores. The participation probability is quadratic in age, with a maximum at 25.3 years for the LaLonde sample with the CPS comparison group and a maximum at 20.2 years for the LaLonde sample with the PSID comparison group. As expected given the differences seen in Table 1, being a high school dropout, being black and being Hispanic have strong and statistically significant positive effects on participation. In contrast, being married and being employed in March of 1976 have strong and statistically significant negative effects on participation.[43] Finally, number of children has a strong negative effect on the participation probability, particularly in the CPS sample.

Like the DW scores, the LaLonde scores estimated on different experimental samples are highly correlated; in every case the correlation exceeds 0.97. The

[42] We ran into two small difficulties in replicating LaLonde's (1986) scores that we resolved as follows. First, LaLonde indicates that he includes a dummy variable for residence in an SMSA in his model. Given that everyone in the NSW experimental sample lives in an SMSA, not living in an SMSA is a perfect predictor of not being in the NSW demonstration. Thus, this variable should not be included in the model. We dealt with this in two ways. In one case, we just dropped this variable from the specification. In the other, we set the participation probability to zero for everyone not in an SMSA and then estimated the model on those who remained. The scores produced in these two ways had a correlation of 0.9734 in the combined LaLonde (1986) experimental sample and CPS comparison group sample and a correlation of 0.9730 in the combined sample with the PSID. The estimates presented in Table 4 are for the specification that sets the probability to zero for all CPS and PSID comparison group members not living in an SMSA.

The second issue concerns missing values of the variables for the number of children. There are missing values for observations in the experimental sample and in the CPS comparison group, but not in the PSID sample. As a result of the asymmetry between the two comparison groups in this regard, we adopt separate strategies in the two cases. In estimating the LaLonde propensity score model with the CPS comparison group, we set missing values of the number of children to zero and include an indicator variable set to one for observations with a missing value and zero otherwise. In the PSID case, we impute missing values of the number of children variable in the experimental data by running a regression of number of children on a set of exogenous covariates (including interactions of age and age squared with race and ethnicity).

[43] The latter variable represents an attempt to capture one aspect of the NSW eligibility rules. Nonetheless, it is somewhat problematic, given that some members of the NSW sample are randomly assigned in January and February of 1976, and therefore some treatment group members could be employed as part of the program by March of 1976. Given the sign and magnitude of the estimated coefficient, this concern appears to be a minor one.

Table 4
LaLonde (1986) propensity score model coefficient estimates (estimated standard errors in parentheses)

Variable	LaLonde experimental sample		DW experimental sample		Early RA experimental sample	
	CPS	PSID	CPS	PSID	CPS	PSID
Age	0.3445	0.1739	0.3932	0.2204	0.4412	0.3436
	(0.0588)	(0.0716)	(0.0689)	(0.0801)	(0.0924)	(0.1070)
Age squared	−0.0068	−0.0043	−0.0072	−0.0047	−0.0081	−0.0068
	(0.0010)	(0.0011)	(0.0011)	(0.0012)	(0.0015)	(0.0017)
Years of schooling	−0.0126	−0.0311	0.0147	−0.0258	−0.0042	−0.1177
	(0.0362)	(0.0502)	(0.0435)	(0.0568)	(0.0550)	(0.0729)
High school dropout	2.0993	1.6396	2.2222	1.5613	2.1959	1.3237
	(0.1972)	(0.2306)	(0.2438)	(0.2664)	(0.2986)	(0.3108)
Black	3.9569	2.0614	4.1637	2.1835	3.9714	1.7441
	(0.1623)	(0.1911)	(0.2126)	(0.2363)	(0.2687)	(0.2855)
Hispanic	2.1891	2.3517	2.1930	2.4690	1.9834	2.0859
	(0.2150)	(0.3282)	(0.2889)	(0.3887)	(0.3713)	(0.4387)
Working in 1976	−2.1184	−2.4017	−2.4166	−2.5784	−1.9932	−2.0762
	(0.1396)	(0.1635)	(0.1739)	(0.1861)	(0.2104)	(0.2203)
Number of children	−1.0608	−0.3826	−1.0392	−0.3639	−0.9028	−0.2343
	(0.0648)	(0.0777)	(0.0809)	(0.0898)	(0.0986)	(0.1014)
Missing children variable	2.6233	N.A.	3.2783	N.A.	3.4188	N.A.
	(0.3512)		(0.3813)		(0.4422)	
Intercept	−6.9687	−1.3695	−8.8816	−2.0868	−9.6280	−3.5263
	(0.9800)	(1.1894)	(1.1759)	(1.3367)	(1.5639)	(1.7471)

prediction rates are similar as well. For the LaLonde scores with the LaLonde experimental sample and the CPS comparison group, 95.4 percent of the participants are correctly predicted along with 94.7 percent of the comparison group. With the PSID, the corresponding values are 95.0 and 92.8 percent. Similar percentages hold for the other experimental samples, but with slightly higher prediction rates for the participants and slightly lower ones for the non-participants. The correlations between the LaLonde scores and the DW scores are between 0.77 and 0.83 for the CPS comparison group and between 0.88 and 0.93 for the PSID comparison group; it is not clear why the correlation is higher in the PSID case. With both samples, but particularly with the CPS, it is clear that the LaLonde scores differ meaningfully from the DW scores. Finally, Fig. 1 shows that the LaLonde scores for all three experimental samples, like the DW scores, are spread out over the full range between zero and one, but the density is quite thin among nonparticipants at the higher scores.

7. Variable selection and the balancing test

Under the conditional mean independence assumption required for application of propensity score matching, the outcome variable must be conditionally mean independent of treatment conditional on the propensity score, $P(Z)$. Implementing matching requires choosing a set of variables Z that plausibly satisfy this condition. This set should include all of the key factors affecting both program participation and outcomes—that is, all the variables that affect both D and Y_0. No mechanical algorithm exists that automatically chooses sets of variables Z that satisfies the identification conditions. Moreover, the set Z that satisfies the matching conditions is not necessarily the most inclusive one. Augmenting a set of variables that satisfies the identification conditions for matching could lead to a violation of those conditions. Adding additional conditioning variables may also exacerbate a common support problem. Finally, it should always be kept in mind that any given data set may contain no combination of variables Z that satisfy the conditional independence assumption. In the latter case, matching is not an appropriate estimator.

To guide in the selection of Z, there is some accumulated empirical evidence on how bias estimates of matching estimators depend on the choice of Z in particular applications. For example, HIT (1997a), HIST (1998a), Heckman and Smith (1999) and Lechner (2002) show that which variables are included in the estimation of the propensity score can make a substantial difference to the performance of the estimator. These papers find, in general, larger biases with cruder conditioning sets. Theory also provides a guide to variables likely to affect both participation and outcomes in particular contexts.

Rosenbaum and Rubin (1983) present a theorem (see their Theorem 2) that does not aid in choosing which variables to include in Z, but which can help in determining which interactions and higher order terms to include for a given set of included Z variables. The theorem states that

$$Z \perp\!\!\!\perp D|\Pr(D=1|Z),$$

or equivalently

$$\mathrm{E}(D|Z,\Pr(D=1|Z)) = \mathrm{E}(D|\Pr(D=1|Z)).$$

The basic intuition is that after conditioning on $\Pr(D=1|Z)$, additional conditioning on Z should not provide new information about D. Thus, if after conditioning on the estimated values of $P(D=1|Z)$ there is still dependence on Z, this suggests misspecification in the model used to estimate $\Pr(D=1|Z)$. Note that the theorem holds for any Z, including sets Z that do not satisfy the conditional independence condition required to justify matching. As such, the theorem is not informative about what set of variables to include in Z.

This theorem motivates a specification test for $\Pr(D=1|Z)$. The general idea is to test whether or not there are differences in Z between the $D=1$ and 0 groups after conditioning on $P(Z)$. The test has been implemented in the literature in a number of ways. Lechner (1999), Lechner (2000) and Eichler and Lechner (2002) use a variant of a measure suggested in Rosenbaum and Rubin (1985) that is based on standardized differences between the treatment and matched comparison group samples in terms of means of each variable in Z, squares of each variable in Z and first-order interaction terms between each pair of variables in Z. An alternative approach used in DW (1999, 2002) divides the observations into strata based on the estimated propensity scores. These strata are chosen so that there is not a statistically significant difference in the mean of the estimated propensity scores between the experimental and comparison group observations within each strata, though how the initial strata are chosen and how they are refined if statistically significant differences are found is not made precise. The problem of choosing the strata in implementing the balancing test is analogous to the problem of choosing the strata in implementing the interval matching estimator, described earlier. Then, within each stratum, *t*-tests are used to test for mean differences in each Z variable between the experimental and comparison group observations. When significant differences are found for particular variables, higher order and interaction terms in those variables are added to the logistic model and the testing procedure is repeated, until such differences no longer emerge.

As described earlier, we use two different model specifications to estimate propensity scores in this paper. The specification based on DW (1999, 2002) was selected using the balancing test strategy described above. The specification based on LaLonde (1986) was selected on the basis of how well the model predicted program participation. We retain LaLonde's (1986) original specification in this paper when we implement the matching estimators to allow for easy comparison of his results with our results, which are based on different estimators and different sample inclusion criteria.

8. Matching estimates

We now present our estimates of the bias obtained when we apply matching to the experimental NSW data and the two different nonexperimental comparison groups. Our estimation strategy differs somewhat from that of LaLonde (1986) and DW (1999, 2002) in that we obtain direct estimates of the bias by applying matching to the randomized-out control group and the nonexperimental comparison groups, whereas the other papers obtain the bias indirectly by applying matching to the treatment and comparison groups and comparing the resulting experimental and the nonexperimental impact estimates. Second, in contrast to DW (1999, 2002), we match on the log-odds ratio rather than on the propensity score itself, so that our estimates are robust to choice-based sampling.

Finally, we impose the common support condition using the trimming method described above, which differs from the method used by DW (1999, 2002) that discards comparison group observations with estimated propensity scores that lie below the minimum or above the maximum of the estimated scores in the experimental sample.[44] The main advantage of this approach is ease of implementation. While somewhat more difficult to implement, our trimming approach has two substantive advantages. First, we do not throw out good matches that lie just below the minimum estimated score in the $D=1$ sample (or just above the estimated maximum). Second, we allow for gaps in the empirical common

[44] See the first column of page 1058 in DW (1999).

support that lie between the extreme values of the estimated propensity scores in the experimental sample. This is important because the nonparametric regression estimators of the counterfactual mean outcomes are unreliable when evaluated at P points where the estimated density is close to zero. In practice, our method of imposing the support condition is somewhat more stringent than that of DW, as we drop five to ten percent of the $D=1$ sample due to the common support condition, in addition to dropping a fraction of the comparison group samples similar to that dropped by DW.

8.1. Cross-sectional matching estimates

Estimates of the bias associated with cross-sectional matching on the propensity score appear in Table 5. These are estimates of the bias expression given previously in Eq. (13). We first consider Panel A of Table 5, which shows the estimates for the CPS comparison group. The outcome variable throughout Table 5 is earnings in calendar year 1978, where January 1978 is at least 5 months after random assignment for all of the controls. The first column of Panel A of Table 5 gives the simple mean difference in 1978 earnings between each experimental control group and the CPS comparison group. The remaining columns present estimates of the bias associated with different matching estimators. The first six rows of the table refer to estimates using the DW propensity score specification, while the final two rows refer to the LaLonde propensity score specification. Each pair of rows presents bias estimates for one experimental sample along with the percentage of the experimental impact estimate for that sample that the bias estimate represents. These percentages are useful for comparisons of different estimators within each row, but are not useful for comparisons across rows given the large differences in experimental impact estimates among the three experimental samples. We present bootstrap standard errors based on 100 replications below each estimate; the standard errors for the percentage impacts assume the experimental impact is constant.

The second through the fifth columns in Table 5 give various estimates based on nearest-neighbor matching, defined above in Section 3.3. The second and third columns present estimates from matching using the one and ten nearest neighbors, respectively, without imposing the common support condition. The fourth and fifth columns present estimates using the same methods but imposing the common support condition. Five important patterns characterize the nearest-neighbor estimates for the CPS comparison group. First, using the DW experimental sample and DW propensity score model, we replicate the low biases that were reported in DW (1999, 2002). Second, when the DW propensity score model is applied to the LaLonde sample or to the Early RA sample, the bias estimates are substantially higher. Indeed, the bias estimates for the DW scores as applied to the Early RA sample are among the largest in the table. Third, the imposition of the common support condition has little effect on the estimates for the LaLonde and DW samples, but does result in a substantial reduction in bias for the Early RA sample. Fourth, increasing the number of nearest neighbors reduces bias in the relatively small Early RA sample, but does little to change the bias estimates for the other two

Table 5
Bias associated with alternative cross-sectional matching estimators. Comparison groups: (A) CPS male sample and (B) PSID male sample. Dependent variable: real earnings in 1978 (bootstrap standard errors in parentheses; trimming level for common support is 2 percent)

Sample and propensity score model	(1) Mean diff.	(2) 1 Nearest neighbor without common support	(3) 10 Nearest-neighbors without common support	(4) 1 Nearest-neighbor with common support	(5) 10 Nearest-neighbors with common support	(6) Local linear matching (bw = 1.0)	(7) Local linear matching (bw = 4.0)	(8) Local linear regression adjusted matching[a] (bw = 1.0)	(9) Local linear regression adjusted matching (bw = 4.0)
(A) *Comparison group*: *CPS male sample*									
LaLonde sample with DW prop. score model	−9757	−555	−270	−838	−1299	−1380	−1431	−1406	−1329
	(255)	(596)	(493)	(628)	(529)	(437)	(441)	(490)	(441)
As % of $886 impact	−1101%	−63%	−30%	−95%	−147%	−156%	−162%	−159%	−150%
	(29)	(67)	(56)	(71)	(60)	(49)	(50)	(55)	(50)
DW sample with DW	−10291	407	−5	−27	−261	−88	−67	−127	−96
prop. score model	(306)	(698)	(672)	(723)	(593)	(630)	(611)	(709)	(643)
As % of $1794 impact	−574%	23%	−0.3%	−1.5%	−15%	−5%	−4%	−5%	−7%
	(17)	(39)	(37)	(40)	(33)	(35)	(34)	(40)	(36)
Early RA sample with	−11101	−7781	−3632	−5417	−2396	−3427	−2191	−3065	−3391
DW prop. score model	(461)	(1245)	(1354)	(1407)	(1152)	(1927)	(1069)	(3890)	(1124)
As % of $2748 impact	−404%	−283%	−132%	−197%	−87%	−125%	−80%	−112%	−123%
	(17)	(45)	(49)	(51)	(42)	(70)	(39)	(142)	(41)
LaLonde sample with	−10227	−3602	−2122	−3586	−2342	−3562	−2708	−3435	−2362
LaLonde prop. score model	(296)	(1459)	(1299)	(1407)	(1165)	(3969)	(1174)	(4207)	(1178)
As % of $886 impact	−1154%	−406%	−240%	405%	264%	402%	306%	388%	−266%
	(33)	(165)	(147)	(159)	(131)	(448)	(133)	(474)	(133)

(B) *Comparison group*: *PSID male sample*									
LaLonde sample with	−587	−2932	−2119	−166	−898	−1237	−1283	−587	−817
DW prop. score model	(264)	(898)	(787)	(959)	(813)	(747)	(633)	(1059)	(681)
As % of $886 impact	−1882%	−331%	−239%	−19%	−101%	−140%	−145%	−66%	−92%
	(30)	(101)	(89)	(108)	(92)	(84)	(71)	(120)	(77)
DW sample with DW	−16999	361	−82	447	−85	−122	143	693	777
prop. score model	(330)	(924)	(1200)	(827)	(1308)	(1362)	(633)	(2092)	(833)
As % of $1794 impact	−947%	20%	−5%	25%	−5%	−7%	8%	39%	43%
	(18)	(52)	(67)	(46)	(73)	(76)	(35)	(117)	(46)
Early RA sample with	−16993	−6132	−3570	−5388	−3337	−1946	−3262	−4602	−4475
DW prop. score model	(555)	(1237)	(1315)	(1487)	(1222)	(1079)	(936)	(1036)	(1010)
As % of $2748 impact	−618%	−223%	−130%	−196%	−121%	−71%	−119%	−256%	−249%
	(20)	(45)	(48)	(54)	(44)	(39)	(34)	(38)	(37)
LaLonde sample with	−16464	−3878	−3054	−3838	−2977	−3689	−3522	−3708	−3512
LaLonde prop. score model	(262)	(872)	(1080)	(872)	(985)	(976)	(964)	(1179)	(1042)
As % of $886 impact	−1858%	−438%	−345%	−433%	−336%	−416%	−397%	−419%	−396%
	(30)	(98)	(122)	(98)	(111)	(110)	(109)	(133)	(118)

[a] Regression adjustment includes race and ethnicity, age categories, education categories and married.

experimental samples. Fifth, when the LaLonde propensity score model is applied to the LaLonde sample, it does quite poorly in terms of bias, though not as poorly as the DW scores in the Early RA sample. Thus, the results obtained by DW (1999, 2002) using simple nearest-neighbor matching with their propensity scores are highly sensitive to changes in the sample composition. Moreover, adopting a reasonable alternative propensity score specification strongly increases the estimated bias in the full LaLonde sample.

The remaining four columns present estimates obtained using local linear matching methods. The sixth and seventh columns report estimates obtained using regular local linear matching with two different bandwidths. Increasing the bandwidth will, in general, increase the bias and reduce the variance associated with the estimator by putting a heavier weight on the information provided by more distant observations in constructing the counterfactual for each $D = 1$ observation. Interestingly, in Panel A of Table 5, both the variance and the overall average bias usually decrease when we increase the bandwidth. The final two columns present estimates obtained using regression-adjusted local linear matching, again with two different bandwidths. The notes to Table 5 list the variables used to do the regression adjustment.

The lessons from the local linear matching estimates are largely the same as those from the nearest-neighbor estimates. The DW scores do well in their sample, but have much larger biases in the LaLonde sample and in the Early RA sample. The LaLonde scores have large biases in his sample. Once again, the results in DW (1999, 2002) are sensitive to changes in the sample and, in the full LaLonde sample, an alternative propensity score specification yields much larger biases. The one additional finding is that, consistent with HIT (1997a), the matching estimates do not show much sensitivity, at least in terms of the qualitative conclusion they provide, to changing the fixed bandwidth from 1.0 to 4.0 and the local linear matching results do not change much when we use regression adjusted matches.

Panel B of Table 5 presents estimates analogous to those in Panel A of Table 5 but constructed using the PSID comparison group. The unadjusted mean differences shown in the first column are substantially larger here than with the CPS comparison group, presumably due to the sample restrictions imposed in constructing the CPS sample but not in the PSID sample. Thus, at some level, matching faces a tougher challenge with this comparison group. In practice, despite the larger raw mean differences, the bias estimates in Panel B of Table 5 are comparable to those in Panel A of Table 5.

Overall, the performance of the cross-sectional matching estimators is a bit worse than that found in HIT (1997a) and HIST (1998a). These estimators reduce the bias substantially relative to an unadjusted comparison of means, but the bias that remains after matching is typically somewhat larger than the corresponding experimental impact estimate. For the DW scores applied to the DW sample, we find that the matching estimators perform extremely well. However, as discussed above, the DW sample is somewhat peculiar in only including persons randomized after April of 1976 who had zero earnings in months 13–24 prior to randomization.

Because we find it difficult to motivate this type of sample inclusion criteria, we do not believe that the evidence that matching performs well on this particular sample can be generalized. Clearly, the performance of the matching estimators is much less impressive when applied to samples other than that analyzed in DW (1999, 2002).

8.2. *Difference-in-differences matching estimates*

Panels A and B of Table 6 present difference-in-differences matching estimates for the CPS and PSID comparison groups, respectively. These estimators have not previously been applied to the NSW data. As described in Section 3.3, difference-in-differences matching differs from cross-sectional matching in that it removes any time-invariant differences between the $D = 1$ and $D = 0$ groups conditional on $P(Z)$. This is accomplished in our context by subtracting a cross-sectional matching estimate of the pre-random-assignment bias from a cross-sectional matching estimate of the post-random assignment bias. In constructing the difference-in-differences matching estimates presented in Table 6, we use the same matching methods used in Table 5.

Consider Panel A of Table 6 and the CPS comparison group first. Four major patterns emerge. First, all of the difference-in-differences matching estimators perform well with the DW scores applied to the DW sample. This finding mirrors that for the cross-sectional matching estimators. Second, the bias associated with the difference-in-differences matching estimators is lower in most cases for the DW scores and the Early RA sample and in all cases with the LaLonde scores applied to the LaLonde sample. As a result, the biases associated with difference-in-differences propensity score matching are of the same order of magnitude as the impact (or smaller) for all of the samples and scores in Panel A of Table 6. Third, as in Panel A of Table 5 for the cross-sectional matching estimators, the particular estimator selected, the imposition of the common support condition and the choice of bandwidth all have no consistent effect on the estimated bias. Finally, and most importantly, when either the score model or the sample is changed, the estimated bias increases substantially, though less than in the case of the cross-sectional matching estimators considered in Table 5. Once again, the bias estimates are not robust to perturbations in the sample or in the propensity score model, mirroring the findings for the cross-sectional matching estimators.

The estimates for the PSID comparison group, presented in Panel B of Table 6, reveal even stronger patterns. While the biases for the DW sample with the DW scores get a bit larger with differencing, the biases for the other three combinations of scores and samples presented in the table all get substantially smaller. Especially dramatic are the changes for the Early RA sample with the DW scores and for the LaLonde sample with the LaLonde scores, where the biases often fall from several thousand dollars to only a few hundred. As was the case with the CPS comparison group, the biases show no consistent pattern in response to the choice of matching procedure, the imposition of the common support condition or the selection of the bandwidth.

Table 6

Bias associated with alternative difference-in-differences matching estimators. Comparison groups: (A) CPS male sample and (B) PSID male sample. Difference between real earnings in 1978 and real earnings in 1975 (bootstrap standard errors in parentheses; trimming level for common support is 2 percent)

Sample and propensity score model	(1) Mean diff.	(2) 1 Nearest neighbor without common support	(3) 10 Nearest-neighbors without common support	(4) 1 Nearest-neighbor with common support	(5) 10 Nearest-neighbors with common support	(6) Local linear matching (bw = 1.0)	(7) Local linear matching (bw = 4.0)	(8) Local linear regression adjusted matching[a] (bw = 1.0)	(9) Local linear regression adjusted matching (bw = 4.0)
(A) *Comparison group*: *CPS male sample*									
LaLonde sample with DW	867	−1527	−1317	−929	−1064	−1212	−1271	−1212	−1271
prop. score model	(314)	(563)	(520)	(554)	(539)	(483)	(472)	(524)	(475)
As % of $886 impact	98%	−172%	−149%	−105%	−120%	−137%	−143%	−137%	−143%
	(35)	(64)	(59)	(63)	(61)	(55)	(53)	(59)	(54)
DW sample with DW prop.	2093	45	−101	−607	−417	−88	−75	−88	−75
score model	(365)	(781)	(689)	(784)	(681)	(629)	(621)	(848)	(719)
As % of $1794 impact	117%	3%	−6%	−34%	−23%	−5%	−4%	−5%	−4%
	(20)	(44)	(38)	(44)	(38)	(35)	(34)	(47)	(40)
Early RA sample with DW	598	1398	1041	1689	3200	2993	2909	1876	1461
prop. score model	(549)	(1342)	(1166)	(1212)	(1108)	(3152)	(917)	(4021)	(1521)
As % of $2748 impact	22%	51%	38%	61%	116%	109%	106%	68%	53%
	(20)	(49)	(42)	(44)	(40)	(115)	(33)	(146)	(55)
LaLonde sample with	897	−463	1317	−21	1229	192	927	193	928
LaLonde prop. score model	(333)	(1290)	(878)	(1092)	(862)	(1102)	(801)	(3970)	(1466)
As % of $886 impact	101%	−52%	149%	−2%	138%	22%	105%	−16%	105%
	(38)	(146)	(99)	(123)	(97)	(124)	(90)	(448)	(165)

(B) *Comparison group*: *PSID male sample*									
LaLonde sample with DW prop. score model	−383 (318)	−1644 (1033)	−148 (931)	608 (1070)	−568 (939)	188 (823)	79 (686)	−344 (1127)	−318 (733)
As % of $886 impact	−43% (36)	−186% (117)	−17% (105)	69% (121)	−64% (106)	21% (93)	9% (77)	−39% (127)	−36% (83)
DW Sample with DW prop. score model	797 (362)	537 (1031)	725 (1208)	568 (906)	737 (1366)	286 (1414)	803 (792)	287 (2173)	803 (1058)
As % of $1794 impact	44% (20)	30% (57)	40% (67)	32% (51)	41% (76)	16% (79)	145% (44)	16% (121)	45% (59)
Early RA sample with DW prop. score model	−133 (629)	−46 (1131)	1135 (1266)	316 (1276)	1153 (1273)	2118 (1016)	1018 (993)	207 (1375)	111 (1352)
As % of $2748 impact	−5% (23)	−2% (41)	41% (46)	11% (46)	42% (46)	77% (37)	37% (36)	8% (50)	4% (49)
LaLonde sample with LaLonde prop. score model	−427 (311)	−381 (980)	263 (1115)	−364 (929)	238 (1063)	−204 (969)	39 (1009)	−204 (1323)	39 (1172)
As % of $886 impact	−48% (35)	−43% (111)	30% (126)	−41% (105)	27% (120)	−23% (109)	4% (114)	−23% (149)	4% (132)

[a] Regression adjustment includes race and ethnicity, age categories, education categories and married.

While the cross-sectional matching estimates presented in Table 5 reveal the extreme sensitivity of the results in DW (1999, 2002), the estimates in Table 6 show fairly stable performance for the difference-in-differences matching estimators. These results differ from the findings in HIT (1997a) and HIST (1998a) in the sense that for most demographic groups in the JTPA data, the biases associated with difference-in-differences matching are quite similar to those associated with cross-sectional matching. The difference between the findings here and those from the JTPA data is consistent with the view that the differencing is eliminating time-invariant bias in the NSW data due to geographic mismatch and/or different ways of measuring earnings in the experimental control and nonexperimental comparison groups, which were not sources of bias with the JTPA data. The very limited set of conditioning variables Z available in the NSW data, compared to the rich set of conditioning variables available in the JTPA data, may also help to explain the much larger difference between the cross-sectional and difference-in-differences matching estimates of the bias obtained using the NSW data.

9. Regression-based estimates

We next present bias estimates obtained using a number of standard, regression-based impact estimators for each of the three experimental samples and both comparison groups. We seek answers to two questions. First, how well do these estimators perform in the different samples? We have argued that the DW sample may implicitly present a less difficult selection problem than the original LaLonde sample due to its inclusion of persons randomly assigned late in the experiment only if they had zero earnings in months 13–24 prior to random assignment. Second, is it the matching estimator or just selection of the right conditioning variables that accounts for the low bias estimates when cross-sectional propensity score matching estimators are applied to the DW sample with the DW scores? Both matching and standard regression adjustment seek to correct for selection on observable characteristics, Y_0. Differences between the two are that matching, unlike regression, does not assume a linear (in the parameters) functional form and does not require $\mathrm{E}(U|X,D)=0$.

Panels A and B of Table 7 give the bias estimates for the CPS and PSID comparison group samples, respectively. In each table, each pair of rows contains the estimates of the bias and of the bias as a percentage of the impact for one of the three experimental samples. The first column presents the simple mean difference in earnings in 1978. The next four columns present bias estimates for cross-sectional regression specifications based on Eq. (4) in Section 3. The models containing varying sets of conditioning variables, including the variables from the LaLonde propensity scores, the DW propensity scores, the DW scores without the "Real Earnings in 1974" variable and a richer specification that includes additional interaction terms found to be significant in an investigation of alternative propensity score models. An exact variable list for each specification appears in the table notes. The last four columns of Table 7 show bias estimates from the difference-in-differences estimators and unrestricted difference-in-differences estimators examined in Table 5 of LaLonde (1986). The difference between the two pairs of estimators is that in the first pair, based on Eq. (5), the dependent variable is the difference between earnings in 1978 and earnings in 1975, while in the second pair, the dependent variable is earnings in 1978 and earnings in 1975 is included as a right-hand-side variable. The latter formulation relaxes the restriction implicit in the former that the coefficient on 1975 earnings equals one.[45]

The estimates in Table 7 gives clear answers to both questions raised. Comparing the bias estimates from the LaLonde and Early RA samples reveals that for the standard regression estimators and the unrestricted difference-in-difference estimators, the bias is smallest in the DW sample in every case but one. This strongly suggests that the sub-sampling strategy employed by DW (1999, 2002) results in a sample with a selection problem that is less difficult to solve.[46] The exceptions to this rule are the two standard difference-in-differences estimators. Having selected into the sample persons who may have transitorily, rather than permanently, low earnings, it is perhaps not surprising that differencing does relatively poorly in the DW sample. This pattern is also consistent with the fact that difference-in-differences matching tends to increase the bias (a bit for the CPS comparison group and a bit more for the PSID comparison group) relative to cross-sectional matching for the DW sample, but not for the LaLonde and Early RA samples.[47]

In regard to the second question, the results differ between the CPS and PSID comparison groups. In the CPS sample, the bias estimate from a regression of earnings in 1978 on an NSW indicator (equal to one for the control group members and zero otherwise) and the covariates from the DW propensity score model is $-\$34$ (2% of the experimental impact). Thus, for the CPS comparison group, the key to the low bias estimates found in DW (1999, 2002) is picking the right subsample and the right covariates, not matching. In contrast, in the PSID, matching makes a big difference. The bias estimate from nearest-neighbor matching with ten nearest neighbors (and imposing the common support condition) is $-\$85$, compared to a bias estimate from a regression using the same variables of \$1285. For the PSID, the linearity restriction implicit in the regression has some bite.

[45] We also estimated the bias for the before–after estimator, described in Section 3.2, associated with each experimental sample. In each case, the bias was on the order of several thousand dollars. We do not present estimates from the Heckman (1979) two-step estimator of the bivariate normal selection model examined by LaLonde (1986) as his estimates do not take account of choice-based sampling.

[46] This finding is implicit in Table 2 of DW (1999). Compare the estimated coefficients (not biases) for LaLonde's sample to those for their sample both with and without including the "Real Earnings in 1974" variable among the covariates for the CPS-1 and PSID-1 comparison groups.

[47] It is also of interest to note that the estimated biases for the regression-adjustment and unrestricted difference-in-differences models are almost always lower with the CPS comparison group than with the PSID comparison group. This indicates the value of the additional sample restrictions imposed on the CPS comparison group when the estimator employed is simple regression adjustment.

Table 7

Bias associated with alternative regression-based estimators. Comparison groups: (A) CPS male sample and (B) PSID male sample. Dependent variable: real earnings in 1978 (estimated standard errors in parentheses)

Sample and propensity score model	(1) Mean diff.	(2) Regression with LaLonde covariates[a]	(3) Regression with DW covariates[b]	(4) Regression with DW covariates without RE74	(5) Regression with rich covariates[c]	(6) Difference in-differences	(7) Differences-in differences with age included	(8) Unrestricted difference-in differences[d]	(9) Unrestricted difference-in differences
(A) *Comparison group*: *CPS male sample*									
LaLonde sample	−9756 (470)	−1616 (410)	−1312 (388)	−1466 (393)	−974 (451)	868 (379)	−522 (371)	−2405 (357)	−1906 (388)
As % of $886 impact	−1101%	−182%	−148%	−165%	−110%	98%	−60%	−271%	−215%
DW sample	−10292 (600)	−690 (505)	−34 (486)	−238 (489)	625 (555)	2092 (481)	802 (470)	−1691 (454)	−1089 (479)
As % of $1794 impact	−574%	−38%	−2%	−13%	35%	117%	45%	−94%	−61%
Early RA sample	−10238 (811)	−1384 (655)	−1132 (620)	−1179 (629)	−301 (707)	1136 (649)	−5 (634)	−2337 (608)	−1723 (625)
As % of $2748 impact	−373%	−50%	−41%	−43%	−11%	41%	−0%	−85%	−63%

(B) *Comparison group*: *PSID male sample*									
LaLonde sample	−16037 (668)	−2632 (783)	−2540 (756)	−2448 (751)	−2111 (808)	−427 (543)	−1836 (573)	−3263 (580)	−3192 (665)
As % of $886 impact	−1810%	−297%	−287%	−276%	−238%	−48%	−207%	−368%	−360%
DW sample	−17796 (846)	−920 (940)	−1285 (960)	−1076 (920)	−492 (993)	797 (683)	−497 (704)	−2172 (720)	−1969 (791)
As % of $1794 impact	−992%	−51%	−72%	−60%	−27%	44%	−28%	−121%	−110%
Early RA sample	−16945 (1311)	−1850 (1161)	−1949 (1072)	−1720 (1057)	−820 (1139)	−159 (920)	−1347 (929)	−2951 (936)	−2824 (981)
As % of $2748 impact	−617%	−67%	−71%	−63%	−30%	−6%	−49%	−107%	−103%

[a] The LaLonde Covariates are the variables from the LaLonde (1986) propensity score model.

[b] The DW Covariates are the variables from the Dehejia and Wahba (2002) propensity score model.

[c] The Rich Covariates model includes indicators for age categories, interactions between the age categories and racial and ethnic group, education categories, a marriage indicator, interactions between the marriage indicator and race and ethnicity, real earnings in 1975 and its square, an indicator for zero earnings in 1975, number of children, and number of children interacted with race and ethnicity.

[d] Unrestricted difference-in-differences refers to a regression with real earnings in 1978 on the left-hand side and real earnings in 1975 on the right-hand side. In the specification with covariates, the covariates are age, age squared, years of schooling, high school dropout, and indicators for black and hispanic. This specification follows that in LaLonde (1986).

10. Specification tests

As discussed in Section 2, Heckman and Hotz (1989) found that when they applied two types of specification tests to the NSW data that they were able to rule out those estimators that implied a different qualitative conclusion than the experimental impact estimates. In this section, we apply one of the specification tests that they use to the cross-sectional matching estimators presented in Table 5. The test we apply is the pre-program alignment test, in which each candidate estimator is applied to outcome data from a period prior to the program (i.e., to random assignment). Note that this test actually tests the joint null that the outcome and participation processes are the same in the pre-program and post-program periods and that the estimator being tested successfully corrects for selection bias.[48]

We implement the test by applying the matching estimators to earnings in 1975, keeping the same propensity scores. If the estimated bias is statistically different from zero in the pre-program period, then we reject the corresponding estimator. Because we lack reliable earnings data for two pre-program periods, we are unable to apply the test to the difference-in-differences matching estimators in Table 6.[49]

Panels A and B of Table 8 present the pre-program estimates for the CPS and PSID comparison groups, respectively. Consider first Panel A of Table 8. The pre-program test rejects every estimator for the Early RA sample with the DW scores, which is good, as the biases are all quite high for this sample in Panel A of Table 5. It also rejects all but one of the estimators for the LaLonde sample with the LaLonde scores (though two are rejected only at the 10 percent level), which is of course desirable given the large bias values. The test does not reject any of the very low bias estimators for the DW sample with the DW scores. In the case of the LaLonde sample with the DW scores, where the biases are of moderate size, the first two of the eight estimators in Panel A of Table 5 are rejected. Overall, the pre-program test applied to the CPS comparison group does a good job of eliminating the estimators with the highest estimated biases in the post-program period and not rejecting the estimators with low or moderate estimated biases.

Similar patterns are observed in Panel B of Table 8 for the PSID comparison group. The pre-program test solidly rejects all of the matching estimators as applied to the Early RA sample with the DW scores and to the LaLonde sample with the LaLonde scores. All of these estimators have very large estimated biases in the post-program period. The test does not reject any of the matching estimators for the DW scores applied to the DW sample, which have low estimated biases in the post-program period. Finally, the test results for the DW scores applied to the LaLonde sample are again a mixed bag, though in this case the four estimators eliminated by

[48] See Heckman and Hotz (1989) for a more detailed discussion of the test and Heckman et al. (1999) for a discussion of caveats regarding its use. Ham et al. (2003) apply pre-program specification tests in the context of controlling for selectivity in estimating the returns to migration and find them very useful.

[49] Recall that we are not using the grouped data on SSA earnings that Heckman and Hotz (1989) use in their paper, and which allow them to apply the pre-program test to longitudinal estimators where it requires multiple periods of pre-program data.

the pre-program test are the four with the highest estimated biases in the post-program period. Overall, for both comparison group samples, our results confirm the effectiveness of the pre-program test at calling attention to estimators likely to lead to highly biased estimates. Thus, we reach for cross-sectional matching estimators a similar conclusion to that reached by Heckman and Hotz (1989) in regard to the standard regression-based estimators they examined.

11. Summary and conclusions

Our analysis of the data from the National Supported Work Demonstration yields three main conclusions. First, our evidence leads us to question recent claims in the literature by DW (1999, 2002) and others regarding the general effectiveness of matching estimators relative to more traditional econometric methods. While we are able to replicate the low bias estimates reported in the DW (1999, 2002) studies, we conclude that their evidence is not generalizable. When we apply the same methods to other reasonable samples from the NSW data, the low bias results disappear. When we construct estimates using a modestly different propensity score specification and the full LaLonde sample, we obtain much larger biases than with the DW propensity score specification. The sample inclusion rules employed by DW (1999, 2002) in creating their sample simplify the selection problem by differentially including individuals with zero earnings in the pre-program period. Indeed, in some cases even very simple regression-adjustment estimators have low bias values when applied to the DW sample. Thus, their evidence clearly cannot be construed as showing the superiority of matching over more traditional econometric estimators. More generally, we argue that their study, like much of the earlier literature in this area, implicitly poses the wrong question. The question is not which estimator is the best estimator always and everywhere. Estimators differ in their identifying assumptions, and the assumptions underlying a given estimator will sometimes hold in the data and sometimes fail to hold. Instead of engaging in a hopeless search for a magic bullet estimator, the goal of theoretical and empirical investigation should be to develop a mapping from the characteristics of the data and institutions available in particular evaluation contexts to the optimal nonexperimental estimators for those contexts. In some contexts, particularly those with high-quality data rich in variables related to participation and outcomes, matching may be the best choice. In other cases, such as the NSW data, our results show that matching makes a poor choice.

Second, we find that the difference-in-differences matching estimators introduced in HIT (1997a) and HIST (1998a) perform substantially better than the corresponding cross-sectional matching estimators. This finding is consistent with the elimination of time-invariant biases between the NSW sample and the comparison group sample due to geographic mismatch and differences in the measurement of the dependent variable. Matching methods do not perform well in eliminating these sources of bias, a task for which they were not designed. The positive findings regarding difference-in-differences matching again highlight the importance of choosing a nonexperimental method consistent with the features of

Table 8
Bias associated with alternative cross-sectional matching estimators. Comparison groups: (A) CPS male sample and (B) PSID male sample. Dependent variable: real earnings in 1975 (bootstrap standard errors in parentheses; trimming level for common support is 2 percent)

Sample and propensity score model	(1) Mean diff.	(2) 1 Nearest neighbor without common support	(3) 10 Nearest-neighbors without common support	(4) 1 Nearest-neighbor with common support	(5) 10 Nearest-neighbors with common support	(6) Local linear matching (bw = 1.0)	(7) Local linear matching (bw = 4.0)	(8) Local linear regression adjusted matching[a] (bw = 1.0)	(9) Local linear regression adjusted matching (bw = 4.0)
(A) *Comparison group*: *CPS male sample*									
LaLonde sample with DW prop. score model	−10624	972	1047	91	−235	−168	−160	−194	−58
	(254)	(314)	(258)	(399)	(342)	(315)	(333)	(280)	(270)
As % of $886 impact	−1199%	110%	118%	10%	−27%	−19%	−18%	−22%	−7%
	(29)	(35)	(29)	(45)	(39)	(36)	(38)	(32)	(30)
DW sample with DW prop. score model	−12383	362	96	580	156	0	8	−39	−21
	(172)	(248)	(199)	(339)	(268)	(196)	(203)	(274)	(212)
As % of $1794 impact	−690%	20%	5%	33%	9%	0%	0%	−2%	−1%
	(10)	(14)	(11)	(19)	(15)	(11)	(11)	(15)	(12)
Early RA sample with DW prop. score model	−11700	−9179	−4673	−7106	−5596	−6420	−5100	−4941	−4852
	(354)	(1769)	(1132)	(1357)	(953)	(3903)	(939)	(1427)	(982)
As % of $2748 impact	−426%	−334%	−170%	−259%	−204%	−234%	−186%	−180%	−177%
	(13)	(64)	(41)	(49)	(35)	(142)	(34)	(52)	(36)
LaLonde sample with LaLonde prop. score model	−11124	−3139	−3439	−3565	−3571	−3754	−3635	−3628	−3290
	(224)	(1845)	(1090)	(1889)	(1078)	(4507)	(1103)	(1679)	(770)
As % of $886 impact	−1255%	−354%	−388%	−402%	−403%	−424%	−410%	−409%	−371%
	(25)	(208)	(123)	(213)	(122)	(509)	(124)	(190)	(87)

(B) *Comparison group*: *PSID male sample*									
LaLonde sample with DW	−16293	−1288	−1971	−442	−1466	−161	−1362	−243	−499
prop. score model	(238)	(673)	(524)	(631)	(524)	(547)	(456)	(507)	(435)
As % of $886 impact	−1839%	−145%	−222%	−50%	−165%	−18%	−154%	−27%	−56%
	(27)	(76)	(59)	(71)	(59)	(62)	(51)	(57)	(49)
DW sample with DW prop.	−17796	−176	−807	−121	−822	−408	−660	406	−26
score model	(194)	(443)	(676)	(304)	(746)	(518)	(435)	(962)	(644)
As % of $1794 impact	−992%	−10%	−45%	−7%	−46%	−23%	−37%	23%	−1%
	(11)	(25)	(38)	(17)	(42)	(29)	(24)	(54)	(36)
Early RA sample with DW	−16780	−6086	−4705	−5704	−4490	−4064	−4280	−4809	−4586
prop. score model	(374)	(771)	(778)	(984)	(770)	(690)	(701)	(853)	(811)
As % of $2748 impact	−611%	−221%	−171%	−208%	−163%	−148%	−156%	−268%	−255%
	(14)	(28)	(28)	(36)	(28)	(25)	(26)	(31)	(30)
LaLonde sample with	−16036	−3497	−3317	−3474	−3215	−3485	−3561	−3504	−3551
LaLonde prop. score model	(213)	(624)	(712)	(779)	(740)	(597)	(629)	(604)	(623)
As % of $886 impact	−1810%	−395%	−374%	−392%	−363%	−393%	−402%	−395%	−401%
	(24)	(70)	(80)	(88)	(84)	(67)	(71)	(68)	(70)

[a] Regression adjustment includes race and ethnicity, age categories, education categories and married.

the data and institutions present in a given context. In the NSW context, the data are weak in covariates, fail to place comparison group members in the same local labor markets as participants and rely on different measures of earnings for participants and non-participants. These features strongly suggest that matching should work poorly in the NSW data and that differences-in-differences matching should work better, which is precisely what we find. As this example shows, knowledge regarding which estimators work given the characteristics of the available data and the institutional context has begun to accumulate in the literature and should be used in designing and evaluating evaluation research.

Third, we find that while the choice between cross-sectional matching and difference-in-differences matching makes a big difference to the estimated biases, the details of the matching procedure in general do not. Thus, the choice between nearest neighbor and local linear matching, or the choice of bandwidth for local-linear matching (within reasonable limits), do not have strong or consistent effects on the estimated biases. This finding comports with the findings in a number of other empirical studies. The imposition of the common support condition represents the one (partial) exception in our context, as it affects the estimated biases in some cases but not others.

Acknowledgements

We thank Robert LaLonde for providing us with the data from his 1986 study. We thank Rajeev Dehejia for providing us with information helpful in reconstructing the samples used in the Dehejia and Wahba (1999, 2002) studies. We thank seminar participants at Boston College, the CEA meetings, CILN, the Department of Family and Community Services of Australia, Econometric Society (European meetings), the GAAC conference on the Evaluation of Active Labor Market Policies, IFS, IRP, IZA, Kentucky, Laval, Mathematica, McGill, Pittsburg, PSI, Princeton, St. Gallen, SEA meetings, Stockholm, UCSB, UNC, Wisconsin, WRNET, and Yale for their helpful comments. We thank Sascha Becker, Dan Black, Per-Anders Edin, Markus Frölich, John Ham, Robert LaLonde, Michael Lechner, Thomas Lemieux, Miana Plesca, Mike Veall and two anonymous referees for useful comments. Jingjing Hsee and Miana Plesca provided excellent research assistance. We are grateful to James Heckman for his encouragement and for financial resources to support Jingjing Hsee. Smith's participation in this project was supported by the Social Science and Humanities Research Council of Canada and the CIBC Chair in Human Capital and Productivity at the University of Western Ontario and Todd's by the U.S. National Science Foundation (SBR-9730688).

References

Angrist, J., 1990. Lifetime earnings and the Vietnam draft lottery: evidence from Social Security Administrative records. American Economic Review 80 (3), 313–335.

Angrist, J., 1998. Estimating the labor market impact of voluntary military service using Social Security data on military applicants. Econometrica 66 (2), 249–288.

Angrist, J., Hahn, J., 1999. When to control for covariates? Panel asymptotics for estimates of treatment effects. NBER Technical Working Paper No. 241.

Ashenfelter, O., 1978. Estimating the effect of training programs on earnings. Review of Economics and Statistics 60, 47–57.

Ashenfelter, O., 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., 1987. The impact of CETA programs on earnings: a review of the literature. Journal of Human Resources 22, 157–193.

Bassi, L., 1984. Estimating the effects of training programs with nonrandom selection. Review of Economics and Statistics 66, 36–43.

Blundell, R., Costa-Dias, M., 2000. Evaluation methods for non-experimental data. Fiscal Studies 21 (4), 427–468.

Burtless, G., 1995. The case for randomized field trials in economic and policy research. Journal of Economic Perspectives 9 (2), 63–84.

Burtless, G., Orr, L., 1986. Are classical experiments needed for manpower policy? Journal of Human Resources 21, 606–639.

Card, D., Sullivan, D., 1988. Measuring the effect of subsidized training programs on movements in and out of employment. Econometrica 56 (3), 497–530.

Cochran, W., Rubin, D., 1973. Controlling bias in observational studies. Sankyha 35, 417–446.

Couch, K., 1992. New evidence on the long-term effects of employment and training programs. Journal of Labor Economics 10 (4), 380–388.

Dehejia, R., Wahba, S., 1999. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. Journal of the American Statistical Association 94 (448), 1053–1062.

Dehejia, R., Wahba, S., 2002. Propensity score matching methods for nonexperimental causal studies. Review of Economics and Statistics 84 (1), 151–161.

Devine, T., Heckman, J., 1996. The structure and consequences of eligibility rules for a social program: a study of the job training partnership act (JTPA). In: Polacheck, S. (Ed.), Research in Labor Economics, Vol. 15. JAI Press, Greenwich, CT, pp. 111–170.

Eberwein, C., Ham, J., LaLonde, R., 1997. The impact of being offered and receiving classroom training on the employment histories of disadvantaged women: evidence from experimental data. Review of Economic Studies 64 (4), 655–682.

Eichler, M., Lechner, M., 2002. An evaluation of public employment programmes in the East German state of Sachsen-Anhalt. Labour Economics 9, 143–186.

Fan, J., 1992a. Design adaptive nonparametric regression. Journal of the American Statistical Association 87, 998–1004.

Fan, J., 1992b. Local linear regression smoothers and their minimax efficiencies. The Annals of Statistics 21, 196–216.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman & Hall, New York.

Fraker, T., Maynard, R., 1987. The adequacy of comparison group designs for evaluations of employment related programs. Journal of Human Resources 22, 194–227.

Friedlander, D., Robins, P., 1995. Evaluating program evaluations: new evidence on commonly used nonexperimental methods. American Economic Review 85 (4), 923–937.

Frölich, M., 2004. Finite-sample properties of propensity score matching and weighting estimators. Review of Economics and Statistics 86 (1), 77–90.

Hahn, J., 1998. On the role of the propensity score in efficient estimation of average treatment effects. Econometrica 66 (2), 315–331.

Ham, J., Li, X., Reagan, P., 2003. Propensity score matching, a distance-based measure of migration, and the wage growth of young men. Working paper, Department of Economics, Ohio State University.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153–161.

Heckman, J., 1992. Randomization and social policy evaluation. In: Manski, C., Garfinkel, I. (Eds.), Evaluating Welfare and Training Programs. Harvard University Press, Cambridge, MA, pp. 201–230.

Heckman, J., 1997. Randomization as an instrumental variables estimator: a study of implicit behavioral assumptions in one widely-used estimator. Journal of Human Resources 32, 442–462.

Heckman, J., 2001. Micro data, heterogeneity, and the evaluation of public policy: nobel lecture. Journal of Political Economy 109 (4), 673–748.

Heckman, J., Hotz, J., 1989. Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. Journal of the American Statistical Association 84 (408), 862–880.

Heckman, J., Navarro-Lazano, S., 2004. Using matching, instrumental variables and control functions to estimate economic choice models. Review of Economics and Statistics 86 (1), 30–57.

Heckman, J., Robb, R., 1985. Alternative methods for evaluating the impact of interventions. In: Heckman, J., Singer, B. (Eds.), Longitudinal Analysis of Labor Market Data. Cambridge University Press, New York, pp. 156–246.

Heckman, J., Smith, J., 1995. Assessing the case for social experiments. The Journal of Economic Perspectives 9, 85–110.

Heckman, J., Smith, J., 1999. The pre-programme earnings dip and the determinants of participation in a social programme: implications for simple programme evaluation strategies. Economic Journal 109 (457), 313–348.

Heckman, J., Todd, P., 1995. Adapting propensity score matching and selection models to choice-based samples. Working Paper, Department of Economics, University of Chicago.

Heckman, J., Vytlacil, E., 2001. Policy relevant treatment effects. American Economic Review 91 (2), 107–111.

Heckman, J., Ichimura, H., Smith, J., Todd, P., 1996. 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 (23), 13416–13420.

Heckman, J., Ichimura, H., Todd, P., 1997a. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Review of Economic Studies 64 (4), 605–654.

Heckman, J., Smith, J., Clements, N., 1997b. Making the most out of social experiments: accounting for heterogeneity in programme impacts. Review of Economic Studies 64 (4), 487–536.

Heckman, J., Ichimura, H., Smith, J., Todd, P., 1998a. Characterizing selection bias using experimental data. Econometrica 66 (5), 1017–1098.

Heckman, J., Ichimura, H., Todd, P., 1998b. Matching as an econometric evaluation estimator. Review of Economic Studies 65 (2), 261–294.

Heckman, J., LaLonde, R., Smith, J., 1999. The economics and econometrics of active labor market programs. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, Vol. 3A. North-Holland, Amsterdam, pp. 1865–2097.

Heckman, J., Hohmann, N., Smith, J., Khoo, M., 2000. Substitution and drop out bias in social experiments: a study of an influential social experiment. Quarterly Journal of Economics 115 (2), 651–694.

Heckman, J., Tobias, J., Vytlacil, E., 2001. Four parameters of interest in the evaluation of social programs. Southern Economic Journal 68 (2), 210–223.

Hollister, R., Kemper, P., Maynard, R., 1984. The National Supported Work Demonstration. University of Wisconsin Press, Madison.

LaLonde, R., 1986. Evaluating the econometric evaluations of training programs with experimental data. American Economic Review 76, 604–620.

Lechner, M., 1999. Earnings and employment effects of continuous off-the-job training in East Germany after unification. Journal of Business and Economic Statistics 17, 74–90.

Lechner, M., 2000. An evaluation of public sector sponsored continuous vocational training programs in East Germany. The Journal of Human Resources 35, 347–375.

Lechner, M., 2002. Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. Journal of the Royal Statistical Society Series A 165 (Part 1), 59–82.

Manski, C., Lerman, S., 1977. The estimation of choice probabilities from choice-based samples. Econometrica 45 (8), 1977–1988.

Raaum, O., Torp, H., 2002. Labour market training in Norway—effect on earnings. Labour Economics 9 (2), 207–247.

Regnér, H., 2002. A nonexperimental evaluation of training programs for the unemployed in Sweden. Labour Economics 9 (2), 187–206.

Rosenbaum, P., Rubin, D., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rosenbaum, P., Rubin, D., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. American Statistician 39, 33–38.

Silverman, B., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall, London.

Zhao, Z., 2004. Using matching to estimate treatment effects: data requirements, matching metrics and Monte Carlo evidence. Review of Economics and Statistics 86 (1), 91–107.