Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?[1]

Jeffrey Smith
University of Western Ontario

Petra Todd
University of Pennsylvania[2]

November 22, 2000

[1]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 (1998,1999) studies. This research was presented at the Institute for Research on Poverty (June 2000) at the Western Research Network on Employment and Training summer workshop (August 2000), at the Canadian International Labour Network meetings (September 2000), the University of North Carolina and the Southern Economic Association meetings (November 2000). We thank Dan Black, Michael Lechner, Thomas Lemieux and Mike Veall 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 Todd's by the U.S. National Science Foundation (SBR-9730688).

[2]Both authors are also affiliated with the National Bureau of Economic Research. They may be contacted through email at jsmith@julian.uwo.ca or petra@athena.sas.upenn.edu.

Abstract

This paper applies recently developed cross-sectional and longitudinal propensity score matching estimators to data from the National Supported Work Demonstration that have been previously analyzed by LaLonde (1986) and Dehejia and Wahba (1998,1999). We find little support for recent claims in the econometrics and statistics literatures that traditional, cross-sectional matching estimators generally provide a reliable method of evaluating social experiments (e.g. Dehejia and Wahba, 1998, 1999). Our results show that program impact estimates generated through propensity score matching are highly sensitive to choice of variables used in estimating the propensity scores and sensitive to the choice of analysis sample. Among the estimators we study, the difference-in-differences matching estimator is the most robust. We attribute its better performance to the fact that it eliminates temporarily-invariant sources of bias that may arise, for example, when program participants and nonparticipants are geographically mismatched or from differences in survey questionnaires, which are both common sources of biases in evaluation studies.

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 distribution of both observed and unobserved characteristics as the treatment group. At the same time, social experimentation also has some drawbacks, such as (a) high cost, (b) the potential to distort the operation of an ongoing program, (c) the common problem of program sites refusing to participate in the experiment and (d) the problem of randomized-out controls seeking alternative forms of treatment.[1] In contrast, evaluation methods that use nonexperimental data tend to be less costly and less intrusive. 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.[2]

In this paper, we use experimental data combined with nonexperimental data to evaluate the performance of alternative nonexperimental estimators. The impact estimates based on experimental data provide a benchmark against which to judge the performance of nonexperimental estimators. Our experimental data come from the National Supported Work (NSW) Demonstration and the nonexperimental data from the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). These same data were used in the influential papers of LaLonde (1986), Heckman and Hotz (1989) and Dehejia and Wahba (hereafter DW) (1998,1999).

We focus on a class of estimators called *propensity score matching estimators*, which are increasingly being used in evaluation studies. Our study finds little support for some recent claims in the literature about the effectiveness of simple matching estimators as a method of controlling for selectivity bias in observational studies. In particular, we find that the low bias estimates obtained by DW (1998,1999) using various cross-sectional matching estimators are highly sensitive to their particular choice of subsample and to the variables used to estimate the propensity scores. We find that difference-in-differences (DID) matching estimators exhibit somewhat better performance than the cross-sectional methods. This may be due to the fact that DID estimators eliminate temporaly invariant sources of bias that may arise, for example, from geographic mismatch of program participants and nonparticipants or from differences in the questionnaires used to gather data from program participants and nonparticipants. In this sense, our findings using the NSW data are consistent with findings reported in Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1996,1998) using the experimental data from the U.S. National Job Training Partnership Act Study.

The plan of the paper is as follows. Section 2 reviews some key papers in the previous literature on the choice among alternative non-experimental estimators. Section 3.1 lays out the evaluation problem and Section 3.2 briefly describes commonly used non-experimental estimators. Section 3.3

[1] See, e.g., Burtless and Orr (1986), Heckman (1992), Burtless (1995), Heckman and Smith (1995), Heckman, LaLonde and Smith (1999) and Heckman, Hohmann, Khoo and Smith (2000).

[2] See, e.g., the sensitivity documented in Ashenfelter (1978), Bassi (1984), Ashenfelter and Card (1985), Lalonde (1986) and Fraker and Maynard (1987).

describes the cross-sectional and difference-in-differences matching estimators that we focus on in our study, while Section 3.4 explains our method of using the experimental data to benchmark the performance of non-experimental estimators. Sections 4 and 5 describe the National Supported Work Demonstration data subsamples that we use. Section 6 presents our estimated propensity scores and Section 7 discusses the related "balancing tests" used in some recent studies to aid in selecting a propensity score specification in recent studies. Sections 8 and 9 present bias estimates obtained using matching and regression-based estimators, respectively. Finally, 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 estimator of 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 is likely to be satisfied. Using a limited set of specification tests, Lalonde (1986) concludes that there is no good way 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 1980's movement towards using experiments to evaluate social programs.(See e.g. Burtless (1986, 1995).)

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.[3] The primary test they consider is based on preprogram data, so its validity depends on the assumption that the outcome and participation processes are similar in pre-program and post-program time periods.[4] When their specification tests are applied to the NSW data, Heckman and Hotz (1989) find that

[3]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 administrative data for the CPS comparison group and survey-based earnings measures for NSW participants and for the PSID comparison group. Because their administrative data do not suffer from attrition problems, their sample of AFDC women includes more of the total sample that participated in the experiment than does LaLonde's sample.

[4]We apply their tests below in section 10.

the tests exclude the estimators that would imply a substantially different qualitative conclusion (impact sign and statistical signficance) 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.

Traditional pairwise matching methods are extended in Heckman, Ichimura and Todd (1997,1998) and Heckman, Ichimura, Smith and Todd (1998) (henceforth HIT and HIST) in several ways. First, kernel and local linear matching estimators are described that use multiple nonparticipants in constructing each of the matched outcomes. The main advantage of these estimators vis-a-vis pairwise matching is a reduction in the variance of the estimator. Second, HIT and HIST propose modified versions of matching estimators that can be implemented when longitudinal or repeated cross-section data are available. These estimators accommodate time-invariant differences between participant and nonparticipant outcomes that are not eliminated by cross-sectional matching.

HIT and HIST 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 they examine 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: (a) 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, (b) participants and nonparticipants reside in the same local labor markets, and (c) the data contain a rich set of variables relevant to modeling the program participation decision. If the comparison group data fails to satisfy these criteria, the performance of the estimators diminishes greatly. Based on this evidence, HIT and HIST 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 and the data were collected using a combination of different survey instruments and administrative data sources.

More recently, DW (1998,1999) use the NSW data (also used by Lalonde) 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 suc-

[5]These tests have also been applied in an evaluation context by Ashenfelter (1978), Bassi (1984), LaLonde (1986), Friedlander and Robins (1995), Regnér (2001) and Raaum and Torp (2001).

ceed 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 (1997) and HIST (1998). They interpret their findings as evidence that matching-on-observables approaches are generally more reliable than the econometric estimators that Lalonde used, some of which were designed to control for biases arising from selection on observables and unobservables.

In this paper, we use the same NSW data to evaluate the performance of both traditional, pairwise matching methods and of the newer methods developed in HIT (1997, 1998) and HIST (1998). We find that a major difference between the DW (1998, 1999) studies and the LaLonde (1986) study is that DW exclude about 40 percent of Lalonde's observations 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 and has the effect of eliminating many of the higher earners in Lalonde's original sample, which makes the selection problem easier to solve. In fact, almost any conventional evaluation estimator applied to the DW samples exhibits low bias. Matching estimators perform much less well when applied to the full data sample that Lalonde (1986) used. Their performance is also highly sensitive to the choice of variables included in the propensity score model.

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. This missing data problem – that the researcher seeking to evaluate the impact of a program only observes one of the two potential outcomes for each person – 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]

$$TT = E(\Delta|X, D = 1) = E(Y_1 - Y_0|X, D = 1) = E(Y_1|X, D = 1) - E(Y_0|X, D = 1). \quad (1)$$

This parameter estimates the average impact among those participating in the program. It is the parameter on which LaLonde (1986) and DW (1998,1999) focus and is a central parameter in many

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

evaluations.[7] When Y represents earnings, a comparison of the mean impact of treated on the treated with the average per-person cost of the program indicates whether or not the program's benefits outweigh its costs, which is of a key question of interest in many evaluations.

Data on program participants identifies the mean outcome in the treated state, $E(Y_1|X, D = 1)$. In a social experiment, where persons who would otherwise participate are randomly denied access to the program, the randomized-out control group provides a direct estimate of $E(Y_0|X, D = 1)$. However, in nonexperimental (or observational) studies, no direct estimate of this counterfactual mean is available. In the next section, we discuss common approaches for estimating the missing counterfactual mean.

3.2 Three Commonly-Used Nonexperimental Estimators

Nonexperimental estimators use two types of data to impute counterfactual outcomes for program participants: (1) data on participants prior to entering the program and (2) data on nonparticipants. Three common evaluation estimators are the *before-after*, *cross-section* and *difference-in-difference* 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 $E(U_{1it}) = 0$ and $E(U_{0it}) = 0$. The observed outcome is $Y_{it} = D_i Y_{1it} + (1 - D_i)Y_{0it}$, which can be written as

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

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 both 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 that is 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 is constant across individuals (i.e. the comment effect assumption $\varphi_1(X_{it}) = \varphi_0(X_{it}) + \alpha^*$). 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 ($\hat{\alpha}_{BA}$) to α^* in

$$Y_{it} - Y_{it'} = \varphi_0(X_{it}) - \varphi_0(X_{it'}) + \alpha^* + U_{it} - U_{it'}.$$

[7] See Heckman, Smith and Clements (1997), Heckman, LaLonde and Smith (1999) and Heckman and Vytlacil (2000) for discussions of other parameters of interest.

For $\hat{\alpha}_{BA}$ to be a consistent estimator, we require that $E(U_{it} - U_{it'}) = 0$ and $E((U_{it} - U_{it'})(\varphi(X_{it}) - \varphi(X_{it'}))) = 0$. A special case where this assumption would be satisfied is if $U_{it} = f_i + v_{it}$, 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] Estimates can also be sensitive to the choice of base time period due to the commonly observed pattern that the mean earnings of program participants decline during the period just prior to participation (see the discussions of the so–called "Ashenfelter's Dip" in Ashenfelter, 1978, Heckman and Smith, 1999, and Heckman LaLonde and Smith, 1999).

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

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

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

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 $D = 0$ persons. The difference-in-differences estimator $\hat{\alpha}_D$ corresponds to the least squares solution for α^* in

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

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

3.3 Matching Methods

Traditional matching estimators pair each program participant with an observably similar non-participant and interpret the difference in their outcomes as the effect of the program (see, e.g., Rosenbaum and Rubin, 1983). Matching estimators are often justified under the assumption that program outcomes are independent of program participation conditional on a set of observables. That is, it is assumed that there exists a set of observable conditioning variables Z (which may

[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 over time or from shocks to the economy. In this example, α^* is confounded with $\gamma_t - \gamma_{t'}$.

be a subset or a superset of X) for which the non-participation outcome Y_0 is independent of participation status D conditional on Z,[9]

$$Y_0 \perp\!\!\!\perp D \mid Z \ . \tag{4}$$

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.,

$$0 < \Pr(D = 1|Z) < 1. \tag{5}$$

This assumption implies that a match can be found for all $D = 1$ persons. If assumptions (**??**) and (**??**) are satisfied, then the Y_0 distribution observed for the matched non-participant group can be substituted for the missing Y_0 distribution for participants.

Assumption (**??**) 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:

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

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 non-participant. The TT parameter requires only the possibility of a non-participant analogue for each participant. For completeness, the required condition is

$$\Pr(D = 1|Z) < 1.^{10} \tag{7}$$

Under these assumptions – either (4) and (5) or (6) and (7) – the mean impact of the program can be written as

$$\begin{aligned} \Delta &= E(Y_1 - Y_0|D = 1) \\ &= E(Y_1|D = 1) - E_{Z|D=1}\{E_Y(Y|D = 1, Z)\} \\ &= E(Y_1|D = 1) - E_{Z|D=1}\{E_Y(Y|D = 0, Z)\}, \end{aligned}$$

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

In a social experiment, (**??**) and (**??**) 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 (1997) and HIST (1996,1998) in their application of matching methods to the JTPA data and of DW (1998, 1999) in their application to the NSW data is that (**??**) 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 $D = 0$ 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 P-values lie within the overlapping support region.

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

[11] An advantage of experiments noted by Heckman (1997), as well as HIT (1997) and HIST (1998), is that they guarantee that the supports are equal across treatments and controls, so that the mean impact of the program can be estimated over the entire support.

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 the matching method. They show that for random variables Y and Z and a discrete random variable D

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

so that $E(D|\ Y, Z) = E(D|Z) = \Pr(D = 1|Z)$ implies $E(D|Y, \Pr(D = 1|Z)) = E(D|\Pr(D = 1|Z))$. 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 parametrically (or semiparametrically at a rate faster than the nonparametric rate), the dimensionality of the matching problem is reduced by matching on the univariate propensity score. For this reason, much of the recent evaluation literature on matching focuses on propensity score matching methods.[13]

3.3.2 Matching Estimators

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

$$\hat{\alpha}_M = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} [Y_{1i} - \hat{E}(Y_{0i}|D = 1, P_i)] \tag{8}$$

where

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

and where I_1 denotes the set of program participants, I_0 the set of non-participants, S_P the region of common support (see below for ways of constructing this set), n_1 denotes 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 non-participants, where the weights $W(i, j)$ depend on the distance between P_i and P_j.

Define a neighborhood $C(P_i)$ for each i in the participant sample. Neighbors for i are non-participants $j \in I_0$ 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.

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

[13] HIT (1998) and Hahn (1998) consider whether it is better in terms of efficiency to match on $P(X)$ or on X 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.

Nearest Neighbor matching Traditional, pairwise matching, also called *nearest-neighbor matching*, sets

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

That is, the non-participant 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. Also, in traditional applications of this estimator it was common not to impose any common support condition. We implement this method in our empirical work using both single nearest neighbor and ten nearest neighbors. When multiple neighbors are used, each receives equal weight in constructing the counterfactual mean. 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).

Caliper matching *Caliper matching* (Cochran and Rubin, 1973) is a variation 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 (1998) employ a variant of caliper matching that they call "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.[14]

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

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

in HIT (1997, 1998) and HIST (1998), which is given by

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

where $G(\cdot)$ is a kernel function and a_n is a bandwidth parameter. In terms of equation (8), the weighting function, $W(i,j)$, is equal to $\frac{G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n}\right)}$. For a kernel function bounded between -1 and 1, the neighborhood is $C(P_i) = \{|\frac{P_i - P_j}{a_n}| \leq 1\}$, $j \in I_0$. Under standard conditions on the bandwidth and kernel , $\frac{\sum_{j \in I_0} Y_{0j} G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n}\right)}$ is a consistent estimator of $E(Y_0|D=1, P_i)$.[15]

In this paper, we implement a generalized version of kernel matching, called local linear matching. Recent research by Fan (1992a,b) has demonstrated advantages of local linear estimation over more standard kernel estimation methods.[16] The local linear weighting function is given by

$$W(i,j) = \frac{G_{ij} \sum\limits_{k \in I_0} G_{ik}(P_k - P_i)^2 - [G_{ij}(P_j - P_i)][\sum\limits_{k \in I_0} G_{ik}(P_k - P_i)]}{\sum\limits_{j \in I_0} G_{ij} \sum\limits_{k \in I_0} G_{ij}(P_k - P_i)^2 - \left(\sum\limits_{k \in I_0} G_{ik}(P_k - P_i)\right)^2}. \tag{9}$$

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 evaluation. The weights depend on the distance (as adjusted by the kernel) 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.

To implement the matching estimator given by equation (8), the region of common support S_P needs to be determined. To determine the support region, we use only those values of P that have positive density within both the $D=1$ and $D=0$ distributions. The common support region can be estimated by

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

where $\hat{f}(P|D=d)$, $d \in \{0,1\}$ are nonparametric density estimators given by

$$\hat{f}(P|D=d) = \sum_{k \in I_d} G\left(\frac{P_k - P}{a_n}\right),$$

[15] We require that $G(\cdot)$ integrates to one, has mean zero and that $a_n \to 0$ as $n \to \infty$ and $na_n \to \infty$.

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

and where a_n is a bandwidth parameter.[17] To ensure that the densities are strictly greater than zero, we require that the densities be strictly positive density exceed zero by a certain 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 percentage 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 level that satisfies:

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

where J is the number of observed values of P that lie in $I_1 \cap \hat{S}_P$. That is, matches are constructed only for the program participants for which the propensity scores lie in $\hat{S}_q$.

HIST (1998) and HIT (1997) also implement a variation of local linear matching which they call "regression-adjusted matching." In this variation, the residual from a regression of Y_{0j} on a vector of exogenous covariates replaces Y_{0j} as the dependent variable in the matching.(For a detailed discussion see HIST (1998) and HIT (1998)). 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 (without regression adjustment) in Sections 7 and 8 below.

Difference-in-difference 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, for a variety of reasons there may be systematic differences between participant and nonparticipant outcomes, even after conditioning on observables, that could lead to a violation of the identification conditions required for matching. Such differences may arise, for example, (a) because of program selectivity on unmeasured characteristics, (b) because of levels differences in earnings across different labor markets in which the participants and nonparticipants reside, or (c) because earnings outcomes for participants and nonparticipants are measured in different ways (as when data are collected using different survey instruments).

A difference-in-differences (DID) matching strategy, as defined in HIT (1997) and HIST (1998), 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

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

[17] In implementation, we select the bandwidth parameter using Silverman's (1986) so-called rule-of-thumb method.

where t and t' are time periods after and before the program enrollment date. This estimator also requires the support condition given in (7), which must hold in both periods t and t' (a non-trivial assumption given the attrition present in many panel data sets). The local linear difference-in-difference estimator is given by

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

where the weights can correspond to either the kernel or the local linear weights defined above. If repeated cross-section data are available, instead of longitudinal data, the estimator can be implemented as

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

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

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.[18] 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-based sampled and the sampling weights are unknown, so we match on the odds ratio, $P/(1-P)$.[19]

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 (**??**) and (**??**) are satisfied. Suppose only a subset

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

[19] With 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 (1998, 1999) studies. However, for methods that take account of the absolute distance between observations, such as kernel matching or local linear matching, it does matter.

$Z_0 \subset Z$ of the variables required for matching is observed. The propensity score matching estimator based on Z_0 then converges to

$$\alpha'_M = E_{P(Z_0)|D=1} \left(E(Y_1|P(Z_0), D=1) - E(Y_0|P(Z_0), D=0) \right). \tag{8}$$

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

$$bias_M = E(Y_0|D=1) - E_{P(Z_0)|D=1}\{E(Y_0|P(Z_0), D=0)\}.$$

HIST (1998) show that what variables are included in the propensity score matters in practice for the estimated bias. They found that the lowest bias values were obtained when the Z data included a rich set of variables relevant to modeling the program participation decision. Higher bias values were obtained for a cruder set of Z variables. Similar findings about nonrobustness of matching when cruder conditional variables are used are reported in Lechner (2000) and below in 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, data on a randomized-out control group makes it possible to separate out the bias. First, subject to the caveats discussed in Heckman and Smith (1995) and Heckman, LaLonde and Smith (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.[20] Therefore, the performances of nonexperimental estimators can be evaluated by applying the estimator to data from the randomized-out control group and from the nonexperimental comparison group and then checking whether the resulting estimates yield and estimated impact equal to zero.

4 The National Supported Work Demonstration

The National Supported Work (NSW) Demonstration[21] was a transitional, subsidized work experience program that operated for four years at fifteen locations throughout the United States.[22] It

[20] 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 (1998,1999). 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 (1997) and HIST (1998).

[21] See Hollister, Kemper and Maynard (1984) for a detailed description of the NSW demonstration and Couch (1992) for long-term experimental impact estimates.

[22] The data we use in this paper comes from the sites in Atlanta, Chicago, Hartford, Jersey City, Newark, New York, Oakland, Philadelphia, San Francisco, and Wisconsin.

served four target groups: female long-term AFDC recipients, ex-drug-addicts, ex-offenders, and young school dropouts. The program provided work in a sheltered training environment and assisted in job placement. 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 persons 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 four weeks preceeding the time of selection for the program), and (2) the person must have spent no more than three months on one regular job of at least 20 hours per week during the preceding six 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.

¿From April 1975 to August 1977[23] the NSW program in 10 cities 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. The experimental sample includes 6,616 treatment and control observations for which data were gathered through a retrospective baseline interview and four follow-up interivews. These interviews covered 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.[24]

5 Samples

In this study, we consider three experimental samples and two non-experimental comparison groups. All of the samples are based on the male samples from LaLonde (1986).[25] LaLonde's (1986) experimental sample includes male respondents in the NSW's 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 male NSW participants were almost all minorities (mostly African American), 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.[26] 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

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

[24] 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.

[25] We do not examine LaLonde's (1986) sample of AFDC women as it is no longer available due to data storage problems. We plan to reconstruct it from the original MDRC data files in future work.

[26] 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 (1998,1999).

in months 13 to 24 prior to the 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 (1998,1999). The variable "Zero Earnings in 1974" is an indicator variable equal to one when the "Real Earnings in 1974" variable equals zero.[27] 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 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 (1998,1999), which is about 40% smaller than Lalonde's original sample due to additional restrictions they impose. In order to include two years of pre-program earnings in their model for program participation, DW omit 40% of Lalonde's (1986) original sample for which that information was missing.[28] While DW (1998, 1999) 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. [29] Table 2 illustrates the sample inclusion criteria that we found (partly through trial and error) which correctly accounts for all but one observation in their sample.[30] The table is a cross-tabulation of LaLonde's (1986) sample with month of random assignment as rows and zero earnings in months 13 to 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 to 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 to 24 before random assignment. This rule corresponds fairly closely to the verbal statement in DW (1999), though we are puzzled as to the reasoning behind the second rule. The stated intent is to use "earnings in 1974" as an additional conditioning variable,

[27]This is the variable denoted "U74" in DW (1998,1999); note that it corresponds to non-employment rather than unemployment.

[28]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 discussion in Angrist (1990,1998), Ashenfelter (1978), Ashenfelter and Card (1985) and Card and Sullivan (1988) on this point.

[29]See footnote 5, page 11 of DW (1998) or the discussion at the bottom of the first column of page 1054 of DW (1999) for their descriptions.

[30]Dehejia provided us with both their version of the LaLonde sample and a version of their sample in separate files. However, neither file included 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 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.

but as already noted, earnings in months 13 to 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 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 (1998,1999). It is a proper sub sample of the DW sample that excludes the persons who were randomized after April of 1976, because we find their second rule–to include persons randomized after April of 1976 only if they had zero earnings in months 13 to 24–to be problematic. Our "Early RA" sample consists of persons randomly assigned during January through April of 1976, or equivalently the observations shown in the bottom four rows of Table 2. This sample includes 108 treatment group members and 142 control group members and is a proper subset of the DW sample. 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 to 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 (1998,1999). Both are representative national samples drawn from throughout the United States. This implies that the vast majority of comparison group members, even of 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, better educated (70 percent completed high school), more 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 poses a tough problem for any non-experimental 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.[31] 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.

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.[32] As discussed in Heckman, LaLonde and Smith (1999), such estimators (like estimators consisting of crude matching followed by some other non-experimental estimator) do not have clear economic or econometric justifications. 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.

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 (1998,1999); 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.

The estimated coefficients and associated estimated standard errors for the propensity scores based on the DW (1998) specification appear in Table 3.[33] We estimate six sets of scores, one for each pair of experimental and comparison group samples. We follow DW in including a slightly different set of higher order and interaction terms in the specifications for the CPS and PSID

[31] Following DW (1998,1999), we drop the three persons from LaLonde's sample who are missing data on education.

[32] 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.

[33] DW (1999) use slightly different specifications for both the CPS and PSID comparison groups. Compare the notes to Tables 2 and 3 in DW (1998) with the notes to Table 3 in DW (1999).

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 (1998,1999) appear to use only the treatment group.[34] Second, DW (1998) apparently did not include a constant term in the logistic model which we do include.

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 non-linear 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 qualititative, 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 non-participants.[35] 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 comparison group members are correctly predicted to be non-participants 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.

Figure 1 presents histograms of the log-odds ratio for the DW propensity score model applied to each of the three experimental samples with each of the two comparison groups. These figures allow a graphical assessment of the 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 differ-

[34] 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.

[35] This metric is discussed in Heckman and Smith (1999) and HIST (1998). For caveats, see Lechner and Smith (2000).

ences in observable characteristics reported in Table 1. However, the support problem here is not as strong as in the JTPA data examined in HIST (1996,1998), 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 a handful of comparison group observations.

Table 4 presents the coefficient estimates from the participation model in LaLonde (1986).[36] 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.[37] 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 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, Figure 1 shows that the LaLonde scores for all three experimental

[36]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 MSA 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).

[37]The latter variable is a bit of an odd choice for inclusion, 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.

samples, like the DW scores, are spread out over the full range between zero and one, but are quite thin among non-participants at the higher scores.

7 Variable Selection and the Balancing Test

An important consideration in implementing propensity score matching is how to choose which variables to include in estimating the propensity score. HIST (1998), HIT (1999) and Lechner (2000) show that which variables are included in the estimation of the propensity score can make a substantial difference to the performance of the estimator. In practice, these papers found that biases tended to be higher when cruder sets of conditioning variables where used, but theory does not provide any guidance as to how to choose the set Z. The set Z that satisfies the matching conditions is not necessarily the one the most inclusive one, as augmenting a set that satisfies the identification conditions for matching could lead to a violation of the conditions. Also, using more conditioning variables could exacerbate a common support problem, which is another consideration.

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 variables Z. The theorem states that

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

or equivalently

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

The basic intuition is that after conditioning on $\Pr(D = 1|Z)$, additional conditioning on Z should not provide any 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 (given in equation (4)). 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 $D = 0$ groups after conditioning on $P(Z)$. The test has been implemented in the literature a number of ways. Eichler and Lechner (2001) 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 (1998,1999) 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. Within each stratum, t-tests are used to detect 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 are added to the logistic model and the testing procedure is repeated, until such differences no longer emerge.

In this paper, we implement the balancing test for each combination of experimental and comparison group sample. In each case, we use control and comparison group members in the common support who get matched in the course of calculating our nearest neighbor matching estimates. We next break each pair of matched samples into quintiles or deciles based on the estimated values of $P(Z)$ in the $D = 1$ group. Then we do a Hotelling T^2 test for differences in the means of the Z used to estimate the scores (including any interaction or higher order terms) within each interval.[38] Our focus on just the first moments of the Z variables follows DW (1998), who report little difference in results from also testing higher order terms and interactions. We then report the number of intervals (out of five or ten) where the F-test rejects the null as well as provide information about the strength of the rejections, if any. (These tests will appear in the next version of the paper).

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(1998, 1999) in that we obtain direct estimates of the bias by applying matching to the randomized-out control group and nonexperimental group, whereas the other papers obtain the bias indirectly by applying matching to the treatment and comparison groups and comparing the experimental and the nonexperimental estimates. Second, 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 (1998,1999) that discards treatment group observations with estimated propensity scores that lie below the minimum or above the maximum of the estimated scores in the experimental sample.[39] The main advantage of this approach is ease of implementation. While somewhat more difficult to implement, our 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 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.

[38] DW (1998) suggest but do not implement an F-test in this capacity.

[39] See page 12 of DW (1998) and the first column of page 1058 in DW (1999).

8.1 Cross-Sectional Matching Estimates

Estimates of the bias associated with cross-sectional matching on the propensity score appear in Tables 5A and 5B. We first consider Table 5A, which shows the estimates for the CPS comparison group. The outcome variable throughout both Tables 5A and 5B is earnings in calendar year 1978, where January 1978 is at least five months after random assignment for all of the controls. The first column of Table 5A 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.

The second through the fifth columns in Tables 5A and 5B 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 (1998, 1999). 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 LaLonde and DW samples, but does result in a substantial bias 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 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 (1998,1999) using simple nearest neighbor matching on their sample are highly sensitive both to changes in the sample composition and to changes in the variables included in the propensity score model.

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 Table 5A, 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 5A 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 (1998,1999) are sensitive on both dimensions: the experimental sample employed and the variables used to estimate the propensity score. The one additional finding is that, consistent with HIT (1997), the matching estimates do not show much sensitivity, at least in terms of the qualitative conclusion they provide, to either the matching method used or to the bandwidth used within the subclass of local linear matching estimators.

Table 5B presents estimates analogous to those in Table 5A 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 initial raw mean differences, the bias estimates in Table 5B are comparable to those in Table 5A. Overall, the performance of the cross-sectional matching estimators is similar to that found in HIT (1997) and HIST (1998). These estimators reduce the bias relative to an unadjusted comparison of means, but the bias that remains after matching is typically of the same order of magnitude as the experimental impact. 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 1975 who had zero earnings in months 13 to 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 different data subsamples.

8.2 Difference-in-Differences Matching Estimates

Tables 6A and 6B present difference-in-differences matching estimates for the CPS and PSID comparison groups, respectively, which were not considered in earlier work. 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. 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 assignmnet bias. In constructing the difference-in-differences matching estimates presented in Tables 6A and 6B, we use the same matching methods used in Tables 5A and 5B.

Consider Table 6A and the CPS comparison group first. Four major patterns emerge. First, all of the difference-in-dfferences 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 Table 6A. Third, as in Table 5A 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. Results are not robust to perturbations in the sample or in the propensity score model, mirroring the findings for the cross-sectional matching estimator.

The results with the PSID comparison group, presented in Table 6B, 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.

While the cross-sectional matching estimates presented in Tables 5A and 5B reveal the extreme sensitivity of the results in DW (1998,1999), the estimates in Tables 6A and 6B show fairly stable performance for the difference-in-differences matching estimators. These results differ from the findings in HIT (1997) and HIST (1998) 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 non-experimental comparison groups, which were not sources of bias with the JTPA 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 to 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 functional form and does not require $E(U|X, D) = 0$.[40]

Tables 7A and 7B give the bias estimates for the CPS and PSID comparison group samples, respectively. In each table, each pair of rows contains the bias and bias-as-a-percentage-of-the-impact estimates 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 regression specifications containing varying sets of covariates, 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. The final four columns in Tables 7A and 7B show bias estimates from the difference-in-differences estimator and unrestricted difference-in-differences estimator examined in Table 5 of LaLonde (1986). The difference between the two pairs of estimators is that in the first two, 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 are included as a right-hand-side variable. The latter formulation relaxes the restriction implicit in the former that the coefficient on 1975 earnings equal -1.[41]

The estimates in Tables 7A and 7B 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 (1998,1999) results in a sample with a selection problem that is less difficult to solve.[42] The exception to this rule are the two 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.[43]

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 (1998,1999) is picking the right subsample and the right covariates, not matching. In contrast, in the PSID, matching makes a big difference. The bias estimate with tenth nearest neighbor matching (imposing common support)

[40] Of course, with a sufficient number of interaction and higher order terms, this difference goes away.

[41] 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.

[42] 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 using the "Real Earnings in 1974" variable among the covariates using the CPS-1 and PSID-1 comparison groups.

[43] 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.

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.

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 Tables 5A and 5B. 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 sucessfully corrects for selection bias.[44]

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 Tables 6A and 6B. [45]

Tables 8A and 8B present the pre-program estimates for the CPS and PSID comparison groups, respectively. Consider first Table 8A. 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 Table 5A. 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, where the biases are of moderate size, the first two of the eight estimators in Table 5A 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 Table 8B for the PSID comparison group in Table 8B. 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

[44] See Heckman and Hotz (1989) for a more detailed discussion of the test and Heckman, LaLonde and Smith (1999) for a discussion of important caveats regarding its use.

[45] 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.

sample are again a mixed bag, though in this case the four estimators eliminated by 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 also employed in LaLonde's (1986) influential paper on the performance of non-experimental evaluation methods yields three main conclusions. First, our evidence leads us to question recent claims in the literature (DW (1998, 1999)) about the effectiveness of matching estimators and about their better performance over traditional econometric methods. While we are able to replicate the low bias estimates reported in the DW studies, we conclude that their evidence is not generalizable because the estimators were applied to a sample that was much smaller than Lalonde's original sample and that imposed some peculiar sample inclusion criteria. In particular, the DW sample only includes persons who were randomized at a calendar date late in the experiment if they had zero earnings in the 13-24 months prior to randomization. Differentially including these zero earners makes a huge difference to the bias estimates and has the effect of making the selection problem easier to solve. Indeed, 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. When we apply matching methods to the more inclusive Lalonde sample or to a subsample of the DW sample that does not impose the problematic zero-earner restriction, we obtain very different results that indicate large biases for cross-sectional matching procedures. The estimates also tend to be highly sensitive to changes in the model for the propensity scores.

Second, we find that the difference-in-differences matching estimators introduced in HIT (1997) and HIST (1998) perform substantially better than the corresponding cross-sectional matching estimators and are more generally robust to perturbations in the analysis sample and in the propensity score model. This 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. Third, we find that the details of the matching procedure, such as which particular form of matching is used and what bandwidth is selected for local-linear matching, do not have a strong or consistent effect on the estimated biases.

The implications of our findings for evaluation research are clear. Matching is not a magic bullet that will solve all evaluation problems. When the estimators are applied to high quality data, matching methods have been found to perform well in coming close to replicating the results of experiments.(See e.g. HIT, 1997). However, the CPS and PSID comparison groups in the NSW evaluation suffer from the problems such as geographic mismatch and from variables being measured in different ways across different survey instruments. It is perhaps not surprising that

matching methods do not perform well in eliminating these various sources of biases, a purpose for which they were not designed. Among the methods considered in this paper, difference-in-difference matching estimators come the closest to providing a reliable evaluation strategy, although even with this estimator the magnitude of the bias sometimes exceeds the magnitude of the experimental impact.

References

[1] Angrist, Joshua (1990): "Lifetime Earnings and the Vietnam Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80(3), 313-335.

[2] Angrist, Joshua (1998): "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica*, 66(2), 249-288.

[3] Ashenfelter, Orley (1978): "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47-57.

[4] Ashenfelter, Orley and David Card (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 67, 648-660.

[5] Barnow, B., G. Cain, and A. Goldberger (1980): "Issues in the Analysis of Selectivity Bias," in Ernst Stromsdorfer and George Farkas, eds., *Evaluation Studies Review Annual Volume 5* (San Fransisco: Sage), 290-317.

[6] Bassi, Lauri (1984): "Estimating the Effects of Training Programs with Nonrandom Selection, " *Review of Economics and Statistics*, 66, 36-43.

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

[8] Burtless, Gary (1995): "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, 9(2), 63-84.

[9] Burtless, Gary and Larry Orr (1986): "Are Classical Experiments Needed for Manpower Policy?," *Journal of Human Resources*, 21, 606-639.

[10] Card, David and Daniel Sullivan (1988): "Measuring the Effect of Subsidized Training Programs on Movements in and out of Employment," *Econometrica*, 56(3), 497-530.

[11] Cochran, W. and Donald Rubin (1973): "Controlling Bias in Observational Studies," *Sankyha*, 35, 417-446.

[12] Couch, Kenneth (1992): "New Evidence on the Long-Term Effects of Employment and Training Programs," *Journal of Labor Economics*, 10(4), 380-388.

[13] Dehejia, Rajeev and Sadek Wahba (1998): "Propensity Score Matching Methods for Nonexperimental Causal Studies," NBER Working Paper No. 6829.

[14] Dehejia, Rajeev and Sadek Wahba (1999): "Causal Effects in Noexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94(448), 1053-1062.

[15] Devine, Terry and James J. Heckman (1996): "The Structure and Consequences of Eligibility Rules for a Social Program: A Study of the Job Training Partnership Act (JTPA)", *Research in Labor Economics*, *Volume. 15*, ed. by S. Polachek. Greenwich, CT: JAI Press, pp. 111-170.

[16] Eichler, Martin and Michael Lechner (2001): "An Evaluation of Public Employment Programmes in the East German State of Sachsen-Anhalt," *Labour Economics*, forthcoming.

[17] Fan, J. (1992a): "Design Adaptive Nonparametric Regression, " *Journal of the American Statistical Association,* 87, 998-1004.

[18] Fan, J. (1992b): : "Local Linear Regression Smoothers and their Minimax Efficiencies," *The Annals of Statistics,* 21, 196-216.

[19] Fraker, Thomas and Rebecca Maynard (1987): "The Adequacy of Comparison Group Designs for Evaluations of Employment Related Programs," *Journal of Human Resources,* 22, 194-227.

[20] Friedlander, Daniel and Philip Robins (1995): "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods," *American Economic Review*, 85(4), 923-937.

[21] Hahn, Jinyong (1998): "On the Role of the Propensity Score in Efficient Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315-331.

[22] Heckman, James (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153-161.

[23] Heckman, James (1992): "Randomization and Social Policy Evaluation," in Charles Manski and Irwin Garfinkle, eds., *Evaluating Welfare and Training Programs* (Cambridge, Mass.: Harvard University Press), 201-230.

[24] Heckman, James (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.

[25] Heckman, James, Neil Hohmann and Jeffrey Smith, with Michael Khoo (2000): "Substitution and Drop Out Bias in Social Experiments: A Study of an Influential Social Experiment," *Quarterly Journal of Economics*, 115(2), 651-694.

[26] Heckman, James and Joseph Hotz (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.

[27] Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1996): "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evdience on the Effectiveness of Matching as a Program Evaluation Method," *Proceedings of the National Academy of Sciences*, 93(23), 13416-13420.

[28] Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica* , 66(5), 1017-1098.

[29] Heckman, James, Hidehiko Ichimura and Petra Todd (1997): "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies,* 64(4), 605-654.

[30] Heckman, James, Hidehiko Ichimura and Petra Todd (1998), "Matching As An Econometric Evaluation Estimator," *Review of Economic Studies,* 65(2), 261-294.

[31] Heckman, James, Robert Lalonde and Jeffrey Smith (1999): "The Economics and Econometrics of Active Labor Market Programs" in Orley Ashenfelter and David Card, eds., *Handbook of Labor Economics Volume 3A* (Amsterdam: North-Holland), 1865-2097.

[32] Heckman, James and Richard Robb (1985): "Alternative Methods for Evaluating the Impact of Interventions," in James Heckman and Burton Singer, eds., *Longitudinal Analysis of Labor Market Data (*Cambridge, England: Cambridge University), 156-246.

[33] Heckman, James and Jeffrey Smith (1995): "Assessing the Case the Randomized Social Experiments," *The Journal of Economic Perspectives,* 9, 85-110.

[34] Heckman, James and Jeffrey Smith (1999): "The Pre-Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies," *Economic Journal*, 109(457), 313-348.

[35] Heckman, James and Jeffrey Smith, with Nancy Clements (1997): "Making the Most Out of Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64(4), 487-536.

[36] Heckman, James and Petra Todd (1995): "Adapting Propensity Score Matching and Selection Models to Choice-based Samples," manuscript, University of Chicago.

[37] Heckman, James and Edward Vytlacil (2000): "Causal Parameters, Structural Equations, Treatment Effects and Randomized Evaluations of Social Programs," manuscript, University of Chicago.

[38] Hirano, Keisuke, Imbens, Guido and Geert Ridder (2000): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," manuscript, UCLA.

[39] Hollister, Robinson, Peter Kemper and Rebecca Maynard. 1984. *The National Supported Work Demonstration* (Madison: University of Wisconsin Press).

[40] LaLonde, Robert (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review,* 76, 604-620.

[41] Lechner, Michael (2000):

[42] Lechner, Michael and Jefrey Smith (2000): "Some Exogenous Information Should Not Be Used in Evaluation Studies," manuscript, University of Western Ontario.

[43] Manski, Charles and Steven Lerman (1977): "The Estimation of Choice Probabilities from Choice-Based Samples," *Econometrica*, 45(8), 1977-1988.

[44] Raaum, Oddbjørn and Hege Torp (2001): "Labour Market Training in Norway – Effect on Earnings," *Labour Economics*, forthcoming.

[45] Regnér, Håkan (2001): "A Nonexperimental Evaluation of Training Programs for the Unemployed in Sweden," *Labour Economics*, forthcoming.

[46] Rosenbaum, Paul and Donald Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects, "*Biometrika,* 70,41-55.

[47] Rosenbaum, Paul and Donald Rubin (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *American Statistician,* 39, 33-38.

[48] Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis* (London: Chapman and Hall).

TABLE 1
Descriptive Statistics for Adult Male Experimental and Comparison Group Samples

	NSW Experimental Samples			Comparison Groups	
Variable	**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)	(1.6)	(2.87)	(3.08)
Black	0.80	0.84	0.82	0.07	0.25
	(.40)	(0.37)	(0.38)	(0.26)	(0.43)
Hispanic	0.11	0.09	0.10	0.07	0.03
	(0.31)	(0.28)	(.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	3631	2102	3742	14017	19429
in 1974"	(6221)	(5364)	(6718)	(9570)	(13407)
Real Earnings	3043	1377	2415	13651	19063
in 1975	(5066)	(3151)	(3894)	(9270)	(13597)
Real Earnings	5455	5301	5796	14847	21554
in 1978	(6253)	(6632)	(7582)	(9647)	(15555)
Real Earnings	…	…	…	14730	…
in 1979				(11028)	
"Zero Earnings	0.45	0.73	0.524	0.12	0.09
in 1974"	(0.50)	(0.44)	(0.50)	(0.32)	(0.28)
Zero Earnings	0.40	0.65	0.41	0.11	0.10
in 1975	(0.49)	(0.48)	(0.49)	(0.31)	(0.30)
Experimental	886	1794	2748	…	…
Impact (1978	(488)	(670)	(1005)		
earnings)					
Sample Size	297 Treatments	185 Treatments	108 Treatments	15992	2490
	425 Controls	260 Controls	142 Controls		

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

TABLE 2
Dehejia and Wahba (1998,1999) Sample Composition
Month of Random of Random Assignment and
Earnings 13-24 Months Before Random Assignnment
Number in Cell, Row Percentage and Overall Percentage
Shaded Area Indicates DW Sample

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

TABLE 3
Dehejia and Wahba (1999a) Propensity Score Model
Coefficient Estimates
(Estimated Standard Errors in Parentheses)

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

TABLE 4
LaLonde Propensity Score Model
Coefficient Estimates
(Estimated Standard Errors in Parentheses)

	LaLonde Experimental Sample		DW Experimental Sample		Early RA Experimental Sample	
Variable	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
	(01972)	(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)
Married	-1.4815	-1.9434	-1.5414	-1.9610	-1.4920	-1.8271
	(0.1531)	(0.1804)	(0.1908)	(0.2115)	(0.2355)	(0.2513)
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)

TABLE 5A
Bias Associated with Alternative Cross-Sectional Matching Estimators
Comparison Group: CPS Adult Male Sample
Dependent Variable: Real Earnings in 1978
(bootstrap standard errors shown in parentheses, trimming level used to determine common support is 2%)

Sample and Propensity Score Model	Mean Diff.	1 Nearest Neighbor Without Common Support	10 Nearest Neighbor Estimator without Common Support	1 Nearest Neighbor Estimator with Common Support	10 Nearest Neighbor Estimator with Common Support	Local Linear Matching (bw = 1.0)	Local Linear Matching (bw =4.0)	Local Linear Regression Adjusted Matching[a] (bw =1.0)	Local Linear Regression Adjusted Matching (bw =4.0)
Lalonde Sample with DW Prop. Score Model	-9757 (255)	-555 (596)	-270 (493)	-838 (628)	-1299 (529)	-1380 (437)	-1431 (441)	-1406	-1329
as % of $886 impact	-1101%	-63%	-30%	-95%	-147%	-156%	-162%	-159%	-150%
DW Sample with DW Prop. Score Model	-10291 (306)	407 (698)	-5 (672)	-27 (723)	-261 (593)	-88 (630)	-67 (611)	-96	-127
as % of $1794 impact	-574%	23%	-0.3%	-1.5%	-15%	-5%	-4%	-5%	-7%
Early RA sample with DW Prop. Score Model	-11101 (461)	-7781 (1245)	-3632 (1354)	-5417 (1407)	-2396 (1152)	-3427 (1927)	-2191 (1069)	-3065	-3391
as % of $2748 impact	-404%	-283%	-132%	-197%	-87%	-125%	-80%	-112%	-123%
Lalonde Sample with Lalonde Prop. Score Model	-10227 (296)	-3602 (1459)	-2122 (1299)	-3586 (1407)	-2342 (1165)	-3562 (3969)	-2708 (1174)	-3435	-2362
as % of $886 impact	-1154%	-406%	-240%	405%	264%	402%	306%	388%	-266%

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

TABLE 5B
Bias Associated with Alternative Cross-Sectional Matching Estimators
Comparison Group: PSID Adult Male Sample
Dependent Variable: Real Earnings in 1978
(bootstrap standard errors shown in parentheses, trimming level used to determine common support is 2%)

Sample and Propensity Score Model	Mean Diff	1 Nearest Neighbor Without Common Support	10 Nearest Neighbor Estimator without Common Support	1 Nearest Neighbor Estimator with Common Support	10 Nearest Neighbor Estimator with Common Support	Local Linear Matching (bw = 1.0)	Local Linear Matching (bw = 4.0)	Local Linear Regression Adjusted Matching[a] (bw = 1.0)	Local Linear Regression Adjusted Matching (bw = 4.0)
Lalonde Sample with DW Prop. Score Model	-16676 (264)	-2932 (898)	-2119 (787)	-166 (959)	-898 (813)	-1237 (747)	-1283 (633)	-587	-817
as % of $886 impact	-1882%	-331%	-239%	-19%	-101%	-140%	-145%	-66%	-92%
DW Sample with DW Prop. Score Model	-16999 (330)	361 (924)	-82 (1200)	447 (827)	-85 (1308)	-122 (1362)	143 (633)	693	777
as % of $1794 impact	-947%	20%	-5%	25%	-5%	-7%	8%	39%	43%
Early RA sample with DW Prop. Score Model	-16993 (555)	-6132 (1237)	-3570 (1315)	-5388 (1487)	-3337 (1222)	-1946 (1079)	-3262 (936)	-3065	-3391
as % of $2748 impact	-618%	-223%	-130%	-196%	-121%	-71%	-119%	-112%	-123%
Lalonde Sample with Lalonde Prop. Score Model	-16464 (262)	-3878 (872)	-3054 (1080)	-3838 (872)	-2977 (985)	-3689 (976)	-3522 (964)	-3708	-3512
as % of $886 impact	-1858%	-438%	-345%	-433%	-336%	-416%	-397%	-419%	-396%

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

TABLE 6A
Bias Associated with Alternative Difference-in-Difference Matching Estimators
Comparison Group: CPS Adult Male Sample
Difference Between Real Earnings in 1978 and Real Earnings in 1975
(bootstrap standard errors shown in parentheses, trimming level used to determine common support is 2%)

Sample and Propensity Score Model	Mean Diff	1 Nearest Neighbor Without Common Support	10 Nearest Neighbor Estimator without Common Support	1 Nearest Neighbor Estimator with Common Support	10 Nearest Neighbor Estimator with Common Support	Local Linear Matching (bw =1.0)	Local Linear Matching (bw = 4.0)	Local Linear Regression Adjusted Matching[a] (bw =1.0)	Local Linear Regression Adjusted Matching (bw = 4.0)
Lalonde Sample with DW Prop. Score Model	867	-1527 (563)	-1317 (520)	-929 (554)	-1064 (539)	-1212 (483)	-1271 (472)	-1212	-1271
as % of $886 impact	98%	-172%	-149%	-105%	-120%	-137%	-143%	-137%	-143%
DW Sample with DW Prop. Score Model	2093	45 (781)	-101 (689)	-607 (784)	-417 (681)	-88 (629)	-75 (621)	-88	-75
as % of $1794 impact	117%	3%	-6%	-34%	-23%	-5%	-4%	-5%	-4%
Early RA Sample with DW Prop. Score Model	598 (549)	1398 (1342)	1041 (1166)	1689 (1212)	3200 (1108)	2993 (3152)	2909 (917)	1876	1461
as % of $2748 impact	22%	51%	38%	61%	116%	109%	106%	68%	53%
Lalonde Sample with Lalonde Prop. Score Model	897 (333)	-463 (1290)	1317 (878)	-21 (1092)	1229 (862)	192 (1102)	927 (801)	-145	928
as % of $886 impact	101%	-52%	149%	-2%	138%	22%	105%	-16%	105%

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

TABLE 6B
Bias Associated with Difference-in-Difference Matching Estimators
Comparison Group: PSID Adult Male Sample
Difference Between Real Earnings in 1978 and Real Earnings in 1975
(bootstrap standard errors shown in parentheses, trimming level is 2%)

Sample and Propensity Score Model	Mean Diff	1 Nearest Neighbor Without Common Support	10 Nearest Neighbor Without Common Support	1 Nearest Neighbor Estimator with Common Support	10 Nearest Neighbor Estimator with Common Support	Local Linear Matching (bw =1.0)	Local Linear Matching (bw =4.0)	Local Linear Regression Adjusted Matching[a] (bw =1.0)	Local Linear Regression Adjusted Matching (bw =4.0)
Lalonde Sample with DW Prop. Score Model	-383 (318)	-1644 (1033)	-148 (931)	608 (1070)	-568 (939)	188 (823)	79 (686)	-344	-318
as % of $886 impact	-43%	-186%	-17%	69%	-64%	21%	9%	-39%	-36%
DW Sample with DW Prop. Score Model	797 (362)	537 (1031)	725 (1208)	568 (906)	737 (1366)	286 (1414)	803 (792)	287	803
as % of $1794 impact	44%	30%	40%	32%	41%	16%	45%	16%	45%
Early RA Sample with DW Prop. Score Model	-133 (629)	-46 (1131)	1135 (1266)	316 (1276)	1153 (1273)	2118 (1016)	1018 (993)	207	111
as % of $2748 impact	-5%	-2%	41%	11%	42%	77%	37%	8%	4%
Lalonde Sample with Lalonde Prop. Score Model	-427	-381	263	-364	238	-204	39	-204	39
as % of $886 impact	-48%	-43%	30%	-41%	27%	-23%	4%	-23%	4%

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

TABLE 7A
Bias Associated with Alternative Regression-Based Estimators
Comparison Group: CPS Adult Male Sample
Dependent Variable: Real Earnings in 1978
(estimated standard errors shown in parentheses)

Sample and Propensity Score Model	Mean Diff.	Regression With LaLonde Covariates[a]	Regression With DW Covariates[b]	Regression With DW Covariates Without RE74	Regression With Rich Covariates[c]	Difference-In-Differences	Difference-In-Differences With Age Included	Unrest. Difference-In-Differences[a]	Unrest. Difference-In-Differences With Covariates
LaLonde Sample	-9756	-1616	-1312	-1466	-974	868	-522	-2405	-1906
	(470)	(410)	(388)	(393)	(451)	(379)	(371)	(357)	(388)
as % of $886 impact	-1101%	-182%	-148%	-165%	-110%	98%	-60%	-271%	-215%
DW Sample	-10292	-690	-34	-238	625	2092	802	-1691	-1089
	(600)	(505)	(486)	(489)	(555)	(481)	(470)	(454)	(479)
as % of $1794 impact	-574%	-38%	-2%	-13%	35%	117%	45%	-94%	-61%
Early RA Sample	-10238	-1384	-1132	-1179	-301	1136	-5	-2337	-1723
	(811)	(655)	(620)	(629)	(707)	(649)	(634)	(608)	(625)
as % of $2748 impact	-373%	-50%	-41%	-43%	-11%	41%	-0%	-85%	-63%

a) The "LaLonde Covariates" are the variables from the LaLonde propensity score model.
b) The "DW Covariates" are the variables from the Dehejia and Wahba (1999a) 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).

TABLE 7B
Bias Associated with Alternative Regression-Based Estimators
Comparison Group: PSID Adult Male Sample
Dependent Variable: Real Earnings in 1978
(estimated standard errors shown in parentheses)

Sample and Propensity Score Model	Mean Diff.	Regression With LaLonde Covariates[a]	Regression With DW Covariates[b]	Regression With DW Covariates Without RE74	Regression With Rich Covariates[c]	Difference-In-Differences	Difference-In-Differences With Age Included	Unrest. Difference-In-Differences[a]	Unrest. Difference-In-Differences With Covariates
LaLonde Sample	-16037	-2632	-2540	-2448	-2111	-427	-1836	-3263	-3192
	(668)	(783)	(756)	(751)	(808)	(543)	(573)	(580)	(665)
as % of $886 impact	-1810%	-297%	-287%	-276%	-238%	-48%	-207%	-368%	-360%
DW Sample	-17796	-920	-1285	-1076	-492	797	-497	-2172	-1969
	(846)	(940)	(960)	(920)	(993)	(683)	(704)	(720)	(791)
as % of $1794 impact	-992%	-51%	-72%	-60%	-27%	44%	-28%	-121%	-110%
Early RA Sample	-16945	-1850	-1949	-1720	-820	-159	-1347	-2951	-2824
	(1311)	(1161)	(1072)	(1057)	(1139)	(920)	(929)	(936)	(981)
as % of $2748 impact	-617%	-67%	-71%	-63%	-30%	-6%	-49%	-107%	-103%

a) The "LaLonde Covariates" are the variables from the LaLonde propensity score model.
b) The "DW Covariates" are the variables from the Dehejia and Wahba (1999a) 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).

TABLE 8A
Bias Associated with Alternative Cross-Sectional Matching Estimators
Comparison Group: CPS Adult Male Sample
Dependent Variable: Real Earnings in 1975
(bootstrap standard errors shown in parentheses, trimming level used to determine common support is 2%)

Sample and Propensity Score Model	Mean Diff.	1 Nearest Neighbor Without Common Support	10 Nearest Neighbor Estimator without Common Support	1 Nearest Neighbor Estimator with Common Support	10 Nearest Neighbor Estimator with Common Support	Local Linear Matching (bw = 1.0)	Local Linear Matching (bw =4.0)	Local Linear Regression Adjusted Matching[a] (bw =1.0)	Local Linear Regression Adjusted Matching (bw =4.0)
Lalonde Sample with DW Prop. Score Model	-1064	972 (314)	1047 (258)	91 (399)	-235 (342)	-168 (315)	-160 (333)	-194	-58
as % of $886 impact	-120%	110%	118%	10%	-27%	-19%	-18%	-22%	-7%
DW Sample with DW Prop. Score Model	-12383 (172)	362 (248)	96 (199)	580 (339)	156 (268)	0 (196)	8 (203)	-39	-21
as % of $1794 impact	-690%	20%	5%	33%	9%	0%	0%	-2%	-1%
Early RA Sample with DW Prop. Score Model	-11700 (354)	-9179 (1769)	-4673 (1132)	-7106 (1357)	-5596 (953)	-6420 (3903)	-5100 (939)	-4941	-4852
as % of $2748 impact	-426%	-334%	-170%	-259%	-204%	-234%	-186%	-180%	-177%
Lalonde Sample with Lalonde Prop. Score Model	-11124 (224)	-3139 (1845)	-3439 (1090)	-3565 (1889)	-3571 (1078)	-3754 (4507)	-3635 (1103)	-3628	-2362
as % of $886 impact	-1255%	-354%	-388%	-402%	-403%	-424%	-410%	-409%	-267%

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

TABLE 8B
Bias Associated with Alternative Cross-Sectional Matching Estimators
Comparison Group: PSID Adult Male Sample
Dependent Variable: Real Earnings in 1975
(bootstrap standard errors shown in parentheses, trimming level used to determine common support is 2%)

Sample and Propensity Score Model	Mean Diff	1 Nearest Neighbor Without Common Support	10 Nearest Neighbor Estimator without Common Support	1 Nearest Neighbor Estimator with Common Support	10 Nearest Neighbor Estimator with Common Support	Local Linear Matching (bw = 1.0)	Local Linear Matching (bw = 4.0)	Local Linear Regression Adjusted Matching[a] (bw = 1.0)	Local Linear Regression Adjusted Matching (bw = 4.0)
Lalonde Sample with DW Prop. Score Model	-16293 (238)	-1288 (673)	-1971 (524)	-442 (631)	-1466 (524)	-161 (547)	-1362 (456)	-243	-499
as % of $886 impact	-1839%	-145%	-222%	-50%	-165%	-18%	-154%	-27%	-56%
DW Sample with DW Prop. Score Model	-17796 (194)	-176 (443)	-807 (676)	-121 (304)	-822 (746)	-408 (518)	-660 (435)	406	-26
as % of $1794 impact	-992%	-10%	-45%	-7%	-46%	-23%	-37%	23%	-1%
Early RA sample with DW Prop. Score Model	-16780 (374)	-6086 (771)	-4705 (778)	-5704 (984)	-4490 (770)	-4064 (690)	-4280 (701)	-4941	-4852
as % of $2748 impact	-611%	-221%	-171%	-208%	-163%	-148%	-156%	-180%	-177%
Lalonde Sample with Lalonde Prop. Score Model	-16036 (213)	-3497 (624)	-3317 (712)	-3474 (779)	-3215 (740)	-3485 (597)	-3561 (629)	-3504	-3551
as % of $886 impact	-1810%	-395%	-374%	-392%	-363%	-393%	-402%	-395%	-401%

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

Figure 1a: Distribution of Estimated Log Odds Ratios

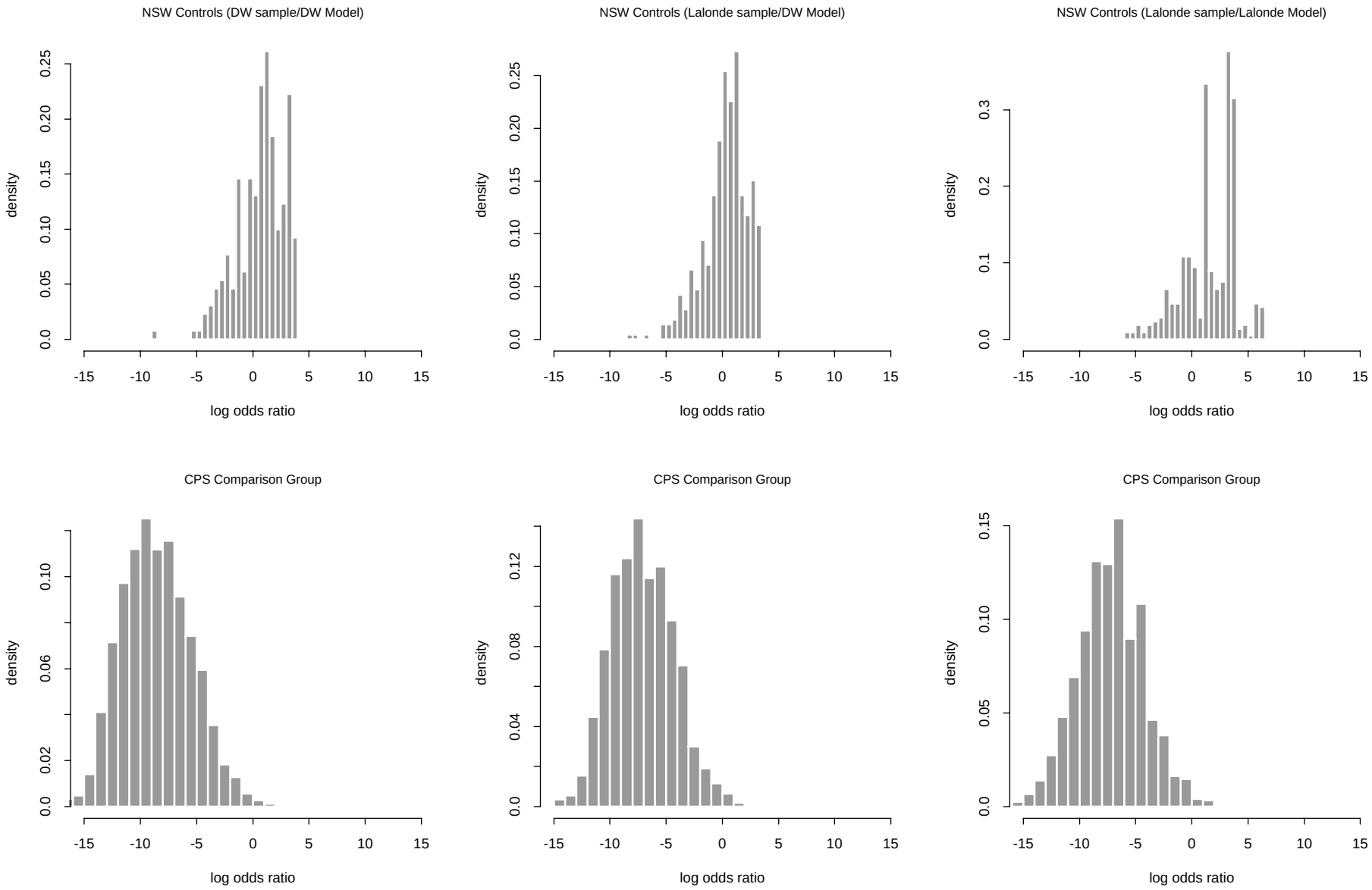

Figure 1b: Distribution of Estimated Log Odds Ratios