ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Zhao, Zhong

Working Paper

Matching estimators and the data from the national supported work demonstration again

IZA Discussion Papers, No. 2375

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Zhao, Zhong (2006) : Matching estimators and the data from the national supported work demonstration again, IZA Discussion Papers, No. 2375, Institute for the Study of Labor (IZA), Bonn,
https://nbn-resolving.de/urn:nbn:de:101:1-20090406224

This Version is available at:
https://hdl.handle.net/10419/33848

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

DISCUSSION PAPER SERIES

IZA DP No. 2375

Matching Estimators and the Data from the National Supported Work Demonstration Again

Zhong Zhao

October 2006

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Matching Estimators and the Data from the National Supported Work Demonstration Again

Zhong Zhao
IZA Bonn

Discussion Paper No. 2375
October 2006

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of the institute. Research disseminated by IZA may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit company supported by Deutsche Post World Net. The center is associated with the University of Bonn and offers a stimulating research environment through its research networks, research support, and visitors and doctoral programs. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 2375
October 2006

ABSTRACT

Matching Estimators and the Data from the National Supported Work Demonstration Again *

We use the data from the National Supported Work Demonstration to study performance of non-propensity-score-matching estimators, and to compare them with propensity score matching. We find that all matching estimators we studied here are sensitive to the choice of data set. Propensity score methods are sensitive to smoothing parameters, and they usually have larger standard error. Difference-in-differences and bias-corrected matching improve the performance of the matching estimators considered here. Our results suggest that the 1974 earnings are important for Dehejia and Wahba's PSID data but not for their CPS data in replicating experiment results. After decomposing the selection bias, we find that a sizable selection bias on unobservables is present in all data sets.

JEL Classification: C14, C21, I38

Keywords: treatment effect, matching estimators, NSW data, selection bias

Corresponding author:

Zhong Zhao
IZA
P.O. Box 7240
D-53072 Bonn
Germany
E-mail: zhao@iza.org

* I am grateful to Carl Christ, Bruce Hamilton, Robert Moffitt, Peter Mueser, Geert Ridder, John Rust, and other seminar participants at Johns Hopkins, Washington University in St. Louis, Manpower Demonstration Research Corporation, Abt Associations, Inc., the 14th Annual Meeting of Midwest Econometrics Group, the 2006 Far Eastern Meeting of the Econometric Society, and 2006 European Economic Association and the Econometric Society European Meetings, for their helpful comments. All errors are mine. I would like to thank Rajeev Dehejia for making the data used in this paper available through his web page.

I. Introduction

Using matching methods to estimate treatment effects under the assumption of selection on observables is attracting much attention from economists (see Imbens, 2004, for an excellent survey, and the references therein). Rosenbaum and Rubin (1983) show that matching on covariates and matching on the propensity score will both balance the distribution of the covariates in the treated group and the comparison group.

The pros and cons of propensity score matching methods are a hotly debated topic in the literature, as in the exchange among Dehejia, Wahba, Smith, and Todd; see Dehejia and Wahba (1999, 2002), Dehejia (2005a, 2005b), and Smith and Todd (2005a, 2005b). Also see the empirical evidence in Mueser, Troske, and Gorislavsky (2005) from administrative data. The consensus that has emerged from this debate is "that propensity score methods are a valuable tool in the research's arsenal and that these methods are not a silver bullet fix to all evaluation problems" (Dehejia, 2005b).

Besides propensity score matching methods, there are other matching estimators available in the literature. For example, Imbens (2004) and Zhao (2004) discuss other matching metrics, including the Mahalanobis metric. Abadie and Imbens (2006) have derived large-sample properties of simple matching estimators based on a Euclidean-type metric. However, there is little empirical evidence on these matching estimators, which do not use propensity scores or the Mahalanobis metric (Imbens 2004).

The main focus of this paper is using the data from the National Supported Work (NSW) Demonstration to present some evidence on the performance of non-propensity-score-matching methods, and to compare them with propensity score methods.

We apply different matching metrics to the NSW data. For each metric, we also consider cross-section matching, difference-in-differences (DID) matching (Heckman, Ichimura, and Todd, 1997; Heckman, Ichimura, Smith, and Todd, 1998), and the bias-correction technique (Rubin, 1973; Abadie and Imbens, 2002). Our study differs from Smith and Todd (2005a), though they use same data set as we do; their paper focuses on propensity score matching methods. Mueser, Troske, and Gorislavsky (2005) study both propensity score matching and Mahalanobis-metric matching, but they use administrative data.

We are interested in studying other matching estimators besides propensity score methods, first, because propensity score matching does not dominate covariate-matching and nonmatching estimators (Hahn, 1998; Angrist and Hahn, 2004; Frölich, 2004; Zhao, 2004). Second, because the large-sample standard error is available for covariate matching (Abadie and Imbens, 2006), but is not available for nearest neighborhood propensity score matching when the propensity score is unknown (Imbens, 2004).[1] The widely used bootstrapping technique is invalid for calculating the standard error for nearest neighborhood matching estimators (Abadie and Imbens, 2005). Third, because matching based on a Euclidean-type metric does not need to specify a model, but matching based on propensity scores needs to specify a function for the propensity score if that is unknown. Balancing tests are usually carried out to select the proper specification for the propensity score. However, what is the best way to do the balancing test is still an unsolved problem (Smith and Todd, 2005b).[2]

[1] Heckman, Ichimura and Todd (1997) give asymptotic standard error for kernel matching.
[2] Even if the propensity score is estimated semi-parametrically or non-parametrically, such as in Kordas and Lehrer (2003), it is still beneficial to carry out the balancing test.

The choice of the NSW data set has been driven by its importance in the evaluation literature since the work of LaLonde (1986), by the fact that there has accumulated considerable knowledge on evaluating the nonexperimental estimators using these data (see Section 3 on this), and by the experimental nature of the NSW data, which enables us to calculate the benchmark treatment effect. However, the NSW data also have some drawbacks, such as small sample size, lack of common support, and high variance of outcome variables.[3]

The rest of the paper is organized as follows. Section II sets up the model using the potential-outcome framework, and discusses cross-section matching, DID matching, matching metrics other than the propensity score, and the bias-correction technique used in the paper. Section III describes the NSW data. Section IV presents matching results from different estimators using different metrics, with and without bias correction. Section V investigates two issues related to why matching estimators perform better in Dehejia and Wahba's data than in LaLonde's data. One issue concerns Ashenfelter's dip, and the other is whether selection on observables is a valid assumption. Section VI concludes the paper.

II. Model Setup and Methodology

1. Model Setup

We set up the model using the standard potential outcome framework as in Rubin (1974). This approach can be traced back to Neyman (1923), Roy (1951), and Quandt (1973). It is also referred as Fisher-Neyman-Roy-Quandt-Rubin model. We assume each

[3] More information on the NSW data can be found in Hollister (1984), LaLonde (1986), Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002), Smith and Todd (2005a).

individual has two potential outcomes (Y_{0i}, Y_{1i}) for a treatment, such as job training, education, or a welfare program. Y_{1i} is the outcome if individual i is treated, and Y_{0i} is the outcome if individual i is not treated. Let $D_i = 1$ indicate that individual i is treated, and $D_i = 0$ indicate the contrary. With (Y_{0i}, Y_{1i}) we can define different treatment effects, such as those in Heckman and Vytlacil (1999, 2005), as follows:

$\Delta_i = Y_{1i} - Y_{0i}$ Treatment effect for individual i ,

$\Delta_{ATE} = E[\Delta_i]$ Average treatment effect for the population (ATE),

$\Delta_S = E[\Delta_i \mid i \in S]$ Average treatment effect for the subpopulation S .

When $S = \{i : D_i = 1\}$, Δ_S is the treatment effect on the treated, denoted as Δ_{TT} . In this paper, we will focus on estimating Δ_{TT} .

2. Cross-Section Matching

That the selection bias is only due to observables is formally characterized by the following two assumptions (Rosenbaum and Rubin, 1983):

M-1: $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$ Conditional-independence assumption,

M-2: $0 < prob(D = 1 \mid X) < 1$ Common-support assumption.

where $\perp\!\!\!\perp$ is the notation for statistical independence as in Dawid (1979). M-1 is also commonly referred as the unconfoundedness assumption or the exogeneity assumption. Under M-1 and M-2 ,

$$\Delta_{TT} = E_{x|D=1}\{E[Y_1 \mid D = 1, X = x] - E[Y_0 \mid D = 1, X = x]\}$$

$$= E_{x|D=1}\{E[Y_1 \mid D = 1, X = x] - E[Y_0 \mid D = 0, X = x]\}.$$

As pointed out by Heckman, Ichimura, and Todd (1997) and Smith and Todd (2005), M-1 can be weakened to H-1 if Δ_{TT} is the parameter of interest:

$$H\text{-}1: E[Y_0 \mid D=0, X] = E[Y_0 \mid D=1, X],$$

$$H\text{-}2: prob(D=1 \mid X) < 1.$$

Using the so-called balancing property

$$prob(X_i \mid T_i = 1, p(X_i) = p) = prob(X_i \mid T_i = 0, p(X_i) = p) = prob(X_i \mid p),$$

Rosenbaum and Rubin (1983) prove that M-1 and M-2 imply

$$P\text{-}1: (Y_0, Y_1) \perp\!\!\!\perp D \mid p(X), \text{ and}$$

$$P\text{-}2: 0 < prob(D=1 \mid p(X)) < 1.$$

It follows from P-1 and P-2 that

$$\Delta_{TT} = E_{p|D=1}\{E[Y_1 \mid D=1, p(X)=p] - E[Y_0 \mid D=1, p(X)=p]\}$$

$$= E_{p|D=1}\{E[Y_1 \mid D=1, p(X)=p] - E[Y_0 \mid D=0, p(X)=p]\}.$$

3. Difference-in-Differences Matching

Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd (1998) extend the cross-section-matching estimators to a longitudinal setting. The following assumptions justify DID covariate matching:

$$DM\text{-}1: Y_{0t} - Y_{0t'} \perp\!\!\!\perp D \mid X \ ,$$

$$DM\text{-}2: 0 < prob(D=1 \mid X) < 1.$$

where the subscript t means the time period after treatment, and t' means before treatment. Under DM-1 and DM-2, we have

$$
\begin{aligned}
\Delta_{TT} &= E_{x|D=1}\{E[Y_{1t} \mid D=1, X=x] - E[Y_{0t} \mid D=1, X=x]\} \\
&= E_{x|D=1}\{E[Y_{1t} \mid D=1, X=x] - E[Y_{0t'} \mid D=1, X=x] \\
&\quad + E[Y_{0t'} \mid D=1, X=x] - E[Y_{0t} \mid D=1, X=x]\} \\
&= E_{x|D=1}\{E[Y_{1t} - Y_{0t'} \mid D=1, X=x] - E[Y_{0t} - Y_{0t'} \mid D=1, X=x]\} \\
&= E_{x|D=1}\{E[Y_{1t} - Y_{0t'} \mid D=1, X=x] - E[Y_{0t} - Y_{0t'} \mid D=0, X=x]\}
\end{aligned}
$$

Δ_{TT} is also identifiable if DM-1 is replaced by the weaker assumption $E[Y_{0t} - Y_{0t'} \mid D=0, X] = E[Y_{0t} - Y_{0t'} \mid D=1, X]$ or $E[Y_{0t} \mid D=0, X] = E[Y_{0t} \mid D=1, X]$.

Following a similar argument to that in Rosenbaum and Rubin (1983), DM-1 and DM-2 imply

$DP\text{-}1: Y_{0t} - Y_{0t'} \perp\!\!\!\perp D \mid p(X)$, and

$DP\text{-}2: 0 < prob(D=1 \mid p(X)) < 1$.

So DID matching can also be done on the propensity score. Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd (1998) use a weaker version of DP-1, $E[Y_{0t} - Y_{0t'} \mid D=0, p(X)] = E[Y_{0t} - Y_{0t'} \mid D=1, p(X)]$, in their paper.

DP-1 can be replaced by $E[Y_{0t} \mid D=0, p(X)] = E[Y_{0t} \mid D=1, p(X)]$ to identify Δ_{TT}.

4. Matching Metrics

Besides the propensity score metric, which uses the absolute difference of propensity score to select observations, we consider five additional metrics here: the standard Euclidean metric d_E, the Mahalanobis metric d_M, the metric d_{AI} used in Abadie and Imbens (2002, 2006), and the metrics d_{Z1} and d_{Z2} in Zhao (2004).

All of these metrics can be written as $(X_{D=1} - X_{D=0})W(X_{D=1} - X_{D=0})'$, where $X_{D=1}$ and $X_{D=0}$ are covariates in the treated group and comparison group, respectively. The only difference is the weighting matrix W. d_E is weighted by an

identity matrix, d_M is weighted by the inverse of the variance-covariance matrix of *X*, d_{AI} is weighted by a diagonal matrix with the inverse of the variances of the *X*'s as its elements, d_{Z1} is weighted by a diagonal matrix with the squares of the coefficients from the estimated propensity score as its elements, and d_{Z2} is weighted by a diagonal matrix with the coefficients from a linear regression of the outcome variable on the covariate as its elements (see Imbens, 2004, and Zhao, 2004 for discussions of these metrics).

In this paper, including the tables, we refer to d_E, d_M, d_{AI}, d_{Z1}, and d_{Z2} as the Euclidean metric, Mahalanobis metric, Abadie-Imbens metric, treatment status metric, and outcome metric, respectively.

5. Bias Correction

When the matching is not exact, the matching results may be biased (Imbens, 2004). Abadie and Imbens (2002) show that when the number of covariates is large, the bias of nearest-neighborhood matching can dominate the variance.

Define the two potential outcome equations and the selection equation as follows:

$Y_{1i} = f_1(X_i) + \varepsilon_{1i}$, where ε_{1i} is iid with $E[\varepsilon_{1i} \mid X_i] = 0$,

$Y_{0i} = f_0(X_i) + \varepsilon_{0i}$, where ε_{0i} is iid with $E[\varepsilon_{0i} \mid X_i] = 0$,

$D_i = I(D_i^* > 0)$, where $I(\cdot)$ is the indicator function, and

$D^* = h(X_i) + \nu_i$, where μ_i is iid with $E[\nu_i \mid X_i] = 0$.

Suppose observation *i* in the treated group matches observation *j* in the comparison group. The idea of matching is to use Y_{0j} to impute the counterfactual of Y_{0i}. The bias from matching for treatment effect on treated is

$$E[Y_{1i} - Y_{0j}] - E[Y_{1i} - Y_{0i}] = E[Y_{0j} - Y_{0i}]$$
$$= E[f_0(X_j) - f_0(X_i)]$$

If the matching is not exact, i.e., if $X_j \neq X_i$, the discrepancy between X_i and X_j may cause bias.

Rubin (1973), Imbens (2004), and Abadie and Imbens (2002) discuss bias-correction methods. The idea is to estimate $f_0(\cdot)$, and to adjust the matching result by $\hat{f}_0(X_j) - \hat{f}_0(X_i)$, where $\hat{f}_0(\cdot)$ is the estimate of $f_0(\cdot)$.[4] In the literature, one usually assumes $f_0(\cdot)$ is a linear function (Imbens, 2004). Rubin (1973) and Imbens (2004) discuss three approaches to adjust the bias. In this paper, we follow Abadie, Drukker, Herr, and Imbens (2004) and only use the matched observations to estimate $f_0(\cdot)$.

As an interesting note, we find that the bias-correction results are identical for non-propensity-score cross-section matching and non-propensity-score DID matching.

6. Discussions on Covariate Matching vs. Propensity Score Matching

Though matching on covariates or on the propensity score can both remove the bias due to observables, if there are many covariates, especially continuous ones, matching on covariates runs into the curse of dimensionality. Since the work of Rosenbaum and Rubin (1983), propensity score matching has dominated the matching literature.

However, it is very possible that individuals with the same propensity score will have very different treatment outcomes. Because of the balancing property, this will not be a problem if the number of observations at each propensity score value *p* is large. This

[4] To estimate the average treatment effect and the treatment effect on the untreated, we also need to estimate $f_1(\cdot)$.

can be easily seen if we compare propensity score matching methods with a randomized experiment. The foundation of a randomized experiment is $prob(X, v \mid treated) = prob(X, v \mid control)$, where X is observable and v is unobservable.

Zhao (2004) notes that the balancing property plays a similar role in propensity score matching, but propensity score matching methods differ from randomization in two important ways. First, propensity score matching only balances the observables. This is why the independence assumption M-1 is needed. Randomization balances both observable and unobservable variables, and equalizes selection bias in treated and control, as pointed out by Heckman (1996).

Second, a randomized experiment balances the distributions for the whole sample, but propensity score matching balances the distributions at each individual propensity score value *p*. In other words, under M-1 and M-2 , the matched sample at each propensity score value *p* is equivalent to a randomized sample.

The estimate from propensity score matching can be thought of as a weighted average of the estimates from many mini "randomized experiments" (at different *p*'s). A substantial sample size is needed to obtain a meaningful estimate from a randomized experiment, and this is translated into a sufficiently large sample size at each *p* for a meaningful propensity score matching estimate.

To facilitate the discussion, we denote by $m(x), m(p)$ the numbers of matched pairs in an *x-cell* with the same covariate *x* and in a *p-cell* with the same propensity score value *p*, respectively. Let r^x be the number of covariate matching cells, and r^p be the number of propensity score matching cells.

When comparing covariate matching with propensity score matching, the advantage of propensity score matching over covariate matching is often characterized by dimensionality reduction, which comprises two aspects. One is that instead of controlling for high-dimensional X, it is enough to control for the propensity score $p(X)$, a scalar. The other is that in general the number of *p*-cells, r^p, is less than the number of *x*-cells, r^x (also see the discussion in Angrist and Hahn, 2004).

Let us consider two polar cases. The first is a randomized experiment. This is the strongest case for propensity score matching. Since $p(X_i)$ is the same for every individual in the randomization, r^p is 1. The advantage of the randomized experiment is the drastic reduction of r^p compared with r^x. A randomized experiment can avoid the empty- or small-cell problem that usually plagues covariate matching when the sample size is small.

The other polar case is that in which the correspondence between $p(X)$ and X is one-to-one. In this case, if exact matching is possible, matching on the propensity score and matching on covariates are equivalent, since in this case people with the same *X* must have the same *p*, and vice versa.

If exact matching is impossible and instead we match on some neighborhood of the propensity score, the story is different. We note the fact that there does *not* exist a one-to-one and bicontinuous (i.e., both the function and its inverse function are continuous) correspondence between R^n space and R^1 space, i.e., R^n space and R^1 space are not homeomorphic. It is natural to assume that $p(X)$ is a continuous function of X. This implies that $p^{-1}(X)$ is *not* a continuous function of p.

The implication of this mathematical fact is shown in Figure 1. On the one hand, if $X's$, say X_1 and X_2, lie in the set *A*, then their $p(X)$'s, viz. p_1 and p_2, must lie in the set *B* [this follows from the continuity of $p(X)$]. On the other hand, there must be always some X's, say X_3 and X_4, that lie outside the set *A*, but whose $p(X)$'s, viz. p_3 and p_4, are in the set *B* [this follows from the discontinuity of $p^{-1}(X)$]. Their corresponding treatment outcomes can be quite different from the ones in the set *A*. Matching by the propensity score on some neighborhood of the propensity score bears the risk of matching p_1 with p_4, whose outcomes, $f(X_1)$ and $f(X_4)$, are quite different even though their propensity scores are similar and the correspondence between X and p is one to one. To average this kind of mismatching out, propensity score matching relies on the balancing property and needs the neighborhood of p to contain a sufficiently large number of observations. In this case, the advantage of matching on covariates is obvious.

The choice between propensity score matching and covariate matching depends on the size of the difference between r^p and r^x is. The combination of r^p, r^x, and $m(p)$ determines the preference between propensity score matching and covariate matching.

Angrist and Hahn (2004) show that when cell sizes are small, when the explanatory power of the covariates to the outcomes is low after controlling for the propensity score, and when the probability of treatment is close to 0 or 1, propensity score matching dominates covariate matching.

Mueser, Troske, and Gorislavsky (2005) compare propensity score matching and Mahalanobis-metric matching using administrative data. Zhao (2004) provides some

Monte Carlo evidence on covariate matching vs. propensity score matching, but there is no empirical study in his paper. In the following sections, we provide empirical evidence on the performance of different matching metrics, and compare them with propensity score matching.

III. The National Supported Work Demonstration Data

The NSW Demonstration is a randomized experiment conducted from 1975 to 1980 to estimate the effects of a "supported" work experience on the disadvantaged population. It has four target groups: women on AFDC, former drug addicts, ex-offenders, and high school dropouts of ages from 17 to 20; see Hollister (1984) for more information. This experiment has 3,214 observations in the treated sample and 3,402 in the control sample.

The NSW data set has played an important role in the treatment effect literature. Lalonde (1986) uses observations of the AFDC group and males from the other three groups in the NSW experiment to evaluate different nonexperimental estimators. He uses the estimate from the NSW data set as the benchmark, and constructs new data sets by combining the NSW treated with the Panel Study of Income Dynamic (PSID) and the Current Population Survey (CPS). LaLonde applies different estimators, e.g. linear regression and difference-in-differences estimator to the constructed data, and finds that these estimators produce very different estimates and often have failed to replicate the benchmark.

Fraker and Maynard (1987) also use the NSW data. They focus on evaluating several approaches for selecting matched comparison samples. They reach similar conclusions to LaLonde's.

As a response, Heckman and Hotz (1989) apply tests to aid the choice among estimators. They point out that different non-experimental estimators impose different assumptions; hence the estimates could be different. The model specification test can be used to choose among different estimators.

They propose three types of tests based on preprogram information on program participants, postprogram experimental information on controls (this is the same exercise as in LaLonde (1986) among others), and overidentified restrictions. Using AFDC women and high school dropouts of the NSW data, they demonstrate that their tests can filter out estimators that produce estimates inconsistent with the experimental benchmark and not reject estimators that produce estimates close to the benchmark.[5]

Using propensity score methods, Dehejia and Wahba (1999, 2002) successfully replicate the benchmark result, but Smith and Todd (2005a) show that their success in doing so has to do with the data selected by them rather than with the propensity score matching method per se, and their results are very sensitive to the covariates included in the scores and to the particular sample used.

The data sets used in LaLonde (1986), Dehejia and Wahba (1999, 2002), and Smith and Todd (2005a) are different.

[5] The sample size in Heckman and Hotz (1989) is larger than the one in LaLonde (1986). There are 566 AFDC women and 800 high school dropouts in the treated groups of the NSW data used by Heckman and Hotz (1989). The outcome variables are 1978 and 1979 earnings.

LaLonde's data includes the AFDC group and males from the other three groups as mentioned above. Dehejia and Wahba's data set is a subset of LaLonde's male sample, and it includes only these males with earnings information in months 13 to 24 before random assignment. For simplicity, we follow Dehejia and Wahba (1999), and refer to this information as 1974 earnings, though in fact it is not. Smith and Todd's data is a subset of Dehejia and Wahba's data set, and their data excludes the observations that were randomized after April 1976; see Smith and Todd (2005a) for detailed discussion of differences among these samples.

Dehejia and Wahba (1999, 2000) also study the LaLonde's data. Smith and Todd (2005a) also present results on the LaLonde's data as well as Dehejia and Wahba's data.

Since we want to examine the effectiveness of different matching methods and do not want other sample selection procedures to contaminate the matching process, our estimation is focused on the whole CPS and PSID samples, i.e., CPS-SSA-1 and PSID-1 in LaLonde (1986), and CPS-1 and PSID-1 in Dehejia and Wahba (1999).

IV. Matching Results[6]

As in the work of Smith and Todd (2005a), we use the propensity score metric, Euclidean metric, Mahalanobis metric, Abadie-Imbens metric, treatment-status metric, and outcome metric to carry out cross-section and DID matching on both LaLonde's data

[6] All matching results, including related standard errors, are estimated using the Stata nnmatch ado file by Abadie, Drukker, Herr, and Imbens (2004) except where otherwise noted. When there are comparison observations with identical propensity score values, the nnmatch routine of Abadie, Drukker, Herr, and Imbens (2004) uses all of them, so numerical results from nnmatch differ slightly from the ones estimated by the psmatch routines of Leuven and Sianesi (2003). Results on propensity score matching using psmatch2 are available from the author upon request.

set and Dehejia and Wahba's data set.[7] The propensity scores are estimated using the treated groups of the NSW along with the comparison groups from CPS or PSID data.

All results are from nearest-neighborhood matching with replacement. Nearest-neighborhood matching assign the same weight to matched comparison observations, while kernel matching (Heckman, Ichimura, and Todd, 1997; Heckman, Ichimura, Smith, and Todd, 1998) weights the matched observations differently according to the selected kernel. Both nearest-neighborhood matching with multiple matches and kernel matching can reduce the standard error.

In the following discussion, we refer the treated group of the LaLonde's NSW sample plus the comparison group from the CPS as LaLonde's CPS data; the treated group of the LaLonde's NSW sample plus the comparison group from the PSID as LaLonde's PSID data; the treated group of the Dehejia and Wahba's NSW sample plus the comparison group from the CPS as the Dehejia and Wahba's CPS data; the treated group of the Dehejia and Wahba's NSW sample plus the comparison group from the PSID as the Dehejia and Wahba's PSID data.

We also refer the LaLonde's CPS and PSID data collectively as the LaLonde's data, and the Dehejia and Wahba's CPS and PSID data collectively as the Dehejia and Wahba's data,

1. Cross-Section Matching Estimates

Panel A in Table 1 to 4 shows results from different nearest-neighborhood (with one, four, and eight matches) cross-section matching estimators.

[7] Smith and Todd (2005a, b) directly estimate bias from control groups and comparison groups. This paper estimate bias indirectly from treated groups, comparison groups and experimental benchmark. Results here are complementary to the ones in Smith and Todd (2005a, b).

Table 1 and Table 2 are results for Dehejia and Wahba's data. The propensity score specifications are the same as the specifications in Dehejia and Wahba (1999).[8] Measured by the closeness to the benchmark (column 4 and 10), the results from different metrics are very similar and there is no evidence that one estimator dominates the other, though the simple OLS has the lowest mean squared error (MSE) in the Dehejia and Wahba's CPS data.

Imposing the common-support condition has little effect on the results. For some estimators it increases the bias, and for others it reduces the bias, but these changes are small. This is not surprising, given that more than 96% of treated observations are in the common support.[9] Unlike subclassification estimators, nearest-neighborhood matching estimators match a treated observation with the nearest comparison observation(s), so the matched comparison observation(s) should not be far away from the common support when almost all the treated are in the common support and the number of matches is small.

The estimates from LaLonde's data are shown in Tables 3 and 4. The propensity score specifications are the same as in Dehejia and Wahba (2005a).

Contrary to the estimates from Dehejia and Wahba's data, all methods except one fail to replicate the NSW experimental benchmark. Most estimates do not even have the correct sign. Smith and Todd (2005a) have the same findings on propensity score matching when they use LaLonde's data.

[8] The specifications in Dehejia and Wahba (1999, 2002) are the same for their CPS data, but are different for their PSID data. The specifications in Dehejia (2005a) differ from the ones in Dehejia and Wahba (1999, 2002). It is worth noting that the correct specification of the propensity score is not unique in theory or in practice.

[9] 99.7%, 96.8%, 97.6%, and 96.8% of treated are in the common support for LaLonde's CPS data, Dehejia and Wahba's CPS data, LaLonde's PSID data, and Dehejia and Wahba's PSID data, respectively.

The MSE's of most estimators are three or four times larger than the MSE from experiment (assuming there is no bias from the experiment).

The estimation from OLS motivated by Barnow, Cain, and Goldberger (1980) is as good as the estimates from propensity score-matching and non-propensity-score-matching methods, if not better. Propensity score matching is sensitive to the smoothing parameter, i.e., the number of matches.

In addition, we use two more criteria to evaluate the performance of different matching estimators.

The first one is whether the estimate falls into the 95% confidence interval of the benchmark. From column 2 and 8 of Table 1 and Table 2, we see almost all estimators perform well according to this criterion. Again, almost all estimators perform poorly in Table 3 and Table 4. Abadie and Imbens (2002) also use this criterion to evaluate the performance of their bias-correction estimator.

The second one is whether the estimator reaches the same conclusion as the benchmark at the 10% significance level, i.e., whether the estimator can detect a significant positive treatment effect at the 10% level.[10]

The majority of the estimators can detect a significant positive effect at the 10% level for Dehejia and Wahba's CPS data (columns 3 and 9 of Table 1), but the majority of them (except outcome metric matching) cannot do so for Dehejia and Wahba's PSID data, even though the experimental benchmark is significant at the 1% level.

For LaLonde's data, none of the matching estimators detects a significant positive effect at the 10% level, and some of them even report a significant negative effect.[11]

[10] Suppose the benchmark is the true treatment effect, and the null hypothesis is the treatment effect is zero, then this criterion is whether the matching estimators make Type II error at 10% significance level.

2. Difference-in-Differences Matching Estimates

Panel A in Table 5 to 8 shows results from DID matching.

First we note that for all data sets, a difference in differences in the 1978 earnings without any adjustment does at least as well as other methods (also see Table 5 in LaLonde, 1986).

As in the findings on DID propensity score matching in Smith and Todd (2005a), DID matching improves the non-propensity-score-matching methods. Comparing columns 3 and 9 of Table 6 and Table 2 clearly shows that DID matching that a majority of the DID estimators that can detect a significant positive effect at the 10% level while cross-section matching cannot.[12]

For LaLonde's data, DID matching cannot replicate the experimental results either.

The advantage of the DID version of propensity score-matching estimators over its cross-section counterpart is not significant. Neither version can detect a significant treatment effect in Dehejia and Wahba's PSID data.

As discussed earlier, the estimate from propensity score matching is the weighted average of the estimates at different propensity score values. The overall quality of the estimation relies on the quality of estimation at each propensity score value. It is interesting to examine more closely the intermediate estimates.

Taking Dehejia and Wahba's CPS data as an example, Figure 2 shows the treatment effect estimated at the pair level. It highlights that people with similar propensity scores can have very different treatment effects.

---

[11] The experimental benchmark is significant at the 6% level in the LaLonde data set.
[12] If measured by the MSE, the picture is not so clear.

We stratify the matched pairs into 18 cells by the propensity score of the treated observation. The width of each cell is 0.05 (since there is no treated observation with propensity score value larger than 0.9, there are 18 cells). Figure 3 shows two estimates of the treatment effects for each cell. One is from the NSW experiment using both the treated and the control observations. The other is from propensity score matching using Dehejia and Wahba's CPS data. There is less volatility than at the pair level, but they are still very noisy. This offers a partial explanation for the high standard error of propensity score matching.

Contrary to the common intuition that the people who have higher propensity score values also have larger treatment effects, it seems that the treatment effect is independent of the propensity score.

3. Bias-Correction Results

Bias-corrected estimates are in Panel B of Table 1 to 8.

There are several results from these tables. First, in general, bias correction, i.e., adjustment of the covariates after matching, can reduce bias. But this is not always the case; for example, in Table 2 there are more incidents of bias increased than of bias decreased after bias correction. Second, the bias-correction technique considered here is more effective in LaLonde's data than in Dehejia and Wahba's data. This might be due to the larger sample size of LaLonde's data. Third, like DID matching, bias-correction improves the estimator's ability to detect a significant effect (Table 2), and reduces the cases of falsely reporting a significant negative effect when the benchmark is significantly positive (Tables 3, 4, and 7). However, after correcting the bias, propensity

score matching still cannot detect a significant positive effect at the 10% level for Dehejia and Wahba's PSID data.

Using Dehejia and Wahba's PSID data, Abadie and Imbens (2002) show matching with bias correction is more robust to the choice of the number of matches.

As shown in Table 5 to 8, the bias-corrected estimates are identical for non-propensity-score cross-section matching and non-propensity-score DID matching.

Overall, we find that the matching results are close to the benchmark in Dehejia and Wahba's data, but estimates using their CPS data tend to underestimate the treatment effect, and estimates using their PSID tend to overestimate it. Given that the average earnings in the CPS are lower than in the PSID (Table 3 of LaLonde, 1986, and Table 1 of Smith and Todd, 2005), if the matched comparison samples were randomly selected from the CPS and PSID, it would be the former that would overestimate the treatment effect. So the matching estimators tend to select lower earners in the PSID than in the CPS.

Another possible explanation for this observation is that the PSID data include an oversampling of poor blacks, and matching estimators may disproportionately select matches from this subsample.

For the different matching approaches considered here, increasing the number of matches usually lowers the standard error, but its effect on bias is ambiguous.

Of the four non-propensity-score metrics in this paper, the standard Euclidean metric performs worst. One major disadvantage of that metric is that it is not unit-free.

DID and bias-correction matching are more effective in Dehejia and Wahba's PSID data, but neither of them can replicate the benchmark in LaLonde's data.[13]

V. Ashenfelter's Dip and the Decomposition of Selection Bias

In general, matching estimators perform better in Dehejia and Wahba's data than in LaLonde's data. Dehejia and Wahba (1999, 2002) and Dehejia (2005a) attribute this to Dehejia and Wahba's data having more information on pretreatment earnings, but Smith and Todd (2005a) argue it is because Dehejia and Wahba's data excludes high earners from the sample, so the matching process is easier.

In this section, we examine two important issues related to this topic: how important is Ashenfelter's dip, and is the matching assumption—selection on observables—valid?

1. Ashenfelter's Dip

Dehejia and Wahba's data only include male observations that have information on the earnings 13 to 24 months prior the randomization (see Smith and Todd, 2005a). It is a nonrandom subsample of LaLonde's data.

The importance of preprogram earning history in the program evaluation has been well known since the discovery of the famous Ashenfelter's dip in Ashenfelter (1978). In order to explore this issue further, we pretend that we do not have the 1974 earning information in Dehejia and Wahba's data set and estimate the treatment effects without using the earning variable of 1974. Dehejia and Wahba (1999) have done a similar analysis using propensity score methods.[14]

[13] DID and bias-correction matching are also more effective in LaLonde's PSID data.
[14] They did not select the specification of propensity score through a balancing test in this analysis.

The propensity score specifications are selected through balancing tests using psmatch2 (Leuven and Sianesi, 2003). Table 9 reports the means of covariates used in the propensity score before and after matching. The means of all covariates are not significantly different between the treated group and the comparison group after matching.

The matching results are reported in Table 10. Comparing Table 10 with Tables 1, 2, 5, and 6, it can be seen that the contribution of the 1974 earning history in improving the estimation of the treatment effects is marginal in Dehejia and Wahba's CPS data. With or without the 1974 earning history, the estimates from Dehejia and Wahba's CPS data set are close to the NSW experimental benchmark, and often fall into the 95% confidence interval of the benchmark.

But the 1974 earning history is important for Dehejia and Wahba's PSID data. Without controlling for this information, though the estimated propensity score has balanced the covariates between treated group and comparison group, propensity score matching fails to replicate the experimental benchmark. Dehejia and Wahba (1999) have made a similar observation.

2. Decomposition of Selection Bias

In order to examine the validity of matching, i.e., whether the selection bias on unobservables is negligible, we apply the approach in Heckman, Ichimura, Smith, and Todd (1998), and decompose the selection bias in LaLonde's data and in Dehejia and Wahba's data.

First, we impose the common-support condition to eliminate selection bias arising from the difference in support between the treated group and the comparison group (see

Heckman, Ichimura, Smith, and Todd, 1998). Then, we further decompose the selection bias within the common support. Instead of decomposing the bias nonparametrically as in Heckman, Ichimura, Smith, and Todd (1998), we adopt a Blinder-Oaxaca-type decomposition approach (Blinder, 1973; Oaxaca, 1973).

Suppose the outcome equations for the treated group and the comparison group are

$$Y_1 = \alpha_1 + X\beta_1 + \varepsilon_1, \qquad D_i = 1 \text{ (treated group)}$$

$$Y_0 = \alpha_0 + X\beta_0 + \varepsilon_0, \qquad D_i = 0 \text{ (comparison group)}$$

The difference between $\overline{\hat{Y}}_1$ and $\overline{\hat{Y}}_0$—the sample means of Y in the treated group and the comparison group, respectively—can be decomposed into three parts: treatment effect on treated, selection bias on observables, and selection bias on unobservables:

$$\begin{aligned} &\overline{\hat{Y}}_1 - \overline{\hat{Y}}_0 \\ &= \hat{\alpha}_1 + \overline{X}_1\hat{\beta}_1 - \hat{\alpha}_0 - \overline{X}_0\hat{\beta}_0 \\ &= \hat{\Delta}_{TT} + \{(\overline{X}_1 - \overline{X}_0)\hat{\beta}_1\} + \{\hat{\alpha}_1 - \hat{\alpha}_0 + \overline{X}_0(\hat{\beta}_1 - \hat{\beta}_0) - \hat{\Delta}_{TT}\} \end{aligned}$$

where $\overline{X}_1$ and $\overline{X}_0$ are the sample means of X in the treated group and the comparison group, respectively. The first term is the treatment effect on the treated, which can be estimated using the NSW experimental data. The second term is the selection bias on observables. The last term is the selection bias on unobservables, and it is estimable given that we know $\hat{\Delta}_{TT}$.

Table 11 summarizes the decomposition results. There are several points worth noting. First, imposing the common-support condition eliminates a large share of the selection bias, from 30% in LaLonde's CPS data to 66% in Dehejia and Wahba's CPS data. Second, selection biases on observables are small compared with other components

of bias, ranging from 8% in Dehejia and Wahba's CPS data to 21% in LaLonde's CPS data. Third, selection biases on unobervables are sizable in all data sets, from 26% in Dehejia and Wahba's CPS data to 49% in LaLonde's CPS data.

Though we have not totally identified the factor(s) behind the different performance of matching estimators between LaLonde's data and Dehejia and Wahba's data, our results suggest that in order to replicate the experimental benchmark, the 1974 earnings are more important for Dehejia and Wahba's PSID data set than for their CPS data set.

After decomposing the selection bias, we find that a sizable selection bias on unobservables is present in all data sets. It is hard to argue that the matching assumption, selection on observables, is valid in any of the samples considered here. Nonetheless, we find that Dehejia and Wahba's CPS data has the smallest selection bias on unobervables; matching estimators perform best in this data set.

VI. Conclusions

In this paper we have studied the performance of non-propensity-score estimators, and compared them with propensity score matching.

Measured by the closeness to the benchmark, results from different estimators are very similar. They are equally good in Dehejia and Wahba's data and are equally bad in LaLonde's data. Our findings are similar to the findings in Smith and Todd (2005a) on propensity score matching methods. There is no evidence that one estimator dominates the other.

If we measure the performance of the estimators by whether they reach the same conclusion as the benchmark does at the 10% significant level, propensity score matching methods perform more poorly than other matching estimators in Dehejia and Wahba's PSID data, and have similar performance to other estimators in other data sets.

DID matching and bias correction can improve the estimation and reduces the standard error. Our results suggest that the bias-correction technique considered here is more effective in LaLonde's data than in Dehejia and Wahba's data. But DID matching and bias-correction matching do not improve the performance of propensity score matching in Dehejia and Wahba's PSID data much, in that both DID and bias-correction versions of propensity score matching fail to detect a significant positive effect at the 10% level in this data set.

We examine two important issues related to why in general matching estimators perform better in Dehejia and Wahba's data than in LaLonde's data: the importance of Ashenfelter's dip, and the validity of the matching assumption. Our results suggest that in order to replicate the experimental benchmark, the 1974 earnings are important for Dehejia and Wahba's PSID data set, but are marginal for their CPS data set.

After controlling for selection bias due to non-overlap of support and for selection on observables, we find that sizable selection bias on unobservables persists in all data sets. It is hard to argue that selection on observables is a valid assumption in any data set studied here. Nonetheless, we find that Dehejia and Wahba's CPS data has the smallest selection bias on unobervables; matching estimators perform best in this data set.

The failure of matching methods in LaLonde's data set highlights the fact that, like that of any nonexperimental estimator, the behavior of matching estimators largely

depends on the data and program. Matching is a useful estimator under suitable conditions, but it is definitely not the estimator for every evaluation. This is the consensus from Heckman, Ichimura, and Todd (1997), Heckman, Ichimura, Smith, and Todd (1998), Smith and Todd (2005a), and Dehejia (2005a, 2005b).

We echo the lesson from Heckman, Ichimura, Smith, and Todd (1998) that there is no easy way in social program evaluation. A successful evaluation study requires detailed knowledge of the program, a thorough understanding of the selection process, a rich data set, and a careful consideration and choice of the estimation strategy.

References

Abadie, Albert, David Drukker, Jane Leber Herr, and Guido W. Imbens (2004), "Implementing Matching Estimators for Average Treatment Effect in Stata," The Stata Journal 4 (3rd Quarter, 2004), 290–311.

Abadie, Albert and Guido W. Imbens (2002), "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," NBER Technical Working Paper T286 (October 2002).

_____ (2005), "On the Failure of the Bootstrap for Matching Estimators," unpublished manuscript.

_____ (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," Econometrica 74 (January 2006), 235–267.

Angrist, Joshua D. and Jinyong Hahn (2004), "When to Control For Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects," Review of Economics and Statistics 86 (February 2004), 58–72.

Ashenfelter, Orley (1978), "Estimating the Effect of Training Programs on Earnings," Review of Economics and Statistics 60 (February 1978), 47–57

Barnow, Burt S., Glen G. Cain, and Arthur S. Goldberger (1980), "Issues in the Analysis of Selection Bias," Evaluation Studies Review Annual 5 (1980), edited by E. Stromsdorfer and G. Farkas.

Blinder, Alan S. (1973), "Wage Discrimination: Reduced Form and Structural Estimates," Journal of Human Resources 8 (Fall 1973), 436–455.

Dawid, A. Philip (1980), "Conditional Independence for Statistical Operations," Annals of Statistics 8 (May 1980), 598–617.

Dehejia, Rajeev (2005a), "Practical Propensity Score Matching: A Reply to Smith and Todd," Journal of Econometrics 125 (March–April 2005), 355–364.

Dehejia, Rajeev (2005b), "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? A Postscript," unpublished manuscript.

Dehejia, Rajeev H. and Sadek Wahba (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," Journal of the American Statistical Association 94 (December 1999), 1053–1062.

_____ (2002), "Propensity Score Matching Methods for Non-Experimental Causal Studies," Review of Economics and Statistics 84 (February 2002), 151–175.

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

Frölich, Markus (2004), "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators," Review of Economics and Statistics 86 (February 2004), 77–90.

Hahn, Jinyong (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," Econometrica 66 (March 1998), 315–331.

Heckman, James J. (1996), "Randomization as an Instrumental Variable," Review of Economics and Statistics 78 (May 1996), 336–341.

Heckman, James J. and Joseph V. 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 (September 1989), 862–874.

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra E. Todd (1998), "Characterizing Selection Bias Using Experimental Dada," Econometrica 66 (September 1998), 1017–1098.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd (1997), "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," Review of Economic Studies 64 (October 1997), 605–654.

Heckman, James J. and Edward Vytlacil (1999), "Local Instrumental Variables and Latent Variables Models for Identifying and Bounding Treatment Effects," Proceedings of the National Academy of Sciences of the USA, 96 (February 1999), 4730–4734.

_____ (2005), "Structural Equations, Treatment Effects and Econometric Policy Evaluation," Econometrica 73 (May 2005), 669–738.

Hollister, Robinson G., Jr. (1984), "The Design and Implementation of the Supported Work Evaluation," Chapter 2 in *The National Supported Work Demonstration*, eds. Robinson G. Hollister, Jr., Peter Kemper, and Rebecca A. Maynard, University of Wisconsin Press, Madison, Wisconsin.

Imbens, Guido W. (2004), "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," Review of Economics and Statistics 86 (February 2004), 4–29.

Kordas, Gregory and Steven F. Lehrer (2003), "Matching using Semiparametric Propensity Scores," unpublished manuscript.

LaLonde, Robert J. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," The American Economic Review 76 (September 1986), 604–620.

Leuven, Edwin and Barbara Sianesi (2003), "PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity Score Matching, Common Support Graphing, and Covariate Imbalance Testing," unpublished manuscript.

Mueser, Peter R., Kenneth R. Troske, and Alexey Gorislavsky (2005), "Using State Administrative Data to Measure Program Performance," unpublished manuscript.

Neyman, Jerzy S., "On the Application of Probability Theory to Agriculture Experiments. Essay on Principles. Section 9.," Statistical Science 5 (1990), 465-485 (Translated from the Polish origin in Roczniki Nauk Rolniczych Tom X, 1923, 1-51).

Oaxaca, Ronald (1973), "Male-Female Wage Differentials in Urban Labor Markets," International Economic Review 14 (October 1973), 693–709.

Quandt, Richard E. "A New Approach to Estimating Switching Regressions," Journal of American Statistical Association 67 (June 1972), 306-310.

Rosenbaum, Paul R. and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika 70 (April 1983), 41–55.

Roy, Andrew D., "Some Thoughts on the Distribution of Earnings," Oxford Economics Paper 3 (1951), 135-146.

Rubin, Donald B. (1973), "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies," Biometrics 29 (March 1973), 185–203.

_____ (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," Journal of Educational Psychology 66 (1974), 688–

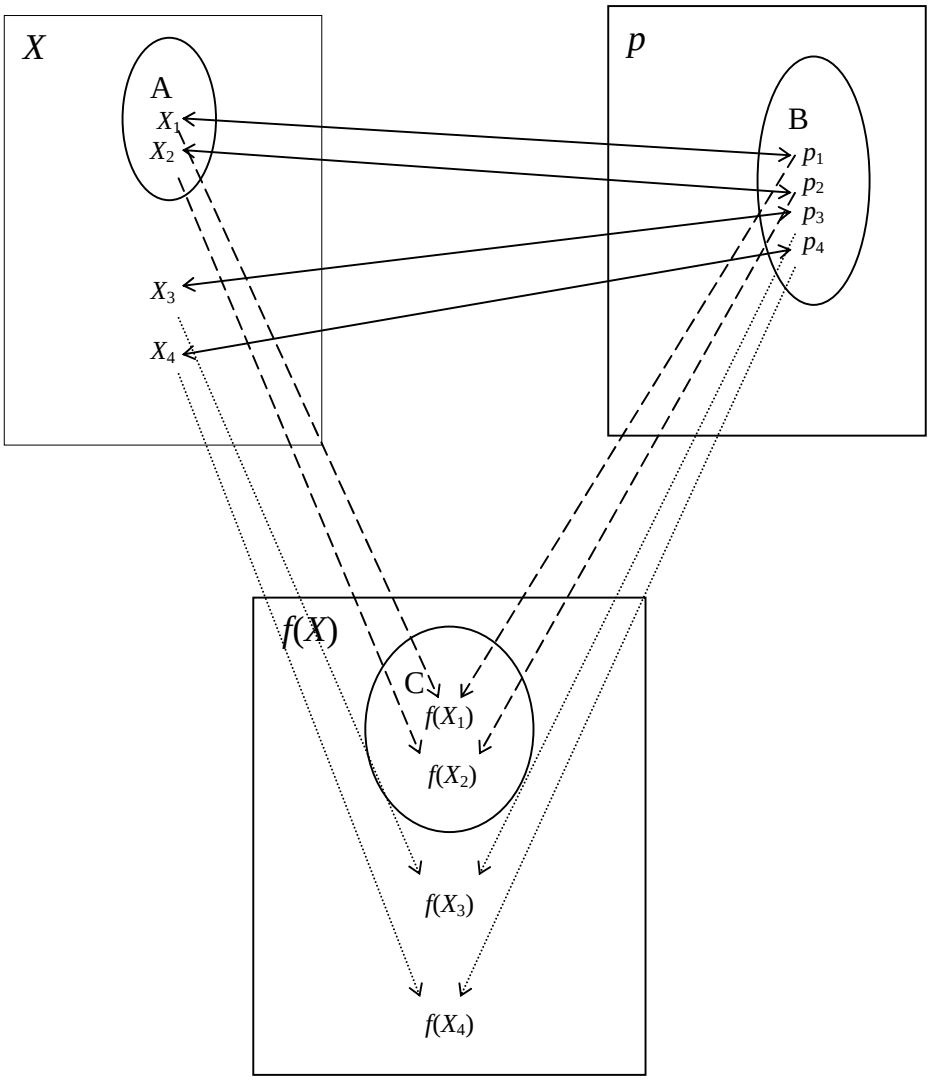
701.

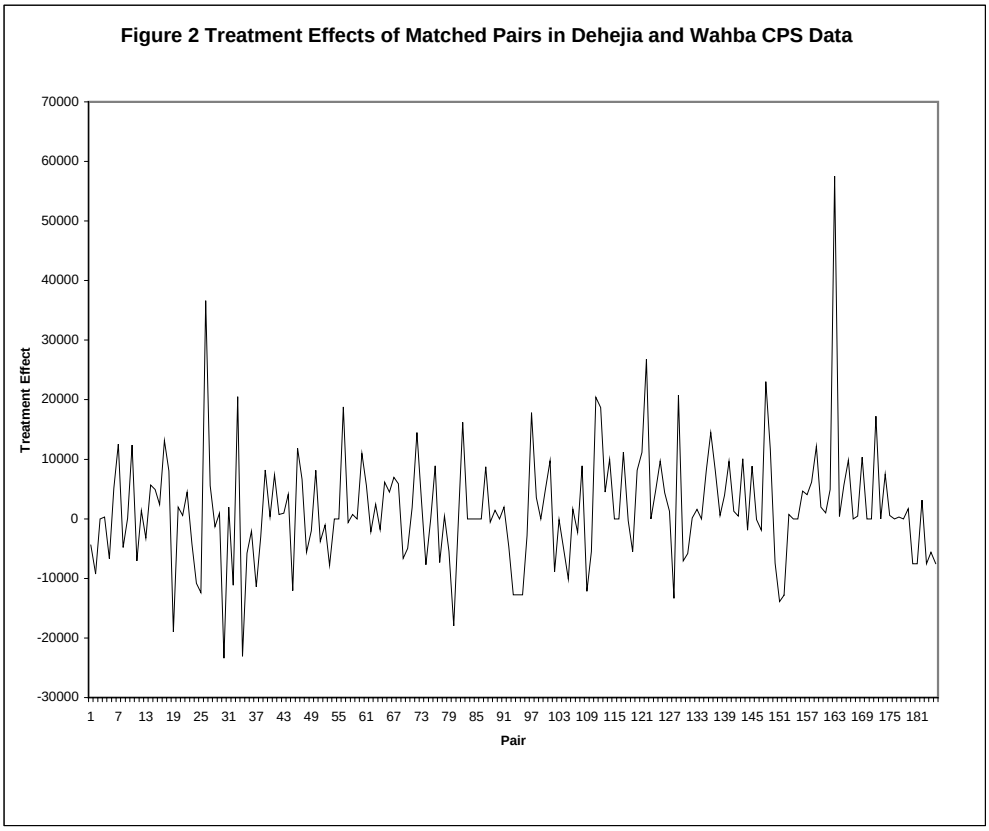
Smith, Jeffrey A. and Petra E. Todd (2005a), “Does Matching Overcome LaLonde’s Critique of Nonexperimental Estimators?” Journal of Econometrics 125 (March–April 2005), 305–353.

_____ (2005b), “Rejoinder,” Journal of Econometrics 125 (March–April 2005), 365–375.

Zhao, Zhong (2004), “Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence,” Review of Economics and Statistics 86 (February 2004), 91–107.

Figure 1 Neighborhood Matching
with One-to-One Correspondence between *X* and *p*

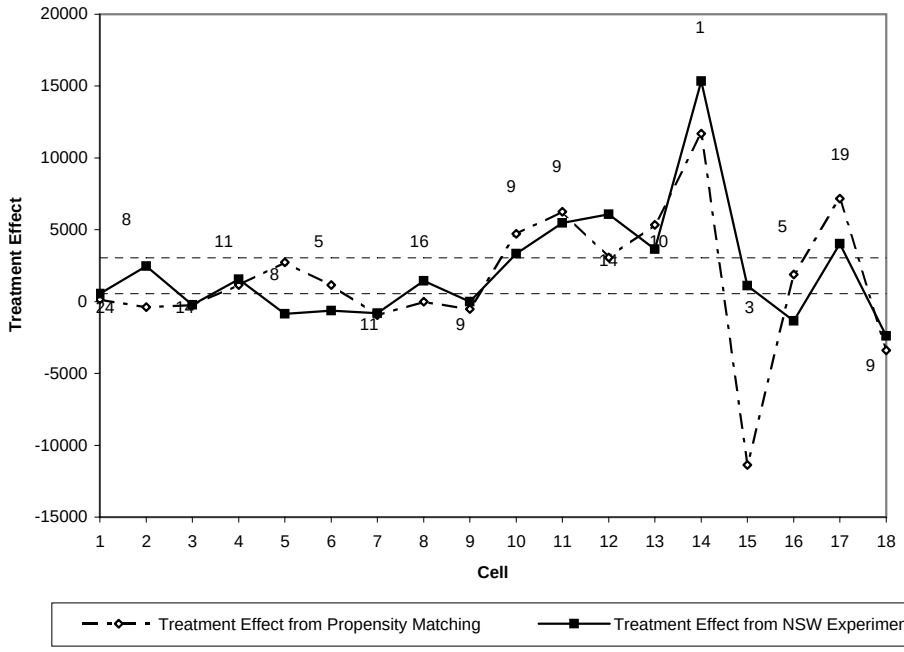




Note: 1. There are total 185 matched pairs.

2. The matched pairs are sorted by the propensity score of the treated.

Figure 3 Treatment Effects of Matched Pairs by Cell in Dehejia and Wahba CPS Data



Note: 1. There are total 185 matched pairs.

2. The matched pairs are sorted by the propensity score of the treated.
3. The width of cell is 0.05. Since there is no observation with propensity score larger than 0.9, there are only 18 cells. The numbers in the plot are the numbers of matched pairs in each cell.
4. Dash lines are 95% confidence interval of experiment benchmark.

Table 1 Estimates from Cross-Section Matching Using Dehejia and Wahba's CPS Data

Panel A: Matching without Bias-Correction											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
NSW Experiment (Benchmark)	1794.34	*	+	0.00	632.85	632.85	1794.34	*	+	0.00	632.85	632.85
Simple Mean Difference	-1739.02		-	-3533.36	606.50	3585.03	-8497.52		-	-10291.86	712.02	10316.46
OLS Regression	1599.42	*	+	-194.92	582.02	613.79	1566.87	*	+	-227.48	556.94	601.60
Propensity Score Matching (1:1)	937.23	*		-857.11	1030.17	1340.10	730.38	*	+	-1063.96	1029.50	1480.50
Propensity Score Matching (1:4)	1719.45	*	+	-74.89	849.13	852.42	1672.93	*	+	-121.42	870.72	879.14
Propensity Score Matching (1:8)	1567.89	*	+	-226.45	803.94	835.23	1513.24	*	+	-281.11	819.80	866.65
Euclidean Metric Matching (1:1)	2026.48	*	+	232.13	825.55	857.56	1313.67	*		-480.67	841.91	969.47
Euclidean Metric Matching (1:4)	1578.14	*	+	-216.20	669.13	703.19	1214.55	*	+	-579.79	659.85	878.38
Euclidean Metric Matching (1:8)	1357.11	*	+	-437.23	625.72	763.35	1078.62	*	+	-715.73	630.08	953.55
Mahalanobis Metric Matching (1:1)	1614.82	*	+	-179.53	989.53	1005.69	1715.67	*	+	-78.67	960.77	963.99
Mahalanobis Metric Matching (1:4)	1511.97	*	+	-282.37	779.08	828.67	1333.56	*	+	-460.79	779.70	905.68
Mahalanobis Metric Matching (1:8)	1184.46	*		-609.88	735.86	955.74	1115.66	*		-678.68	736.73	1001.69
Abadie and Imbens Metric Matching (1:1)	2182.09	*	+	387.75	934.16	1011.43	1723.84	*	+	-70.50	905.55	908.29
Abadie and Imbens Metric Matching (1:4)	1636.66	*	+	-157.69	783.86	799.56	1520.17	*	+	-274.18	785.18	831.67
Abadie and Imbens Metric Matching (1:8)	1398.12	*	+	-396.22	729.66	830.30	1298.87	*	+	-495.48	735.55	886.86
Outcome Metric Matching (1:1)	1819.19	*	+	24.85	971.80	972.12	1623.19	*		-171.16	1019.67	1033.94
Outcome Metric Matching (1:4)	1528.28	*	+	-266.06	792.34	835.82	1363.55	*	+	-430.80	804.03	912.17
Outcome Metric Matching (1:8)	1701.03	*	+	-93.31	719.76	725.78	1567.07	*	+	-227.27	722.96	757.84
Treatment Status Metric Matching (1:1)	1898.96	*	+	104.62	896.48	902.56	1563.19	*	+	-231.15	903.13	932.24
Treatment Status Metric Matching (1:4)	860.67	*		-933.67	731.29	1185.97	618.87	*		-1175.47	712.53	1374.57
Treatment Status Metric Matching (1:8)	596.67	*		-1197.67	670.53	1372.60	313.57			-1480.77	656.68	1619.85
Panel B: Matching with Bias-Correction												
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
Propensity Score Matching (1:1)	932.13	*		-862.21	1030.15	1343.36	728.69	*		-1065.65	1029.32	1481.59
Propensity Score Matching (1:4)	1720.83	*	+	-73.52	849.16	852.34	1681.23	*	+	-113.11	870.68	877.99
Propensity Score Matching (1:8)	1595.45	*	+	-198.89	804.04	828.27	1550.19	*	+	-244.15	819.60	855.19
Euclidean Metric Matching (1:1)	1892.62	*	+	98.28	827.02	832.84	1208.02	*		-586.33	827.91	1014.50
Euclidean Metric Matching (1:4)	1883.99	*	+	89.65	668.05	674.04	1511.52	*	+	-282.82	656.64	714.96
Euclidean Metric Matching (1:8)	1874.09	*	+	79.75	624.13	629.21	1509.86	*	+	-284.48	625.77	687.40
Mahalanobis Metric Matching (1:1)	1654.16	*	+	-140.18	971.94	982.00	1806.98	*	+	12.63	950.35	950.43
Mahalanobis Metric Matching (1:4)	1612.29	*	+	-182.05	773.72	794.85	1467.39	*	+	-326.95	769.30	835.90
Mahalanobis Metric Matching (1:8)	1452.60	*	+	-341.74	731.16	807.08	1349.62	*	+	-444.72	731.10	855.74
Abadie and Imbens Metric Matching (1:1)	2227.29	*	+	432.94	920.98	1017.67	1795.52	*	+	1.17	905.66	905.66
Abadie and Imbens Metric Matching (1:4)	1783.84	*	+	-10.50	779.54	779.61	1734.99	*	+	-59.35	778.82	781.07
Abadie and Imbens Metric Matching (1:8)	1652.51	*	+	-141.83	727.44	741.14	1622.28	*	+	-172.06	730.31	750.31
Outcome Metric Matching (1:1)	1889.73	*	+	95.39	961.25	965.98	1697.62	*	+	-96.72	1006.55	1011.19
Outcome Metric Matching (1:4)	1425.75	*	+	-368.59	785.58	867.75	1258.21	*		-536.13	794.51	958.48
Outcome Metric Matching (1:8)	1449.22	*	+	-345.12	714.93	793.87	1394.54	*	+	-399.80	714.45	818.71
Treatment Status Metric Matching (1:1)	2191.03	*	+	396.68	880.90	966.10	1879.35	*	+	85.01	884.06	888.14
Treatment Status Metric Matching (1:4)	1796.96	*	+	2.62	715.70	715.71	1589.28	*	+	-205.06	693.26	722.95
Treatment Status Metric Matching (1:8)	1843.96	*	+	49.62	656.49	658.37	1755.15	*	+	-39.19	643.90	645.09

Note: 1.The specification of the propensity score is the same as in Table 3 of Dehejia and Wahba (1999), including constant, age, education, no degree, married, black, hispanic, age squared, education squared, re74, re75, age cubed, u74, u75, education*re74.

2. The specification of the OLS is the same as the specification of the propensity score.

3. Standard errors are estimated using nnmatch based on the formula in Abadie and Imbens (2002).

4. The 95% confidence interval from the experiment benchmark is [551,3038]. In column (2) and (8), * indicates the estimate falls into this interval.

5. The benchmark is significant at 1% level. In column (3) and (9), "+"/"-"/blank indicates at 10% level, the estimate positively significant/negatively significant/insignificant.

Table 2 Estimates from Cross-Section Matching Using Dehejia and Wahba's PSID Data

Panel A: Matching without Bias-Correction											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
NSW Experiment (Benchmark)	1794.34	*	+	0.00	632.85	632.85	1794.34	*	+	0.00	632.85	632.85
Simple Mean Difference	-6707.16		-	-8501.50	689.59	8529.42	-15204.78		-	-16999.12	1154.61	17038.29
OLS Regression	26.27			-1768.07	1026.90	2044.65	216.85			-1577.50	1105.87	1926.51
Propensity Score Matching (1:1)	515.90			-1278.44	1900.01	2290.08	552.29	*		-1242.05	1866.97	2242.38
Propensity Score Matching (1:4)	2186.95	*		392.61	1556.17	1604.93	2126.54	*		332.20	1564.49	1599.37
Propensity Score Matching (1:8)	1173.86	*		-620.49	1429.31	1558.18	1068.03	*		-726.32	1452.44	1623.92
Euclidean Metric Matching (1:1)	914.13	*		-880.21	1463.66	1707.94	990.65	*		-803.69	1475.66	1680.32
Euclidean Metric Matching (1:4)	1460.22	*		-334.12	1265.80	1309.16	1485.43	*		-308.91	1254.82	1292.28
Euclidean Metric Matching (1:8)	1512.87	*		-281.47	1184.57	1217.55	1472.61	*		-321.73	1183.71	1226.66
Mahalanobis Metric Matching (1:1)	2375.80	*		581.45	1551.57	1656.94	2050.49	*		256.14	1724.71	1743.63
Mahalanobis Metric Matching (1:4)	1898.70	*		104.36	1303.56	1307.73	1629.94	*		-164.40	1442.21	1451.55
Mahalanobis Metric Matching (1:8)	954.05	*		-840.29	1132.80	1410.43	940.77	*		-853.57	1214.59	1484.53
Abadie and Imbens Metric Matching (1:1)	2453.76	*		659.41	1655.87	1782.33	2037.16	*		242.82	1726.21	1743.20
Abadie and Imbens Metric Matching (1:4)	2206.81	*		412.47	1411.88	1470.90	2040.10	*		245.75	1505.25	1525.18
Abadie and Imbens Metric Matching (1:8)	1791.33	*		-3.01	1168.24	1168.24	1634.21	*		-160.13	1265.23	1275.32
Outcome Metric Matching (1:1)	2915.60	*	+	1121.25	1434.47	1820.69	2886.14	*	+	1091.80	1490.75	1847.80
Outcome Metric Matching (1:4)	2346.68	*	+	552.34	1247.67	1364.47	2204.41	*	+	410.07	1252.49	1317.91
Outcome Metric Matching (1:8)	2077.97	*	+	283.63	1097.50	1133.56	1988.80	*	+	194.45	1098.50	1115.58
Treatment Status Metric Matching (1:1)	1259.79	*		-534.55	1362.17	1463.30	1190.37	*		-603.97	1377.96	1504.51
Treatment Status Metric Matching (1:4)	1233.79	*		-560.55	1155.06	1283.89	1018.93	*		-775.41	1150.38	1387.31
Treatment Status Metric Matching (1:8)	223.13			-1571.21	948.47	1835.29	51.01			-1743.33	937.65	1979.50
Panel B: Matching with Bias-Correction												
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
Propensity Score Matching (1:1)	514.34			-1280.01	514.34	1379.48	550.88	*		-1243.46	1860.69	2237.93
Propensity Score Matching (1:4)	2179.60	*		385.25	1549.71	1596.88	2121.15	*		326.81	1557.88	1591.79
Propensity Score Matching (1:8)	1258.56	*		-535.78	1410.40	1508.73	1162.21	*		-632.13	1431.77	1565.11
Euclidean Metric Matching (1:1)	2164.44	*		370.10	1504.18	1549.04	2018.86	*		224.52	1520.70	1537.18
Euclidean Metric Matching (1:4)	2687.07	*	+	892.72	1278.97	1559.72	2707.34	*	+	912.99	1271.57	1565.39
Euclidean Metric Matching (1:8)	2503.46	*	+	709.11	1204.80	1398.00	2446.61	*	+	652.27	1201.94	1367.52
Mahalanobis Metric Matching (1:1)	2793.25	*	+	998.91	1536.63	1832.77	2519.92	*		725.58	1702.16	1850.35
Mahalanobis Metric Matching (1:4)	2529.17	*	+	734.83	1275.46	1472.00	2458.96	*	+	664.62	1377.00	1529.00
Mahalanobis Metric Matching (1:8)	2461.20	*	+	666.85	1113.36	1297.80	2438.26	*	+	643.92	1114.19	1286.87
Abadie and Imbens Metric Matching (1:1)	2537.02	*		742.68	1686.76	1843.03	2374.69	*		580.34	1709.32	1805.15
Abadie and Imbens Metric Matching (1:4)	2396.05	*	+	601.71	1391.83	1516.32	2332.11	*		537.77	1467.61	1563.03
Abadie and Imbens Metric Matching (1:8)	2430.88	*	+	636.53	1139.79	1305.49	2309.19	*	+	514.85	1216.79	1321.23
Outcome Metric Matching (1:1)	2666.31	*	+	871.96	1440.33	1683.71	2634.32	*	+	839.97	1505.66	1724.11
Outcome Metric Matching (1:4)	2570.90	*	+	776.56	1246.51	1468.61	2505.80	*	+	711.46	1250.00	1438.29
Outcome Metric Matching (1:8)	2432.02	*	+	637.68	1094.33	1266.57	2404.62	*	+	610.27	1094.08	1252.78
Treatment Status Metric Matching (1:1)	2111.34	*		317.00	1431.35	1466.03	2019.13	*		224.78	1448.67	1466.00
Treatment Status Metric Matching (1:4)	2577.44	*	+	783.10	1160.75	1400.21	2420.08	*	+	625.73	1159.02	1317.15
Treatment Status Metric Matching (1:8)	2420.76	*	+	626.42	953.74	1141.06	2284.50	*	+	490.16	941.92	1061.83

Note: 1. The specification of the propensity score is the same as in Table 3 of Dehejia and Wahba (1999), including constant, age, education, no degree, married, black, hispanic, age squared, education squared, re74, re75, re74 squared, re75 squared, u74*black.

2. The specification of the OLS is the same as the specification of the propensity score.

3. Standard errors are estimated using nnmatch based on the formula in Abadie and Imbens (2002).

4. The 95% confidence interval from the experiment benchmark is [551,3038]. In column (2) and (8), * indicates the estimate falls into this interval.

5. The benchmark is significant at 1% level. In column (3) and (9), "+"/"-"/blank indicates at 10% level, the estimate positively significant/negatively significant/insignificant.

Table 3 Estimates from Cross-Section Matching Using LaLonde's CPS Data

Panel A: Matching without Bias-Correction												
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
NSW Experiment (Benchmark)	886.30	*	+	0.00	472.09	472.09	886.30	*	+	0.00	472.09	472.09
Simple Mean Difference	-5953.77		-	-6840.07	528.25	6860.44	-8870.31		-	-9756.61	562.48	9772.81
OLS Regression	-944.73		-	-1831.03	457.99	1887.44	-1043.16		-	-1929.46	449.49	1981.12
Propensity Score Matching (1:1)	824.90	*		-61.41	679.00	681.77	824.69	*		-61.62	677.80	680.60
Propensity Score Matching (1:4)	-201.65			-1087.95	585.95	1235.71	-215.17			-1101.48	586.57	1247.92
Propensity Score Matching (1:8)	-243.45			-1129.75	566.09	1263.65	-246.34			-1132.64	566.80	1266.55
Euclidean Metric Matching (1:1)	-1613.46		-	-2499.76	650.68	2583.06	-1609.74		-	-2496.04	648.42	2578.89
Euclidean Metric Matching (1:4)	-1696.34		-	-2582.64	505.57	2631.66	-1736.68		-	-2622.98	504.92	2671.14
Euclidean Metric Matching (1:8)	-1858.59		-	-2744.89	474.60	2785.62	-1859.03		-	-2745.34	473.56	2785.88
Mahalanobis Metric Matching (1:1)	-400.67			-1286.97	688.95	1459.77	-549.78			-1436.08	700.22	1597.70
Mahalanobis Metric Matching (1:4)	-549.50			-1435.80	563.64	1542.47	-481.42			-1367.72	557.90	1477.13
Mahalanobis Metric Matching (1:8)	-450.70			-1337.01	535.73	1440.35	-411.44			-1297.75	528.75	1401.33
Abadie and Imbens Metric Matching (1:1)	-387.41			-1273.71	697.96	1452.41	-491.97			-1378.27	705.26	1548.23
Abadie and Imbens Metric Matching (1:4)	-392.84			-1279.14	565.55	1398.59	-471.44			-1357.75	558.87	1468.27
Abadie and Imbens Metric Matching (1:8)	-365.27			-1251.57	539.89	1363.05	-403.41			-1289.72	536.81	1396.97
Outcome Metric Matching (1:1)	-219.11			-1105.41	703.83	1310.46	-204.06			-1090.36	701.75	1296.66
Outcome Metric Matching (1:4)	-420.02			-1306.33	551.07	1417.80	-424.05			-1310.36	548.81	1420.64
Outcome Metric Matching (1:8)	-365.71			-1252.01	522.08	1356.50	-367.64			-1253.94	520.76	1357.78
Treatment Status Metric Matching (1:1)	430.32	*		-455.98	657.28	799.96	477.23	*		-409.08	655.76	772.90
Treatment Status Metric Matching (1:4)	-325.87			-1212.17	556.81	1333.94	-331.41			-1217.71	557.16	1339.12
Treatment Status Metric Matching (1:8)	-479.83			-1366.13	533.05	1466.44	-487.60			-1373.91	532.86	1473.62
Panel B: Matching with Bias-Correction												
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
Propensity Score Matching (1:1)	828.26	*		-58.05	678.94	681.42	828.97	*		-57.33	677.74	680.16
Propensity Score Matching (1:4)	-190.13			-1076.44	585.96	1225.59	-200.68			-1086.99	586.48	1235.11
Propensity Score Matching (1:8)	-219.48			-1105.78	566.41	1242.41	-218.42			-1104.72	567.07	1241.77
Euclidean Metric Matching (1:1)	-827.86			-1714.17	645.48	1831.67	-781.20			-1667.50	642.16	1786.88
Euclidean Metric Matching (1:4)	-1069.03		-	-1955.33	505.20	2019.54	-1023.06		-	-1909.36	503.49	1974.63
Euclidean Metric Matching (1:8)	-975.75		-	-1862.06	474.01	1921.44	-890.84		-	-1777.15	472.57	1838.90
Mahalanobis Metric Matching (1:1)	-412.63			-1298.93	697.47	1474.34	-557.14			-1443.44	707.00	1607.29
Mahalanobis Metric Matching (1:4)	-515.07			-1401.37	568.70	1512.37	-459.71			-1346.01	562.95	1458.99
Mahalanobis Metric Matching (1:8)	-388.43			-1274.73	538.69	1383.88	-367.28			-1253.58	532.91	1362.16
Abadie and Imbens Metric Matching (1:1)	-389.83			-1276.13	705.13	1457.98	-474.89			-1361.19	712.12	1536.21
Abadie and Imbens Metric Matching (1:4)	-361.83			-1248.13	569.63	1371.97	-432.43			-1318.73	563.66	1434.14
Abadie and Imbens Metric Matching (1:8)	-298.33			-1184.64	544.58	1303.81	-334.00			-1220.30	542.99	1335.65
Outcome Metric Matching (1:1)	-281.57			-1167.88	708.85	1366.16	-260.80			-1147.10	706.68	1347.31
Outcome Metric Matching (1:4)	-479.92			-1366.22	554.33	1474.40	-485.89			-1372.20	552.05	1479.09
Outcome Metric Matching (1:8)	-351.70			-1238.00	524.58	1344.56	-348.58			-1234.88	523.23	1341.15
Treatment Status Metric Matching (1:1)	401.47	*		-484.84	671.07	827.89	473.88	*		-412.42	669.23	786.10
Treatment Status Metric Matching (1:4)	-27.45	*		-913.75	566.70	1075.21	-28.34	*		-914.64	566.68	1075.96
Treatment Status Metric Matching (1:8)	-235.19			-1121.50	540.49	1244.95	-223.35			-1109.65	540.44	1234.26

Note: 1. The specification of the propensity score is the same as in Table 2 of Dehejia (2005a), including constant, age, education, married, black, hispanic, re75, u75*married re75*no degree, age squared.

2. The specification of the OLS is the same as the specification of the propensity score.

3. Standard errors are estimated using nnmatch based on the formula in Abadie and Imbens (2002).

4. The 95% confidence interval from the experiment benchmark is [-41,1813]. In column (2) and (8), * indicates the estimate falls into this interval.

5. The benchmark is significant at 6% level. In column (3) and (9), "+"/"-"/blank indicates at 10% level, the estimate positively significant/negatively significant/insignificant.

Table 4 Estimates from Cross-Section Matching Using LaLonde's PSID Data

Panel A: Matching without Bias-Correction											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
NSW Experiment (Benchmark)	886.30	*	+	0.00	472.09	472.09	886.30	*	+	0.00	472.09	472.09
Simple Mean Difference	-6849.91		-	-7736.21	689.59	7766.89	-15577.57		-	-16463.87	913.33	16489.19
OLS Regression	-1936.29		-	-2822.60	817.37	2938.56	-1345.55		-	-2231.86	805.38	2372.72
Propensity Score Matching (1:1)	705.20	*		-181.10	1116.08	1130.68	852.09	*		-34.22	1100.74	1101.27
Propensity Score Matching (1:4)	-1073.21			-1959.51	1367.79	2389.68	-947.62			-1833.93	1385.74	2298.60
Propensity Score Matching (1:8)	-1142.19			-2028.49	1204.91	2359.36	-1015.67			-1901.97	1214.39	2256.60
Euclidean Metric Matching (1:1)	-2810.01		-	-3696.31	1165.13	3875.60	-2667.13		-	-3553.43	1209.13	3753.51
Euclidean Metric Matching (1:4)	-3341.03		-	-4227.33	1001.85	4344.42	-3250.17		-	-4136.48	1007.30	4257.36
Euclidean Metric Matching (1:8)	-3456.38		-	-4342.69	900.55	4435.08	-3461.29		-	-4347.59	909.39	4441.68
Mahalanobis Metric Matching (1:1)	-474.71			-1361.02	1202.78	1816.32	-967.33			-1853.63	1163.63	2188.60
Mahalanobis Metric Matching (1:4)	-2204.70		-	-3091.00	1296.02	3351.71	-2175.13		-	-3061.43	1280.62	3318.48
Mahalanobis Metric Matching (1:8)	-2899.31		-	-3785.62	1227.67	3979.71	-2464.42		-	-3350.72	1151.28	3542.99
Abadie and Imbens Metric Matching (1:1)	-389.70			-1276.00	1359.42	1864.46	-401.80			-1288.10	1246.25	1792.30
Abadie and Imbens Metric Matching (1:4)	-2265.53		-	-3151.84	1333.00	3422.13	-2302.85		-	-3189.15	1305.39	3445.97
Abadie and Imbens Metric Matching (1:8)	-2039.07		-	-2925.38	1092.94	3122.87	-1733.38		-	-2619.68	1066.76	2828.55
Outcome Metric Matching (1:1)	-1364.81			-2251.11	1197.61	2549.86	-1332.04			-2218.35	1237.08	2539.97
Outcome Metric Matching (1:4)	-1573.80			-2460.10	1238.61	2754.31	-1523.29			-2409.59	1264.66	2721.31
Outcome Metric Matching (1:8)	-1675.30			-2561.61	1125.57	2797.99	-1679.64			-2565.94	1149.32	2811.58
Treatment Status Metric Matching (1:1)	-1017.14			-1903.44	1230.44	2266.51	-919.22			-1805.52	1269.26	2207.02
Treatment Status Metric Matching (1:4)	-2035.87			-2922.17	1451.28	3262.71	-1946.85			-2833.16	1471.92	3192.70
Treatment Status Metric Matching (1:8)	-1987.95		-	-2874.26	1183.12	3108.24	-1979.96		-	-2866.26	1211.43	3111.76
Panel B: Matching with Bias-Correction												
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
Propensity Score Matching (1:1)	703.22	*		-183.09	1115.50	1130.42	850.58	*		-35.72	1100.20	1100.78
Propensity Score Matching (1:4)	-1074.45			-1960.75	1366.07	2389.71	-944.40			-1830.71	1384.13	2295.06
Propensity Score Matching (1:8)	-1109.04			-1995.34	1204.36	2330.64	-973.87			-1860.17	1214.07	2221.31
Euclidean Metric Matching (1:1)	-1160.66			-2046.96	1158.65	2352.13	-677.33			-1563.64	1198.30	1970.00
Euclidean Metric Matching (1:4)	-1595.71			-2482.01	999.59	2675.73	-1598.76			-2485.06	1005.71	2680.86
Euclidean Metric Matching (1:8)	-1707.82		-	-2594.13	893.97	2743.84	-1669.05		-	-2555.36	901.24	2709.63
Mahalanobis Metric Matching (1:1)	-151.08			-1037.39	1214.43	1597.19	-564.17			-1450.48	1181.98	1871.09
Mahalanobis Metric Matching (1:4)	-1150.81			-2037.11	1311.71	2422.90	-1048.23			-1934.53	1289.33	2324.82
Mahalanobis Metric Matching (1:8)	-1376.75			-2263.06	1215.20	2568.68	-1051.45			-1937.75	1130.98	2243.66
Abadie and Imbens Metric Matching (1:1)	-229.58			-1115.89	1378.53	1773.57	-175.71			-1062.01	1265.43	1652.02
Abadie and Imbens Metric Matching (1:4)	-1375.54			-2261.85	1336.16	2627.03	-1338.99			-2225.29	1310.55	2582.53
Abadie and Imbens Metric Matching (1:8)	-961.60			-1847.91	1081.37	2141.05	-742.76			-1629.06	1055.15	1940.92
Outcome Metric Matching (1:1)	-1180.50			-2066.81	1201.94	2390.89	-1099.30			-1985.60	1253.79	2348.32
Outcome Metric Matching (1:4)	-1406.59			-2292.89	1222.42	2598.39	-1324.24			-2210.55	1249.09	2539.04
Outcome Metric Matching (1:8)	-1433.11			-2319.42	1111.55	2572.01	-1366.18			-2252.48	1131.66	2520.78
Treatment Status Metric Matching (1:1)	-945.64			-1831.95	1231.56	2207.43	-851.96			-1738.27	1270.56	2153.11
Treatment Status Metric Matching (1:4)	-1850.22			-2736.52	1455.56	3099.55	-1733.47			-2619.77	1470.07	3004.05
Treatment Status Metric Matching (1:8)	-1789.45			-2675.75	1192.61	2929.50	-1686.99			-2573.29	1216.04	2846.15

Note: 1. The specification of the propensity score is the same as in Table 2 of Dehejia (2005a), including constant, age, education, married, black, hispanic, re75, education*black, re75*hispanic, no degree*education.

2. The specification of the OLS is the same as the specification of the propensity score.

3. Standard errors are estimated using nnmatch based on the formula in Abadie and Imbens (2002).

4. The 95% confidence interval from the experiment benchmark is [-41,1813]. In column (2) and (8), * indicates the estimate falls into this interval.

5. The benchmark is significant at 6% level. In column (3) and (9), "+"/"-"/blank indicates at 10% level, the estimate positively significant/negatively significant/insignificant.

Table 5 Estimates from Difference-in-Differences Matching Using Dehejia and Wahba's CPS Data

Panel A: Matching without Bias-Correction											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
NSW Experiment (Benchmark)	1794.34	*	+	0.00	632.85	632.85	1794.34	*	+	0.00	632.85	632.85
Difference-in-Differences without Adj.	999.54	*	+	-794.80	544.85	963.62	3621.23	*	+	1826.89	570.74	1913.97
OLS Regression	1599.42	*	+	-194.92	582.02	613.79	1566.87	*	+	-227.48	556.94	601.60
Propensity Score Matching (1:1)	878.12	*		-916.22	1083.15	1418.69	673.18	*		-1121.16	1081.45	1557.73
Propensity Score Matching (1:4)	1584.01	*	+	-210.33	880.07	904.86	1547.39	*	+	-246.96	900.83	934.06
Propensity Score Matching (1:8)	1544.93	*	+	-249.42	822.81	859.78	1484.94	*	+	-309.40	837.30	892.64
Euclidean Metric Matching (1:1)	2017.01	*	+	222.67	826.83	856.29	1312.23	*		-482.11	842.32	970.53
Euclidean Metric Matching (1:4)	1553.57	*	+	-240.77	669.99	711.94	1219.13	*	+	-575.21	659.91	875.41
Euclidean Metric Matching (1:8)	1341.04	*	+	-453.30	627.38	774.01	1078.17	*	+	-716.18	630.14	953.93
Mahalanobis Metric Matching (1:1)	1631.53	*	+	-162.82	977.30	990.77	1741.11	*	+	-53.23	958.37	959.85
Mahalanobis Metric Matching (1:4)	1483.05	*	+	-311.29	775.84	835.96	1377.07	*	+	-417.27	771.59	877.19
Mahalanobis Metric Matching (1:8)	1199.42	*	+	-594.92	732.15	943.39	1118.60	*		-675.75	735.78	999.00
Abadie and Imbens Metric Matching (1:1)	2189.12	*	+	394.78	924.38	1005.15	1795.68	*	+	1.34	912.77	912.77
Abadie and Imbens Metric Matching (1:4)	1685.21	*	+	-109.13	785.79	793.33	1640.69	*	+	-153.66	786.70	801.56
Abadie and Imbens Metric Matching (1:8)	1499.09	*	+	-295.25	730.86	788.25	1487.49	*	+	-306.85	734.21	795.76
Outcome Metric Matching (1:1)	1838.55	*	+	44.21	973.17	974.18	1653.95	*	+	-140.39	1018.14	1027.78
Outcome Metric Matching (1:4)	1542.21	*	+	-252.13	790.15	829.40	1396.22	*	+	-398.12	800.57	894.10
Outcome Metric Matching (1:8)	1702.37	*	+	-91.97	718.70	724.56	1578.04	*	+	-216.30	720.57	752.33
Treatment Status Metric Matching (1:1)	2538.09	*	+	743.74	907.26	1173.15	2420.66	*	+	626.32	897.13	1094.13
Treatment Status Metric Matching (1:4)	2036.61	*	+	242.27	724.16	763.61	1986.54	*	+	192.20	702.55	728.36
Treatment Status Metric Matching (1:8)	2046.32	*	+	251.98	659.15	705.67	2010.84	*	+	216.50	648.63	683.81
Panel B: Matching with Bias-Correction												
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
Propensity Score Matching (1:1)	875.66	*		-918.68	1083.17	1420.29	672.47	*		-1121.87	1081.40	1558.21
Propensity Score Matching (1:4)	1584.54	*	+	-209.80	880.12	904.78	1550.56	*	+	-243.78	900.86	933.26
Propensity Score Matching (1:8)	1551.06	*	+	-243.29	822.84	858.05	1492.93	*	+	-301.41	837.27	889.87
Euclidean Metric Matching (1:1)	1892.62	*	+	98.28	827.02	832.84	1208.02	*		-586.33	827.91	1014.50
Euclidean Metric Matching (1:4)	1883.99	*	+	89.65	668.05	674.04	1511.52	*	+	-282.82	656.64	714.96
Euclidean Metric Matching (1:8)	1874.09	*	+	79.75	624.13	629.21	1509.86	*	+	-284.48	625.77	687.40
Mahalanobis Metric Matching (1:1)	1654.16	*	+	-140.18	971.94	982.00	1806.98	*	+	12.63	950.35	950.43
Mahalanobis Metric Matching (1:4)	1612.29	*	+	-182.05	773.72	794.85	1467.39	*	+	-326.95	769.30	835.90
Mahalanobis Metric Matching (1:8)	1452.60	*	+	-341.74	731.16	807.08	1349.62	*	+	-444.72	731.10	855.74
Abadie and Imbens Metric Matching (1:1)	2227.29	*	+	432.94	920.98	1017.67	1795.52	*	+	1.17	905.66	905.66
Abadie and Imbens Metric Matching (1:4)	1783.84	*	+	-10.50	779.54	779.61	1734.99	*	+	-59.35	778.82	781.07
Abadie and Imbens Metric Matching (1:8)	1652.51	*	+	-141.83	727.44	741.14	1622.28	*	+	-172.06	730.31	750.31
Outcome Metric Matching (1:1)	1889.73	*	+	95.39	961.25	965.98	1697.62	*	+	-96.72	1006.56	1011.19
Outcome Metric Matching (1:4)	1425.75	*	+	-368.59	785.58	867.75	1258.21	*		-536.13	794.51	958.48
Outcome Metric Matching (1:8)	1449.22	*	+	-345.12	714.93	793.87	1394.54	*	+	-399.80	714.45	818.71
Treatment Status Metric Matching (1:1)	2191.03	*	+	396.68	880.90	966.10	1879.35	*	+	85.01	884.06	888.14
Treatment Status Metric Matching (1:4)	1796.96	*	+	2.62	715.70	715.71	1589.28	*	+	-205.06	693.26	722.95
Treatment Status Metric Matching (1:8)	1843.96	*	+	49.62	656.49	658.37	1755.15	*	+	-39.19	643.90	645.09

Note: 1.The specification of the propensity score is the same as in Table 3 of Dehejia and Wahba (1999), including constant, age, education, no degree, married, black, hispanic, age squared, education squared, re74, re75, age cubed, u74, u75, education*re74.

2. The specification of the OLS is the same as the specification of the propensity score.
3. Standard errors are estimated using nnmatch based on the formula in Abadie and Imbens (2002).
4. The 95% confidence interval from the experiment benchmark is [551,3038]. In column (2) and (8), * indicates the estimate falls into this interval.
5. The benchmark is significant at 1% level. In column (3) and (9), "+"/"-"/blank indicates at 10% level, the estimate positively significant/negatively significant/insignificant.

Table 6 Estimates from Difference-in-Differences Matching Using Dehejia and Wahba's PSID Data

Panel A: Matching without Bias-Correction											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
NSW Experiment (Benchmark)	1794.34	*	+	0.00	632.85	632.85	1794.34	*	+	0.00	632.85	632.85
Difference-in-Differences without Adj.	1238.08	*	+	-556.27	756.64	939.12	2326.51	*	+	532.16	813.86	972.40
OLS Regression	26.27			-1768.07	1026.90	2044.65	216.85			-1577.50	1105.87	1926.51
Propensity Score Matching (1:1)	502.18			-1292.16	1932.54	2324.73	548.70			-1245.64	1895.34	2268.03
Propensity Score Matching (1:4)	2371.85	*		577.51	1571.43	1674.19	2312.46	*		518.12	1579.24	1662.06
Propensity Score Matching (1:8)	1245.75	*		-548.59	1452.98	1553.09	1154.97	*		-639.37	1473.86	1606.57
Euclidean Metric Matching (1:1)	897.10	*		-897.24	1464.77	1717.73	1002.87	*		-791.47	1476.12	1674.92
Euclidean Metric Matching (1:4)	1427.73	*		-366.62	1270.27	1322.12	1472.99	*		-321.35	1257.83	1298.23
Euclidean Metric Matching (1:8)	1464.11	*		-330.23	1191.81	1236.72	1435.25	*		-359.10	1186.79	1239.93
Mahalanobis Metric Matching (1:1)	2953.85	*	+	1159.51	1559.33	1943.19	2769.89	*		975.55	1715.23	1973.25
Mahalanobis Metric Matching (1:4)	2664.59	*	+	870.25	1284.86	1551.84	2601.45	*	+	807.10	1382.86	1601.16
Mahalanobis Metric Matching (1:8)	2466.65	*	+	672.30	1104.81	1293.29	2345.02	*	+	550.68	2345.02	2408.81
Abadie and Imbens Metric Matching (1:1)	2866.05	*	+	1071.70	1721.54	2027.86	2787.50	*		993.16	1730.05	1994.85
Abadie and Imbens Metric Matching (1:4)	2872.00	*	+	1077.65	1407.25	1772.48	3032.83	*	+	1238.48	1492.72	1939.60
Abadie and Imbens Metric Matching (1:8)	2632.12	*	+	837.77	1148.37	1421.48	2764.04	*	+	969.70	1238.63	1573.06
Outcome Metric Matching (1:1)	2996.34	*	+	1201.99	1428.64	1867.03	2962.22	*	+	1167.88	1485.37	1889.51
Outcome Metric Matching (1:4)	2506.64	*	+	712.30	1236.17	1426.70	2367.33	*	+	572.99	1240.15	1366.12
Outcome Metric Matching (1:8)	2229.21	*	+	434.86	1090.67	1174.17	2148.47	*	+	354.13	1090.26	1146.33
Treatment Status Metric Matching (1:1)	1812.34	*		17.99	1371.75	1371.87	1725.87	*		-68.47	1387.84	1389.53
Treatment Status Metric Matching (1:4)	2258.58	*	+	464.24	1138.77	1229.76	2086.37	*	+	292.03	1134.69	1171.67
Treatment Status Metric Matching (1:8)	1938.39	*	+	144.05	934.84	945.87	1815.32	*	+	20.98	923.38	923.62
Panel B: Matching with Bias-Correction												
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
Propensity Score Matching (1:1)	502.22			-1292.13	1932.66	2324.82	548.65			-1245.70	1895.09	2267.84
Propensity Score Matching (1:4)	2369.35	*		575.01	1569.84	1671.83	2310.56	*		516.22	1577.54	1659.86
Propensity Score Matching (1:8)	1267.26	*		-527.08	1448.24	1541.18	1180.07	*		-614.27	1468.41	1591.71
Euclidean Metric Matching (1:1)	2164.44	*		370.10	1504.18	1549.04	2018.86	*		224.52	1520.70	1537.18
Euclidean Metric Matching (1:4)	2687.07	*	+	892.72	1278.97	1559.72	2707.34	*	+	912.99	1271.57	1565.39
Euclidean Metric Matching (1:8)	2503.46	*	+	709.11	1204.80	1398.00	2446.61	*	+	652.27	1201.94	1367.52
Mahalanobis Metric Matching (1:1)	2793.25	*	+	998.91	1536.63	1832.77	2519.92	*		725.58	1702.16	1850.35
Mahalanobis Metric Matching (1:4)	2529.17	*	+	734.83	1275.46	1472.00	2458.96	*	+	664.62	1377.00	1529.00
Mahalanobis Metric Matching (1:8)	2461.20	*	+	666.85	1113.36	1297.80	2438.26	*	+	643.92	1114.19	1286.87
Abadie and Imbens Metric Matching (1:1)	2537.02	*		742.68	1686.76	1843.03	2374.69	*		580.34	1709.32	1805.15
Abadie and Imbens Metric Matching (1:4)	2396.05	*	+	601.71	1391.83	1516.32	2332.11	*		537.77	1467.61	1563.03
Abadie and Imbens Metric Matching (1:8)	2430.88	*	+	636.53	1139.79	1305.49	2309.19	*	+	514.85	1216.79	1321.23
Outcome Metric Matching (1:1)	2666.31	*	+	871.96	1440.33	1683.71	2634.32	*	+	839.97	1505.66	1724.11
Outcome Metric Matching (1:4)	2570.90	*	+	776.56	1246.51	1468.61	2505.80	*	+	711.46	1250.00	1438.29
Outcome Metric Matching (1:8)	2432.02	*	+	637.68	1094.33	1266.57	2404.62	*	+	610.27	1094.08	1252.78
Treatment Status Metric Matching (1:1)	2111.34	*		317.00	1431.35	1466.03	2019.13	*		224.79	1448.67	1466.00
Treatment Status Metric Matching (1:4)	2577.44	*	+	783.10	1160.75	1400.21	2420.08	*	+	625.73	1159.02	1317.15
Treatment Status Metric Matching (1:8)	2420.76	*	+	626.42	953.74	1141.06	2284.50	*	+	490.16	941.92	1061.83

Note: 1. The specification of the propensity score is the same as in Table 3 of Dehejia and Wahba (1999), including constant, age, education, no degree, married, black, hispanic, age squared, education squared, re74, re75, re74 squared, re75 squared, u74*black.

2. The specification of the OLS is the same as the specification of the propensity score.

3. Standard errors are estimated using nnmatch based on the formula in Abadie and Imbens (2002).

4. The 95% confidence interval from the experiment benchmark is [551,3038]. In column (2) and (8), * indicates the estimate falls into this interval.

5. The benchmark is significant at 1% level. In column (3) and (9), "+"/"-"/blank indicates at 10% level, the estimate positively significant/negatively significant/insignificant.

Table 7 Estimates from Difference-in-Differences Matching Using LaLonde's CPS Data

Panel A: Matching without Bias-Correction											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
NSW Experiment (Benchmark)	886.30	*	+	0.00	472.09	472.09	886.30	*	+	0.00	472.09	472.09
Difference-in-Differences without Adj.	215.35	*		-670.96	447.98	806.77	1714.40	*	+	828.09	452.29	943.56
OLS Regression	-944.73			-1831.03	457.99	1887.44	-1043.16		-	-1929.46	449.49	1981.12
Propensity Score Matching (1:1)	432.82	*		-453.48	717.16	848.51	433.93	*		-452.37	715.89	846.84
Propensity Score Matching (1:4)	-304.95			-1191.25	633.17	1349.07	-316.47			-1202.77	633.76	1359.53
Propensity Score Matching (1:8)	-311.84			-1198.14	603.83	1341.70	-306.73			-1193.04	605.04	1337.69
Euclidean Metric Matching (1:1)	-1654.56		-	-2540.86	657.24	2624.49	-1650.71		-	-2537.01	654.96	2620.19
Euclidean Metric Matching (1:4)	-1737.51		-	-2623.81	510.62	2673.03	-1777.69		-	-2663.99	509.94	2712.36
Euclidean Metric Matching (1:8)	-1899.91		-	-2786.22	479.50	2827.18	-1900.21		-	-2786.51	478.44	2827.28
Mahalanobis Metric Matching (1:1)	-461.87			-1348.17	706.94	1522.28	-608.72			-1495.03	714.79	1657.12
Mahalanobis Metric Matching (1:4)	-563.65			-1449.95	574.23	1559.52	-522.68			-1408.98	568.10	1519.20
Mahalanobis Metric Matching (1:8)	-434.33			-1320.63	544.43	1428.45	-426.54			-1312.85	538.80	1419.11
Abadie and Imbens Metric Matching (1:1)	-417.88			-1304.18	713.12	1486.41	-515.78			-1402.08	720.57	1576.41
Abadie and Imbens Metric Matching (1:4)	-392.96			-1279.26	576.82	1403.29	-466.17			-1352.47	570.37	1467.82
Abadie and Imbens Metric Matching (1:8)	-328.93			-1215.23	550.45	1334.09	-382.95			-1269.25	549.66	1383.16
Outcome Metric Matching (1:1)	-256.20			-1142.50	712.89	1346.67	-240.25			-1126.55	710.81	1332.06
Outcome Metric Matching (1:4)	-461.44			-1347.74	557.15	1458.37	-464.56			-1350.86	554.92	1460.39
Outcome Metric Matching (1:8)	-401.08			-1287.39	527.93	1391.43	-403.15			-1289.45	526.57	1392.82
Treatment Status Metric Matching (1:1)	337.77	*		-548.53	678.95	872.84	400.92			-485.39	677.00	833.02
Treatment Status Metric Matching (1:4)	-434.47			-1320.78	571.92	1439.28	-422.82			-1309.13	571.87	1428.58
Treatment Status Metric Matching (1:8)	-573.18			-1459.48	546.66	1558.50	-569.67			-1455.97	546.41	1555.13
Panel B: Matching with Bias-Correction												
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
Propensity Score Matching (1:1)	433.17	*		-453.13	717.15	848.31	434.41	*		-451.89	715.89	846.58
Propensity Score Matching (1:4)	-305.91			-1192.21	633.16	1349.91	-317.74			-1204.04	633.76	1360.65
Propensity Score Matching (1:8)	-312.56			-1198.86	603.82	1342.34	-307.71			-1194.02	605.03	1338.56
Euclidean Metric Matching (1:1)	-827.86			-1714.17	645.48	1831.67	-781.20			-1667.50	642.16	1786.88
Euclidean Metric Matching (1:4)	-1069.03		-	-1955.33	505.20	2019.54	-1023.06		-	-1909.36	503.49	1974.63
Euclidean Metric Matching (1:8)	-975.75		-	-1862.06	474.01	1921.44	-890.84		-	-1777.15	472.57	1838.90
Mahalanobis Metric Matching (1:1)	-412.63			-1298.93	697.47	1474.34	-557.14			-1443.44	707.00	1607.29
Mahalanobis Metric Matching (1:4)	-515.07			-1401.37	568.70	1512.37	-459.71			-1346.01	562.95	1458.99
Mahalanobis Metric Matching (1:8)	-388.43			-1274.73	538.69	1383.88	-367.28			-1253.58	532.91	1362.16
Abadie and Imbens Metric Matching (1:1)	-389.83			-1276.13	705.13	1457.98	-474.89			-1361.19	712.12	1536.21
Abadie and Imbens Metric Matching (1:4)	-361.83			-1248.13	569.63	1371.97	-432.43			-1318.73	563.66	1434.14
Abadie and Imbens Metric Matching (1:8)	-298.33			-1184.64	544.58	1303.81	-334.00			-1220.30	542.99	1335.65
Outcome Metric Matching (1:1)	-281.57			-1167.88	708.85	1366.16	-260.80			-1147.10	706.68	1347.31
Outcome Metric Matching (1:4)	-479.92			-1366.22	554.33	1474.40	-485.89			-1372.20	552.05	1479.09
Outcome Metric Matching (1:8)	-351.70			-1238.00	524.58	1344.56	-348.58			-1234.88	523.23	1341.15
Treatment Status Metric Matching (1:1)	401.47	*		-484.84	671.07	827.89	473.88	*		-412.42	669.23	786.10
Treatment Status Metric Matching (1:4)	-27.45	*		-913.75	566.70	1075.21	-28.34	*		-914.64	566.68	1075.96
Treatment Status Metric Matching (1:8)	-235.19			-1121.50	540.49	1244.95	-223.35			-1109.65	540.44	1234.26

Note: 1. The specification of the propensity score is the same as in Table 2 of Dehejia (2005a), including constant, age, education, married, black, hispanic, re75, u75*married re75*no degree, age squared.

2. The specification of the OLS is the same as the specification of the propensity score.

3. Standard errors are estimated using nnmatch based on the formula in Abadie and Imbens (2002).

4. The 95% confidence interval from the experiment benchmark is [-41,1813]. In column (2) and (8), * indicates the estimate falls into this interval.

5. The benchmark is significant at 6% level. In column (3) and (9), "+"/"-"/blank indicates at 10% level, the estimate positively significant/negatively significant/insignificant.

Table 8 Estimates from Difference-in-Differences Matching Using LaLonde's PSID Data

Panel A: Matching without Bias-Correction											
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
NSW Experiment (Benchmark)	886.30	*	+	0.00	472.09	472.09	886.30	*	+	0.00	472.09	472.09
Difference-in-Differences without Adj.	-738.52			-1624.82	621.15	1739.51	419.67	*		-466.63	650.58	800.63
OLS Regression	-1936.29		-	-2822.60	817.37	2938.56	-1345.55		-	-2231.86	805.38	2372.72
Propensity Score Matching (1:1)	608.39	*		-277.91	1220.24	1251.48	736.16	*		-150.15	1189.84	1199.27
Propensity Score Matching (1:4)	-1034.80			-1921.10	1431.70	2395.91	-907.97			-1794.28	1443.79	2303.03
Propensity Score Matching (1:8)	-921.43			-1807.73	1284.61	2217.68	-808.16			-1694.46	1288.20	2128.53
Euclidean Metric Matching (1:1)	-2809.53		-	-3695.84	1165.27	3875.18	-2668.45		-	-3554.75	1209.43	3754.86
Euclidean Metric Matching (1:4)	-3334.76		-	-4221.06	1000.66	4338.05	-3250.41		-	-4136.71	1007.46	4257.62
Euclidean Metric Matching (1:8)	-3462.38		-	-4348.68	900.90	4441.02	-3462.92		-	-4349.22	909.35	4443.27
Mahalanobis Metric Matching (1:1)	-200.56			-1086.87	1240.70	1649.43	-585.25			-1471.55	1208.39	1904.12
Mahalanobis Metric Matching (1:4)	-1491.66			-2377.97	1309.34	2714.61	-1178.97			-2065.27	1293.91	2437.12
Mahalanobis Metric Matching (1:8)	-1858.86			-2745.16	1240.35	3012.37	-1136.87			-2023.17	1160.84	2332.54
Abadie and Imbens Metric Matching (1:1)	-224.54			-1110.85	1424.85	1806.70	-212.24			-1098.55	1304.56	1705.48
Abadie and Imbens Metric Matching (1:4)	-1619.24			-2505.54	1353.25	2847.63	-1437.16			-2323.47	-1437.16	2732.02
Abadie and Imbens Metric Matching (1:8)	-1144.41			-2030.71	1107.44	2313.05	-616.12			-1502.42	1082.39	1851.71
Outcome Metric Matching (1:1)	-1350.26			-2236.56	1202.99	2539.57	-1322.87			-2209.17	1244.24	2535.46
Outcome Metric Matching (1:4)	-1568.86			-2455.16	1238.59	2749.90	-1523.33			-2409.63	1267.08	2722.46
Outcome Metric Matching (1:8)	-1634.90			-2521.20	1125.34	2760.95	-1632.11			-2518.42	1148.43	2767.91
Treatment Status Metric Matching (1:1)	-900.18			-1786.49	1243.12	2176.44	-805.74			-1692.04	1282.08	2122.91
Treatment Status Metric Matching (1:4)	-1845.82			-2732.12	1467.23	3101.17	-1745.62			-2631.92	1484.54	3021.73
Treatment Status Metric Matching (1:8)	-1430.75			-2317.05	1204.63	2611.49	-1393.34			-2279.65	1228.87	2589.77
Panel B: Matching with Bias-Correction												
	With Common Support Condition						Without Common Support Condition					
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error	MSE
Propensity Score Matching (1:1)	608.03	*		-278.27	1220.14	1251.47	735.94	*		-150.36	1189.77	1199.23
Propensity Score Matching (1:4)	-1035.16			-1921.46	1431.23	2395.92	-907.06			-1793.36	1443.36	2302.05
Propensity Score Matching (1:8)	-912.34			-1798.65	1284.51	2210.22	-797.03			-1683.33	1288.15	2119.66
Euclidean Metric Matching (1:1)	-1160.66			-2046.96	1158.65	2352.13	-677.33			-1563.64	1198.30	1970.00
Euclidean Metric Matching (1:4)	-1595.71			-2482.01	999.59	2675.73	-1598.76			-2485.06	1005.71	2680.86
Euclidean Metric Matching (1:8)	-1707.82		-	-2594.13	893.97	2743.84	-1669.05		-	-2555.36	901.24	2709.63
Mahalanobis Metric Matching (1:1)	-151.08			-1037.39	1214.43	1597.19	-564.17			-1450.48	1181.98	1871.09
Mahalanobis Metric Matching (1:4)	-1150.81			-2037.11	1311.71	2422.90	-1048.23			-1934.53	1289.33	2324.82
Mahalanobis Metric Matching (1:8)	-1376.75			-2263.06	1215.20	2568.68	-1051.45			-1937.75	1130.98	2243.66
Abadie and Imbens Metric Matching (1:1)	-229.58			-1115.89	1378.53	1773.57	-175.71			-1062.01	1265.43	1652.02
Abadie and Imbens Metric Matching (1:4)	-1375.54			-2261.85	1336.16	2627.03	-1338.99			-2225.29	1310.55	2582.53
Abadie and Imbens Metric Matching (1:8)	-961.60			-1847.91	1081.37	2141.05	-742.76			-1629.06	1055.15	1940.92
Outcome Metric Matching (1:1)	-1180.50			-2066.81	1201.94	2390.89	-1099.30			-1985.60	1253.79	2348.32
Outcome Metric Matching (1:4)	-1406.59			-2292.89	1222.42	2598.39	-1324.24			-2210.55	1249.09	2539.04
Outcome Metric Matching (1:8)	-1433.11			-2319.42	1111.55	2572.01	-1366.18			-2252.48	1131.66	2520.78
Treatment Status Metric Matching (1:1)	-945.64			-1831.95	1231.56	2207.43	-851.96			-1738.27	1270.56	2153.11
Treatment Status Metric Matching (1:4)	-1850.22			-2736.52	1455.56	3099.55	-1733.47			-2619.77	1470.07	3004.05
Treatment Status Metric Matching (1:8)	-1789.45			-2675.75	1192.61	2929.50	-1686.99			-2573.29	1216.04	2846.15

Note: 1. The specification of the propensity score is the same as in Table 2 of Dehejia (2005a), including constant, age, education, married, black, hispanic, re75, education*black, re75*hispanic, no degree*education.

2. The specification of the OLS is the same as the specification of the propensity score.

3. Standard errors are estimated using nnmatch based on the formula in Abadie and Imbens (2002).

4. The 95% confidence interval from the experiment benchmark is [-41,1813]. In column (2) and (8), * indicates the estimate falls into this interval.

5. The benchmark is significant at 6% level. In column (3) and (9), "+"/"-"/blank indicates at 10% level, the estimate positively significant/negatively significant/insignificant.

Table 9 Means of Covariates before and after Matching without Using 1974 Earnings in Dehejia and Wahba's Data

Variable	Sample	(1) Mean in Treated	(2) Mean in Comparison	(3) %bias	(4) % reduction of |bias|	(5) t-statistics	(6) p>|t|
Panel A: CPS Data							
age	Unmatched	25.82	33.23	-79.60		-9.10	0.00
	Matched	25.82	26.34	-5.60	92.90	-0.63	0.53
age squared	Unmatched	717.39	1225.90	-80.30		-8.80	0.00
	Matched	717.39	748.62	-4.90	93.90	-0.64	0.52
education	Unmatched	10.35	12.03	-67.90		-7.94	0.00
	Matched	10.35	10.36	-0.40	99.40	-0.03	0.97
married	Unmatched	0.19	0.71	-123.30		-15.62	0.00
	Matched	0.19	0.19	-1.30	99.00	-0.12	0.91
black	Unmatched	0.84	0.07	242.80		39.66	0.00
	Matched	0.84	0.84	1.70	99.30	0.13	0.90
hispanic	Unmatched	0.06	0.07	-5.10		-0.66	0.51
	Matched	0.06	0.06	0.00	100.00	0.00	1.00
1975 earnings	Unmatched	1532.10	13651.00	-174.60		-17.77	0.00
	Matched	1532.10	1493.10	0.60	99.70	0.10	0.92
1975 earnings squared	Unmatched	1300.00	27000.00	-145.00		-14.30	0.00
	Matched	1300.00	1500.00	-1.60	98.90	-0.41	0.68
age*black	Unmatched	21.91	2.40	187.60		29.10	0.00
	Matched	21.91	21.91	0.00	100.00	0.00	1.00
hispanic*married	Unmatched	0.02	0.05	-20.00		-2.21	0.03
	Matched	0.02	0.02	0.00	100.00	0.00	1.00
Panel B: PSID Data							
age	Unmatched	25.82	34.85	-100.90		-11.57	0.00
	Matched	25.82	25.58	2.70	97.40	0.28	0.78
education	Unmatched	10.35	12.12	-68.10		-7.69	0.00
	Matched	10.35		-5.80	91.50	-0.61	0.54
no degree	Unmatched	0.71	0.31	87.90		11.49	0.00
	Matched	0.71	0.70	2.40	97.30	0.19	0.85
married	Unmatched	0.19	0.87	-184.20		-25.81	0.00
	Matched	0.19	0.19	0.00	100.00	0.00	1.00
black	Unmatched	0.84	0.25	148.00		18.13	0.00
	Matched	0.84	0.85	-1.30	99.10	-0.12	0.91
hispanic	Unmatched	0.06	0.03	12.90		1.94	0.05
	Matched	0.06	0.08	-7.70	39.80	-0.52	0.61
age squared	Unmatched	717.39	1323.50	-97.10		-10.59	0.00
	Matched	717.39	696.90	3.30	96.60	0.38	0.70
education squared	Unmatched	111.06	156.32	-78.50		-8.52	0.00
	Matched	111.06	114.01	-5.10	93.50	-0.61	0.55
1975 earnings	Unmatched	1532.10	19063.00	-177.40		-17.50	0.00
	Matched	1532.10	1742.00	-2.10	98.80	-0.52	0.60
1975 earnings squared	Unmatched	1300.00	55000.00	-82.90		-7.98	0.00
	Matched	1300.00	1400.00	-0.10	99.80	-0.13	0.90
no earning in 1975	Unmatched	0.40	0.90	-122.80		-20.70	0.00
	Matched	0.40	0.43	-8.00	93.50	-0.53	0.60
hispanic*(no earning in 1975)	Unmatched	0.03	0.03	3.00		0.41	0.68
	Matched	0.03	0.06	-15.90	-427.50	-1.03	0.30

Note: 1. re75 squared is divided by 10,000.

2. This table is produced by psmatch2 in Leuven and Sianesi (2003)

Table 10 Estimates from Cross-Section Matching in Dehejia and Wahba's Data without Using 1974 Earings

Panel A: Matching without Bias-Correction, CPS data										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	With Common Support Condition						Without Common Support Cond				
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error
NSW Experiment (Benchmark)	1794.34	*	+	0.00	632.85	632.85	1794.34	*	+	0.00	632.85
Simple Mean Difference	-1956.07		-	-3750.42	598.81	3797.92	-8497.52		-	-10291.86	712.02
OLS Regression	1085.08	*		-709.26	580.77	916.71	1079.08	*		-715.26	568.24
Propensity Score Matching (1:1)	1442.40	*		-351.94	985.43	1046.39	1463.87	*		-330.47	953.72
Propensity Score Matching (1:4)	1255.02	*	+	-539.32	763.58	934.84	1255.02	*	+	-539.32	760.89
Propensity Score Matching (1:8)	1135.37	*		-658.97	715.34	972.60	1112.84	*		-681.50	713.75
Euclidean Metric Matching (1:1)	656.43	*		-1137.91	794.44	1387.80	362.06			-1432.28	825.41
Euclidean Metric Matching (1:4)	227.73			-1566.61	638.08	1691.57	211.39			-1582.95	643.54
Euclidean Metric Matching (1:8)	207.40			-1586.94	608.54	1699.62	130.63			-1663.71	610.55
Mahalanobis Metric Matching (1:1)	1343.99	*		-450.35	851.96	963.67	1514.00	*	+	-280.34	833.73
Mahalanobis Metric Matching (1:4)	1228.47	*	+	-565.87	716.31	912.86	1261.71	*	+	-532.63	707.19
Mahalanobis Metric Matching (1:8)	1101.98	*	+	-692.36	677.95	969.01	1179.81	*	+	-614.53	676.73
Abadie and Imbens Metric Matching (1:1)	1405.80	*	+	-388.55	838.74	924.36	1499.19	*	+	-295.15	835.66
Abadie and Imbens Metric Matching (1:4)	1296.45	*	+	-497.90	705.40	863.42	1227.12	*	+	-567.23	705.53
Abadie and Imbens Metric Matching (1:8)	1091.58	*		-702.76	679.66	977.66	1146.91	*	+	-647.44	685.57
Outcome Metric Matching (1:1)	1036.15	*		-758.20	898.35	1175.54	1036.15	*		-758.20	898.35
Outcome Metric Matching (1:4)	1189.32	*	+	-605.02	716.48	937.76	1149.68	*		-644.66	717.98
Outcome Metric Matching (1:8)	1343.72	*	+	-450.62	680.19	815.92	1297.05	*	+	-497.29	681.36
Treatment Status Metric Matching (1:1)	1928.73	*	+	134.39	814.79	825.80	1950.75	*	+	156.41	816.00
Treatment Status Metric Matching (1:4)	1193.11	*	+	-601.23	718.26	936.69	1203.84	*	+	-590.50	718.63
Treatment Status Metric Matching (1:8)	1169.00	*	+	678.13	546.66	871.03	1161.11	*	+	-633.23	678.19
Panel B: Matching without Bias-Correction, PSID data											
	With Common Support Condition						Without Common Support Cond				
Methods	Coef.			Bias	Std. Error	MSE	Coef.			Bias	Std. Error
NSW Experiment (Benchmark)	1794.34	*	+	0.00	632.85	632.85	1794.34	*	+	0.00	632.85
Simple Mean Difference	-6062.82		-	-7857.16	856.99	7903.76	-15204.78		-	-16999.12	1154.61
OLS Regression	394.23			-1400.12	965.10	1700.51	832.91	*		-961.43	980.44
Propensity Score Matching (1:1)	-691.50			-2485.84	1280.25	2796.15	-570.83			-2365.17	2187.37
Propensity Score Matching (1:4)	-559.63			-2353.97	1326.66	2702.08	-467.05			-2261.39	1616.46
Propensity Score Matching (1:8)	-880.20			-2674.54	1262.77	2957.66	-818.50			-2612.84	1634.23
Euclidean Metric Matching (1:1)	-312.69			-2107.03	1729.70	2726.06	-277.82			-2072.16	1696.15
Euclidean Metric Matching (1:4)	-1387.99			-3182.33	1417.47	3483.74	-1561.72			-3356.07	1451.13
Euclidean Metric Matching (1:8)	-1557.33			-3351.67	1211.61	3563.95	-1728.46			-3522.80	1259.71
Mahalanobis Metric Matching (1:1)	207.41			-1586.93	1444.67	2146.03	-166.75			-1961.10	1396.87
Mahalanobis Metric Matching (1:4)	-852.44			-2646.78	1465.76	3025.54	-1843.32			-3637.66	1417.74
Mahalanobis Metric Matching (1:8)	-1886.98			-3681.33	1368.71	3927.53	-2160.42		-	-3954.77	1152.97
Abadie and Imbens Metric Matching (1:1)	106.82			-1687.52	1548.03	2290.01	-235.51			-2029.85	1496.14
Abadie and Imbens Metric Matching (1:4)	-1372.55			-3166.89	1664.66	3577.75	-2010.37		-	-3804.72	1206.01
Abadie and Imbens Metric Matching (1:8)	-1821.83			-3616.17	1409.64	3881.21	-818.50			-2612.84	1634.23
Outcome Metric Matching (1:1)	169.26			-1625.08	1516.73	2222.92	-106.96			-1901.30	1559.22
Outcome Metric Matching (1:4)	-310.02			-2104.37	1575.40	2628.73	-345.06			-2139.40	1585.97
Outcome Metric Matching (1:8)	-367.71			-2162.05	1504.98	2634.27	-344.04			-2138.39	1504.54
Treatment Status Metric Matching (1:1)	-312.42			-2106.77	1552.43	2616.96	-489.82			-2284.16	1558.66
Treatment Status Metric Matching (1:4)	-1318.04			-3112.38	1580.93	3490.88	-1324.02			-3118.36	1579.48
Treatment Status Metric Matching (1:8)	-1212.07			-3006.41	1329.15	3287.12	-1349.97			-3144.31	1348.89

Note: 1. Propensity score specifications are selected through balancing tests using psmatch2 (Leuven and Sianesi, 2003). For covariates in the propensity scores, please refer to table 9.

2. The specification of the OLS is the same as the specification of the propensity score.

3. Standard errors are estimated using nnmatch based on the formula in Abadie and Imbens (2002).

4. The 95% confidence interval from the experiment benchmark is [551,3038]. In column (2) and (8), * indicates the estimate falls into this interval.

5. The benchmark is significant at 1% level. In column (3) and (9), "+"/"-"/blank indicates at 10% level, the estimate positively significant/negatively significant/insignificant.

Table 11 Decomposition of Selection-Bias

	Treatment Effect on Treated from Experiment	Total Bias	Bias from Non-Overlapping Support	Bias on Observable	Bias on Unobservable
LaLonde CPS Data	886.30	-9756.61	-2916.54	-2083.03	-4757.04
Percent of Total Bias		100.00%	29.89%	21.35%	48.76%
LaLonde PSID Data	886.30	-16463.87	-8727.66	-1736.80	-5999.41
Percent of Total Bias		100.00%	53.01%	10.55%	36.44%
DW CPS Data	1794.34	-10291.86	-6758.50	-867.25	-2666.11
Percent of Total Bias		100.00%	65.67%	8.43%	25.91%
DW PSID Data	1794.34	-16999.12	-8497.63	-2173.10	-6328.40
Percent of Total Bias		100.00%	49.99%	12.78%	37.23%