NBER WORKING PAPER SERIES

CAUSAL EFFECTS IN NON-EXPERIMENTAL STUDIES: RE-EVALUATING THE EVALUATION OF TRAINING PROGRAMS

Rajeev H. Dehejia
Sadek Wahba

Working Paper 6586
http://www.nber.org/papers/w6586

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 1998

This paper is a revised version of a lengthier (1995) paper circulating under the same title. This work was partially supported by a grant from the Social Sciences and Humanities Research Council of Canada (first author) and a World Bank Fellowship (second author). The authors gratefully acknowledge: the support and encouragement of Gary Chamberlain, Guido Imbens, and Donald Rubin; Robert Lalonde, who kindly provided the data from his 1986 study and substantial help in recreating the original data set; Joshua Angrist, George Cave, David Cutler, Lawrence Katz, Caroline Minter Hoxby, seminar participants at Harvard, MDRC, and MIT for many suggestions; and an associate editor and two referees for detailed comments. All remaining errors are the authors' responsibility. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

Re-Evaluating the Evaluation of Training Programs
Rajeev H. Dehejia and Sadek Wahba
NBER Working Paper No. 6586
June 1998
JEL Nos. C81, C14

ABSTRACT

This paper uses propensity score methods to address the question: how well can an observational study estimate the treatment impact of a program? Using data from Lalonde's (1986) influential evaluation of non-experimental methods, we demonstrate that propensity score methods succeed in estimating the treatment impact of the National Supported Work Demonstration. Propensity score methods reduce the task of controlling for differences in pre-intervention variables between the treatment and the non-experimental comparison groups to controlling for differences in the estimated propensity score (the probability of assignment to treatment, conditional on covariates). It is difficult to control for differences in pre-intervention variables when they are numerous and when the treatment and comparison groups are dissimilar, whereas controlling for the estimated propensity score, a single variable on the unit interval, is a straightforward task. We apply several methods, such as stratification on the propensity score and matching on the propensity score, and show that they result in accurate estimates of the treatment impact.

Rajeev H. Dehejia
Department of Economics
Columbia University
420 West 118th Street, 1022 IAB
New York, NY 10027
and NBER

Sadek Wahba
Morgan Stanley and Company
1585 Broadway
New York, NY 10036

1. INTRODUCTION

This paper discusses the estimation of treatment effects in observational studies. This issue, which is of great practical importance because randomized experiments cannot always be implemented, has been addressed previously by Lalonde (1986), whose data we use in this paper. Lalonde estimates the impact of the National Supported Work (NSW) Demonstration, a labor training program, on post-intervention income levels, using data from a randomized evaluation of the program. He then examines the extent to which non-experimental estimators can replicate the unbiased experimental estimate of the treatment impact, when applied to a composite data set of experimental treatment units and non-experimental comparison units. He concludes that standard non-experimental estimators, such as regression, fixed-effect, and latent-variable-selection models, are either inaccurate (relative to the experimental benchmark), or sensitive to the specification used in the regression. Lalonde's results have been influential in renewing the debate on experimental versus non-experimental evaluations (see Manski and Garfinkel 1992) and in spurring a search for alternative estimators and specification tests (e.g., Heckman and Hotz 1989; and Manski, Sandefur, McLanahan, and Powers 1992).

In this paper, we apply propensity score methods (Rosenbaum and Rubin 1983) to Lalonde's data set. Propensity score methods focus on the comparability of the treatment and non-experimental comparison groups in terms of pre-intervention variables. Controlling for differences in pre-intervention variables is difficult when the treatment and comparison groups are dissimilar and when there are many pre-intervention variables. The propensity score (the probability of assignment to treatment, conditional on covariates) summarizes the pre-intervention variables. We can easily control for differences between the treatment and non-experimental comparison groups through the estimated propensity score, a single variable on the unit interval. Using propensity score methods, we are able to replicate the experimental treatment effect for a range of specifications and estimators.

The assumption underlying the method is that assignment to treatment depends only on observable pre-intervention variables (called the ignorable treatment assignment assumption or selection on observables; see Rubin 1974, 1977, 1978; Heckman and Robb 1985; or Holland 1986). Though this is a strong assumption, we demonstrate that propensity score methods are an informative starting point, because they quickly reveal the extent to which the treatment and comparison groups overlap in terms of pre-intervention variables.

The paper is organized as follows. Section 2 reviews Lalonde's data and replicates his results. Section 3 identifies the treatment effect under the potential outcomes causal model, and discusses estimation strategies for the treatment effect. In Section 4, we apply our methods to Lalonde's data set, and in Section 5, we discuss the sensitivity of the results to the methodology. Section 6 concludes the paper.

2. LALONDE'S RESULTS

2.1 The Data

The National Supported Work (NSW) Demonstration (see Manpower Demonstration Research Corporation 1983) was a federally-funded program implemented in the mid-1970s, with the objective of providing work experience for a period of 12 to 18 months to individuals who had faced economic and social problems prior to enrollment in the program. Those randomly selected to join the program participated in various types of work, ranging from operating a restaurant to construction work. Information on pre-intervention variables (pre-intervention earnings as well as education, age, ethnicity, and marital status) was obtained from initial surveys and Social Security Administration records. In this paper we focus on the male participants, since estimates for this group were the most sensitive to functional-form specification, as indicated in Lalonde (1986). Both the treatment and control groups participated in follow-up interviews at specific intervals. The outcome variable of interest is post-intervention (1978) earnings. Unlike typical clinical trials, the eligible candidates did not join the NSW program immediately, but were randomized in over a period of 51 months between March 1975 and June 1977. This introduced

what the administrators of the program have referred to as the "cohort phenomenon" (MDRC 1983, p. 48): individuals who joined early in the program had different characteristics than those who entered later.

Lalonde limits his sample to those assigned between January 1976 and July 1977 in order to achieve homogeneity within the treatment and control groups, reducing the sample to 297 treated observations and 425 control observations for male participants. His sample is limited to one year of pre-intervention earnings data (1975). However, several years of pre-intervention earnings are viewed as important in determining the effect of job training programs (Angrist 1990, 1998; Ashenfelter 1978; Ashenfelter and Card 1985; and Card and Sullivan 1988). Thus, we further limit ourselves to a subset of this data in order to obtain data on earnings in 1974. Our subset, also defined using the month of assignment, includes 185 treated and 260 control observations. Since month of assignment is a pre-treatment variable, this selection does not affect the properties of the experimentally randomized data set: the treatment and control groups still have the same distribution of pre-intervention variables, so that a difference in means remains an unbiased estimate of the average treatment impact.

We present the pre-intervention characteristics of the original sample and of our subset in the first four rows of Table 1. The distribution of pre-intervention variables is very similar across the treatment and control groups for both samples (none of the differences is statistically significant), but our subset differs somewhat from Lalonde's original sample, especially in terms of 1975 earnings. Our propensity score results will be based on our subset of the data, using two years of pre-intervention earnings. In order to render our results comparable to Lalonde's, we replicate his analysis on our subset (both with and without the additional year of pre-intervention earnings data), and show that his basic conclusions remain unchanged. As well, in Section 5, we discuss the sensitivity of our propensity score results to dropping the additional earnings data.

2.2 Lalonde's Results

Lalonde estimates linear regression, fixed-effect, and latent variable selection models of the treatment impact. Since our analysis focuses on the importance of pre-intervention variables, we consider primarily the first of these. Non-experimental estimates of the treatment effect are based on the two distinct comparison groups used by Lalonde (1986), the Panel Study of Income Dynamics (PSID-1) and Westat's Matched Current Population Survey-Social Security Administration File (CPS-1). Lalonde also considers subsets of these two comparison groups, PSID2-3 and CPS2-3.

Table 1 presents the pre-intervention characteristics of the comparison groups. It is evident that both PSID-1 and CPS-1 differ dramatically from the treatment group, especially in terms of age, marital status, ethnicity, and pre-intervention earnings (all the mean differences are statistically significant). In order to bridge the gap between the treatment and the comparison groups in terms of pre-intervention characteristics, Lalonde extracts subsets from PSID-1 and CPS-1 (PSID-2 and -3, and CPS-2 and -3) which resemble the treatment group in terms of single pre-intervention characteristics (such as age or employment status; see Table 1, notes). But as the table indicates, the subsets still remain substantially different from the treatment group (the mean differences in age, ethnicity, marital status, and earnings are smaller, but remain statistically significant).

Table 2 (Panel A) replicates Lalonde's results using his original data and non-experimental comparison groups (the results are identical to those presented in his paper, with the exceptions noted in the footnote of Table 2). Table 2 (Panel B) applies Lalonde's estimators to our reduced experimental sample and the same comparison units. Comparing the two panels, we note that the treatment effect, as estimated from the randomized experiment, is higher in Panel B ($1,794 compared with $886). This is due to the cohort phenomenon --individuals with a later month of assignment seem to have benefitted more from the program. Otherwise, the results are qualitatively similar. The simple difference in means, reported in column (1), yields negative treatment effects for the CPS and PSID comparison groups in both panels (except PSID-3). The

fixed-effect type differencing estimator in column (3) fares somewhat better, although many estimates are still negative or deteriorate when we control for covariates in both panels. The estimates in column (5) are closest to the experimental estimate, consistently closer than those in column (2) which do not control for earnings in 1975. The treatment effect is underestimated by about $1,000 for the CPS comparison groups and $1,500 for the PSID groups. Lalonde's conclusion from Panel A, which also holds for our version in Panel B, is that there is no consistent estimate robust to the specification of the regression or the choice of comparison group.

The inclusion of earnings in 1974 as an additional variable in the regressions in Table 2 (Panel C) does not alter Lalonde's basic message, although the estimates improve when compared with Panel B. In columns (1) to (3), many estimates are still negative, but less so than in Panel B. In columns (4) and (5), the estimates are also closer to the experimental benchmark, off by about $1,000 for PSID1-3 and CPS1-2 and by $400 for CPS-3. Overall, the best results in Table 2 are for CPS-3, Panel C. This raises a number of issues. The strategy of considering subsets of the comparison group more comparable to the treatment group certainly seems to improve matters, provided that we observe the key pre-intervention variables. But Lalonde creates these subsets in an informal manner, based on one or two pre-intervention variables. Table 1 reveals that significant differences remain between the comparison groups and the treatment group. A more systematic means of creating such subsets should improve the estimates from both the CPS and PSID. We undertake this in Sections 3 and 4 with propensity score methods.

3. IDENTIFYING AND ESTIMATING THE AVERAGE TREATMENT EFFECT

3.1 Identification

Let Y_{i1} represent the value of the outcome when unit i is subject to regime 1 (called treatment), and Y_{i0} the value of the outcome when unit i is exposed to regime 0 (called control). Only one of Y_{i0} or Y_{i1} can be observed for any unit, since we can not observe the same unit under both treatment and control. Let T_i be a treatment indicator (=1 if exposed to treatment, =0 otherwise). Then

the observed outcome for unit i is $Y_i = T_i Y_{i1} + (1-T_i)Y_{i0}$. The treatment effect for unit i is $\tau_i = Y_{i1} - Y_{i0}$.

In an observational study, the treatment and comparison groups are often drawn from different populations. In our application the group exposed to the treatment is drawn from the population of interest (welfare recipients eligible for the program). The comparison group is drawn from a different population (in our application both the CPS and PSID are more representative of the general US population). The treatment effect we are trying to identify is therefore the treatment effect for the treated population:

$$\tau|_{T=1} = E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 1).$$

This cannot be estimated directly since Y_{i0} is not observed for the treated units. Assuming selection on observables (Rubin 1974, 1977), namely $\{Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i\}|X_i$ (using Dawid's notation, $\perp\!\!\!\perp$ is independence), we obtain:

$$E(Y_{ij}|X_i, T_i = 1) = E(Y_{ij}|X_i, T_i = 0) = E(Y_i|X_i, T_i = j),$$

for j=0,1. Conditional on the observables, X_i, there is no systematic pre-treatment difference between the groups assigned to treatment and control. This allows us to identify the treatment effect for the treated:

$$\tau|_{T=1} = E\{E(Y_i \mid X_i, T_i = 1) - E(Y_i \mid X_i, T_i = 0) | T_i = 1\}, \quad (1)$$

where the outer expectation is over the distribution of $X_i|T_i$=1, the distribution of pre-intervention variables in the treated population.

One method for estimating the treatment effect stems from (1): estimating $E(Y_i|X_i, T_i = 1)$ and $E(Y_i \mid X_i, T_i = 0)$ as two non-parametric equations. This estimation strategy becomes difficult, however, if the covariates, X_i, are high dimensional. The propensity score theorem provides an intermediate step:

Proposition 1 (Rosenbaum and Rubin 1983)**:** *Let $p(X_i)$ be the probability of unit i having been assigned to treatment, defined as $p(X_i) \equiv \Pr(T_i=1|X_i) = \mathrm{E}(T_i|X_i)$, where $0<p(X_i)<1$, $\forall i$. Then:*

$$\{(Y_{i1}, Y_{i0}) \perp\!\!\!\perp T_i\} | X_i \Rightarrow \{(Y_{i1}, Y_{i0}) \perp\!\!\!\perp T_i\} | p(X_i).$$

Corollary:

$$\tau|_{T=1} = E\{E(Y_i \mid T_i = 1, p(X_i)) - E(Y_i \mid T_i = 0, p(X_i)) \mid T_i = 1\}, \quad (2)$$

where the outer expectation is over the distribution of $p(X_i)|T_i=1$.

One intuition for the propensity score is that, whereas in equation (1) we are trying to condition on X (intuitively, to find observations with similar covariates), in equation (2) we are trying to condition just on the propensity score, because the proposition implies that observations with the same propensity score have the same distribution of the full vector of covariates X.

3.2 The Estimation Strategy

Estimation is in two steps. First, we estimate the propensity score for the sample of experimental treatment and non-experimental comparison units. We use the logistic model, but other standard models yield similar results. An issue is what functional form of the pre-intervention variables to include in the logit. We rely on the following proposition:

Proposition 2 (Rosenbaum and Rubin 1983)**:**

$$X_i \perp\!\!\!\perp T_i \mid p(X_i).$$

Proposition 2 asserts that, conditional on the propensity score, the covariates are independent of assignment to treatment, so that, for observations with the same propensity score, the distribution of covariates should be the same across the treatment and comparison groups. Conditioning on the propensity score, each individual has the same probability of assignment to treatment, as in a randomized experiment.

We use this proposition to assess estimates of the propensity score. For any given specification (we start by introducing the covariates linearly), we group observations into strata defined on the estimated propensity score and check whether we succeed in balancing the covariates within each stratum. We use tests for the statistical significance of differences in the distribution

of covariates, focusing on first and second moments. If there are no significant differences between the two groups, then we accept the specification. If there are significant differences, we add higher-order terms and interactions of the covariates until this condition is satisfied. Section 5 shows that the results are not sensitive to the selection of higher order and interaction variables.

In the second step, given the estimated propensity score, we need to estimate a univariate non-parametric regression $E\left(Y_i | T_i = j, p(X_i)\right)$, for j=0,1. We focus on simple methods for obtaining a flexible functional form, stratification and matching, but in principle one could use any one of the standard array of non-parametric techniques (e.g., see Härdle 1990).

With stratification, observations are sorted from lowest to highest estimated propensity score. The comparison units with an estimated propensity score less than the minimum (or greater than the maximum) estimated propensity score for treated units are discarded. The strata, defined on the estimated propensity score, are chosen so that the covariates within each stratum are balanced across the treatment and comparison units (we know such strata exist from step one). Based on equation (2), within each stratum we take a difference in means of the outcome between the treatment and comparison groups, and weight these by the number of treated observations in each stratum. We also consider matching on the propensity score. Each treatment unit is matched with replacement to the comparison unit with the closest propensity score; the unmatched comparison units are discarded (see Dehejia and Wahba 1997 for more details; also Rubin 1979, and Heckman, Ichimura, and Todd 1997).

There are a number of reasons to prefer this two-step approach rather than estimating equation (1) directly. First, tackling equation (1) directly with a non-parametric regression would encounter the curse of dimensionality as a problem in many data sets, including ours, which have a large number of covariates. This is also true for estimating the propensity score using non-parametric techniques. Hence, we use a parametric model for the propensity score. This is preferable to applying a parametric model to equation (1) directly because, as we will see, the results are less sensitive to the logit specification than regression models, such as those in Table 2 (and because there is a simple criterion for determining which interactions to add to the specification).

Finally, depending on the estimator one adopts (e.g., stratification), an extremely precise estimate of the propensity score is not even needed, since the process of validating the propensity score produces at least one partition structure which balances pre-intervention covariates across the treatment and comparison groups within each stratum, which (by equation (1)) is all that is needed for an unbiased estimate of the treatment impact.

4. RESULTS USING THE PROPENSITY SCORE

Using the method outlined in the previous section, we estimate the propensity score for each comparison group separately. Figure 1 presents a histogram of the estimated propensity scores for the treatment and PSID-1 comparison units, and Figure 2 for CPS-1 comparison units. In Figure 2, we discard 12,611 (out of a total of 15,992) CPS units whose estimated propensity score is less than the minimum for the treatment units. Even then, the first bin (from 0-0.05) contains 2,969 of the remaining comparison units and only 26 treatment units. This provides a snapshot of the fact that the comparison group, although very large, contains relatively few units comparable to the treatment group. A similar pattern is seen in the first bin of Figure 1, but an important difference is that in Figure 1 there is limited overlap in the estimated propensity score between the treatment and PSID groups: there are 98 (more than half the total number of) treated units with an estimated propensity score in excess of 0.8, and only 7 comparison units. Instead, for the CPS, although the treatment units outnumber the comparisons for higher values of the estimated propensity scores, for most bins there are at least a few comparison units.

We use stratification and matching on the propensity score to group the treatment units with the small number of comparison units that are comparable (namely, those comparison units whose estimated propensity scores are greater than the minimum -- or less than the maximum -- propensity score for treatment units). The treatment effect is estimated by summing the within-stratum difference in means between the treatment and comparison observations (of earnings in 1978), where the sum is weighted by the number of treated observations within each stratum (Table 3, column (4)). An alternative is a within-block regression, again taking a weighted sum

over the strata (Table 3, column (5)). When the covariates are well balanced, such a regression should have little effect, but it can help to eliminate the remaining within-block differences. Likewise for matching, we can estimate a simple difference in means between the treatment and matched comparison group for earnings in 1978 (column (7)), and also perform a regression of 1978 earnings on covariates (column (8)).

Table 3 presents the results. For the PSID sample, the stratification estimate is $1,608 and the matching estimate is $1,691, which should be compared against the benchmark randomized-experiment estimate of $1,794. The estimates from a difference in means, or regression control on the full sample, are -$15,205 and $731. The propensity score estimators yield more accurate estimates simply using a difference in means because only those comparison units similar to the treatment group have been used. In columns (5) and (8) controlling for covariates has little impact on the stratification and matching estimates. Likewise for the CPS, the propensity-score-based estimates from the CPS -- $1,713 and $1,582 -- are much closer to the experimental benchmark than estimates from the full comparison sample -- -$8,498 and $972.

Another set of estimates to consider is from the subsets of the PSID and CPS. In Table 2, the estimates tend to improve when applied to narrower subsets. However, as noted above, the estimates still range from -$8,498 to $1,326. In Table 3, the estimates do not improve for the subsets, although the range of fluctuation is much narrower, from $587 to $2,321. Tables 1 and 4 shed light on this.

Table 1 presents the pre-intervention characteristics of the various comparison groups. We note that the subsets PSID-2 and -3, and CPS-2 and -3, though more closely resembling the treatment group, are still considerably different along a number of important dimensions, including ethnicity, marital status, and especially earnings. Table 4 presents the characteristics of the matched subsamples from the comparison groups. The characteristics of the matched subsets of CPS-1 and PSID-1 closely correspond to the treatment group; none of the differences are statistically significant. But as we create the subsets of the comparison groups, the quality of the matches declines, most dramatically for the PSID, with PSID-2 and -3 earnings now increasing

from 1974 to 1975, whereas for the treatment group they decline. The training literature has identified the "dip" in earnings as an important characteristic of participants in training programs (see Ashenfelter 1974, 1978). The CPS sub-samples retain the dip, but for the matched subset of CPS-3 earnings in 1974 are significantly higher than for the treatment group.

This illustrates one of the important features of propensity score methods, namely that the creation of subsamples from the non-experimental comparison group is neither necessary nor desirable, because subsamples created based on single pre-intervention characteristics may dispose of comparison units which nonetheless are good overall comparisons with treatment units. The propensity score sorts out which comparison units are most relevant considering all of the pre-intervention characteristics, not just one characteristic at a time.

Column (3) in Table 3 gives an important insight into how the estimators in columns (4) to (8) succeed in estimating the treatment effect accurately. In column (3) we regress the outcome (earnings in 1978) on a quadratic function of the estimated propensity score and a treatment indicator. The estimates are comparable to those in column (2), where we regress the outcome on all pre-intervention characteristics. This again demonstrates the ability of the propensity score to summarize all pre-intervention variables. The estimators in columns (4) to (8) differ from column (3) in two respects. First, their functional form is more flexible than a low-order polynomial in the estimated propensity score. Second, rather than requiring a constant additive treatment effect, they allow the treatment effect to vary within each stratum (for stratification) or for each individual (for matching).

Finally, it must be noted that even though the estimates presented in Table 3 are closer to the experimental benchmark than those presented in Table 2, with the exception of the adjusted matching estimator, their standard errors are higher: in Table 3, column (5), the standard errors are 1,152 and 1,581 for the CPS and PSID, compared with 550 and 886 in Table 2, column (5). This is because the propensity score estimators use fewer observations. When stratifying on the propensity score, we discard irrelevant controls, and so the strata may contain as few as seven

treated observations. However, the standard errors for the adjusted matching estimator (751 and 809) are similar to those in Table 2.

By summarizing all of the covariates in a single number, the propensity score method allows us to focus on the comparability of the comparison group to the treatment group. Hence, it allows us to address the issues of functional form and treatment effect heterogeneity much more easily.

5. SENSITIVITY ANALYSIS

5.1 Sensitivity to the Specification of the Propensity Score

How sensitive are the estimates presented to the specification of the estimated propensity score? For the stratification estimator, as was suggested in Section 3, the exact specification of the estimated propensity score is not important as long as, within each stratum, the pre-intervention characteristics are balanced across the treatment and comparison groups. Since this was the basis of the specification search suggested in Section 3, either one can find a specification that balances pre-intervention characteristics, or one must conclude the treatment and comparison groups are irreconcilably different.

The upper half of Table 5 demonstrates that the estimates of the treatment impact are not particularly sensitive to the specification used. Specifications 1 and 4 are the same as those in Table 3 (hence, they balance the pre-intervention characteristics). In specifications 2 to 3 and 5 to 6, we drop the squares and cubes of the covariates, and then interactions and dummy variables. In specifications 3 and 6, the logits then simply use the covariates linearly. These estimates are worse than those in Table 3, ranging from $835 to $1,774. But compared with the range of estimates from Table 2, these remain concentrated. Furthermore, we are unable to find a partition structure for the alternative specifications such that the pre-intervention characteristics are balanced within each stratum. There is a well-defined criterion to reject these alternative specifications. Indeed, the specification search begins with a linear specification, and adds higher-order and interaction terms until within-stratum balance is achieved.

5.2 Sensitivity to Selection on Observables

One important assumption underlying propensity score methods is that all of the variables that affect assignment to treatment and are correlated with the potential outcomes, Y_{i1} and Y_{i0}, are observed. This assumption led us to restrict Lalonde's data to the subset for which two (rather than one) years of pre-intervention earnings data is available. In Table 5 (Panel B), we consider how our estimators would fare in the absence of two years of pre-intervention earnings data by re-estimating the treatment impact without making use of earnings in 1974. For PSID-1, the stratification estimators yield less reliable estimates than in Table 3, ranging from -$1,023 to $1,727 as compared with $1,473 to $1,691, although the matching estimator is more robust. In contrast, even though the estimates from the CPS are farther from the experimental benchmark than those in Table 3 ($861 to $1941 compared with $1,582 to $1,774), they are still more concentrated around the experimental estimates than the regression estimates in Panel B of Table 2.

This illustrates that the results are sensitive to the set of pre-intervention variables used. For training programs, a sufficiently lengthy pre-intervention earnings history clearly is important. Table 5 also demonstrates the value of using multiple comparison groups. Even if we did not know the experimental estimate, in looking at Table 5 we would be concerned that the variables that we observe (assuming that earnings in 1974 are not observed) do not control fully for the differences between the treatment and comparison groups, because of variation in the estimates between the CPS and PSID. If all relevant variables are observed, then the estimates from both groups should be similar (as they are in Table 3). When an experimental benchmark is not available, multiple comparison groups are valuable because they can suggest the existence of important unobservables (see Rosenbaum 1987, which develops this idea in more detail).

6. CONCLUSION

This paper demonstrates how to estimate the treatment impact in an observational study using propensity score methods.

These methods are assessed using Lalonde's influential re-creation of a non-experimental setting. Our results show that the estimates of the training effect are close to the benchmark experimental estimate, and are robust to the specification of the comparison group and the functional form used to estimate the propensity score. A researcher using our method would arrive at estimates of the treatment impact ranging from $1,473 to $1,774, very close to the benchmark unbiased estimate from the experiment of $1,794. Furthermore, our methods succeed for a transparent reason: they use only the subset of the comparison group that is comparable to the treatment group, and discard the complement. Although Lalonde attempts to follow this strategy in his construction of other comparison groups, his method relies on an informal selection among the pre-intervention variables. Our application illustrates that even among a large set of potential comparison units, very few may be relevant. But it also illustrates that even a few comparison units can be enough to estimate the treatment impact.

The methods we suggest are not relevant in all situations: there may be important unobservable covariates, for which the propensity score method cannot account. But rather than giving up, or relying on assumptions about the unobserved variables, propensity score methods may offer both a diagnostic on the quality of the comparison group and a means to estimate the treatment impact.

References

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

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

Ashenfelter, O. (1974), "The Effect of Manpower Training on Earnings: Preliminary Results," in *Proceedings of the Twenty-Seventh Annual Winter Meetings of the Industrial Relations Research Association*, (eds.) J. Stern and B. Dennis, Madison: Industrial Relations Research Association.

------ (1978), "Estimating the Effects of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47-57.

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

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

Dehejia, Rajeev H., and Sadek Wahba (1997), "Matching Methods for Estimating Causal Effects in Non-Experimental Studies," University of Toronto, unpublished manuscript.

Härdle, Wolfgang (1990), *Applied Nonparametric Regression*, Econometric Society Monographs, Cambridge: Cambridge University Press.

Heckman, James, and Richard Robb (1985), "Alternative Methods for Evaluating the Impact of Interventions", in *Longitudinal Analysis of Labor Market Data*, Econometric Society Monograph No. 10, (eds.) James Heckman and Burton Singer, Cambridge: Cambridge University Press.

------ and Joseph Hotz (1989), "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862-874.

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

Holland, Paul W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945-960.

Lalonde, Robert (1986), "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604-620.

Manpower Demonstration Research Corporation (1983), *Summary and Findings of the National Supported Work Demonstration*, Cambridge: Ballinger.

Manski, Charles F., and Irwin Garfinkel (1992), "Introduction," in *Evaluating Welfare and Training Programs*, (eds.) Charles Manski and Irwin Garfinkel, Cambridge: Harvard University Press.

------, G. Sandefur, S. McLanahan, and D. Powers (1992), "Alternative Estimates of the Effect of Family Structure During Adolescence on High School Graduation," *Journal of the American Statistical Association*, 87, 25-37.

Rosenbaum, P. (1987), "The Role of a Second Control Group in an Observational Study," *Statistical Science*, 2(3), 292-316

------ and D. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

------ and ------ (1984), "Reducing Bias in Observational Studies Using the Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516-524.

Rubin, Donald (1974), "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

------ (1977), "Assignment to a Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1-26.

------ (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34-58.

------ (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observation Studies," *Journal of the American Statistical Association*, 74, 318-328.

Table 1. Sample Means of Characteristics for NSW and Comparison Samples

Control Sample	No of Obs.	Sample Characteristics: Age	Educ	Black	Hisp	Nodegree	Married	RE74 US$	RE75 US$
NSW/Lalonde:[a]									
Treated	297	24.63	10.38	0.80	0.09	0.73	0.17	--	3,571
Control	425	24.45	10.19	0.80	0.11	0.81	0.16	--	3,672
RE74 subset:[b]									
Treated	185	25.81	10.35	0.84	0.059	0.71	0.19	2,096	1,532
Control	260	25.05	10.09	0.83	0.1	0.83	0.15	2,107	1,267
Comparison groups:[c]									
PSID-1	2,490	34.85	12.11	0.25	0.032	0.31	0.87	19,429	19,063
PSID-2	253	36.10	10.77	0.39	0.067	0.49	0.74	11,027	7,569
PSID-3	128	38.25	10.30	0.45	0.18	0.51	0.70	5,566	2,611
CPS-1	15,992	33.22	12.02	0.07	0.07	0.29	0.71	14,016	13,650
CPS-2	2,369	28.25	11.24	0.11	0.08	0.45	0.46	8,728	7,397
CPS-3	429	28.03	10.23	0.21	0.14	0.60	0.51	5,619	2,467

NOTES:
Data Legend: Age=age in years; Educ=number of years of schooling; Black=1 if black, 0 otherwise; Hisp=1 is Hispanic, 0 otherwise; Nodegree=1 if no high school degree, 0 otherwise; Married=1 if married, 0 otherwise; RE*x*=earnings in calendar year 19*x*; U*x*=1 if unemployed in 19*x*, 0 otherwise.
[a] NSW sample as constructed by Lalonde (1986).
[b] The subset of the Lalonde sample for which RE74 is available.
[c] Definition of Comparison Groups (Lalonde 1986):
PSID-1: All male household heads less than 55 years old who did not classify themselves as retired in 1975.
PSID-2: Selects from PSID-1 all men who were not working when surveyed in the spring of 1976.
PSID-3: Selects from PSID-2 all men who were not working in 1975.
CPS-1: All CPS males less than 55 years of age.
CPS-2: Selects from CPS-1 all males who were not working when surveyed in March 1976.
CPS-3: Selects from CPS-2 all the unemployed males in 1976 whose income in 1975 was below the poverty level.
PSID1-3 and CPS-1 are identical to those used in Lalonde. CPS 2-3 are similar to those used in Lalonde, but Lalonde's original subset could not be re-created.

Table 2. Lalonde's Earnings Comparisons and Estimated Training Effects for the NSW Male Participants Using Comparison Groups from the PSID and the CPS-SSA[a]

Comparison Group	A. Lalonde's original sample					B. RE74 Subsample (results do not use RE74)					C. RE74 Subsample (results use RE74)				
	NSW Treatment Earnings Less Comparison Group Earnings, 1978[b]		Unrestricted Difference in Differences: Quasi-Difference in Earnings Growth: 1975-1978		Controlling for All Variables[f]	NSW Treatment Earnings Less Comparison Group Earnings, 1978[b]		Unrestricted Difference in Differences: Quasi-Difference in Earnings Growth: 1975-1978		Controlling for All Variables[f]	NSW Treatment Earnings Less Comparison Group Earnings, 1978[b]		Unrestricted Difference in Differences: Quasi-Difference in Earnings Growth: 1975-1978		Controlling for All Variables[f]
	Unadjusted	Adjusted[c]	Unadjusted[d]	Adjusted[e]		Unadjusted	Adjusted[c]	Unadjusted[d]	Adjusted[e]		Unadjusted	Adjusted[c]	Unadjusted[d]	Adjusted[e]	
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
NSW	886	798	879	802	820	1,794	1,672	1,750	1,631	1,612	1,794	1,688	1,750	1,672	1,655
	(472)	(472)	(467)	(468)	(468)	(633)	(637)	(632)	(637)	(639)	(633)	(636)	(632)	(638)	(640)
PSID-1	-15,578	-8,067	-2,380	-2,119	-1,844	-15,205	-7,741	-582	-265	186	-15,205	-879	-582	218	731
	(913)	(990)	(680)	(746)	(762)	(1155)	(1175)	(841)	(881)	(901)	(1155)	(931)	(841)	(866)	(886)
PSID-2	-4,020	-3,482	-1,364	-1,694	-1,876	-3,647	-2,810	721	298	111	-3,647	94	721	907	683
	(781)	(935)	(729)	(878)	(885)	(960)	(1082)	(886)	(1004)	(1032)	(960)	(1042)	(886)	(1004)	(1028)
PSID-3	697	-509	629	-552	-576	1,070	35	1,370	243	298	1,070	821	1,370	822	825
	(760)	(967)	(757)	(967)	(968)	(900)	(1101)	(897)	(1101)	(1105)	(900)	(1100)	(897)	(1101)	(1104)
CPS-1	-8,870	-4,416	-1,543	-1,102	-987	-8,498	-4,417	-78	525	709	-8,498	-8	-78	739	972
	(562)	(577)	(426)	(450)	(452)	(712)	(714)	(537)	(557)	(560)	(712)	(572)	(537)	(547)	(550)
CPS-2	-4,195	-2,341	-1,649	-1,129	-1,149	-3,822	-2,208	-263	371	305	-3,822	615	-263	879	790
	(533)	(620)	(459)	(551)	(551)	(671)	(746)	(574)	(662)	(666)	(671)	(672)	(574)	(654)	(658)
CPS-3	-1,008	-1	-1,204	-263	-234	-635	375	-91	844	875	-635	1,270	-91	1,326	1,326
	(539)	(681)	(532)	(677)	(675)	(657)	(821)	(641)	(808)	(810)	(657)	(798)	(641)	(796)	(798)

NOTES:

Panel A replicates Lalonde (1986), Table 5. The estimates for columns (1) to (4) for NSW, PSID1-3, and CPS-1 are identical to Lalonde's. CPS 2-3 are similar, but not identical, because we could not exactly re-create his subset. Column (5) differs because the data file we obtained did not contain all of the covariates used in column (10) of Lalonde (1986), Table 5.

[a] Estimated effect of training on RE78. Standard errors are in parentheses. The estimates are in 1982 dollars.

[b] Based on the experimental data, an unbiased estimate of the impact of training is presented in column (4), $1,794.

[c] The exogenous variables used in the regressions-adjusted equations are age, age squared, years of schooling, high school dropout status, and race (and RE74 in Panel C).

[d] Compares RE78 across the treatment and comparison group, controlling for RE75.

[e] The same as (d), but controls for the additional variables listed under (c).

[f] Controls for all pre-treatment covariates.

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups from PSID and CPS-SSA

	NSW Earnings Less Comparison Group Earnings		NSW Treatment Earnings Less Comparison Group Earnings, Conditional On The Estimated Propensity Score					
			Quadratic in Score[b]	Stratifying on the Score			Matching on the Score	
	(1) Unadjusted	(2) Adjusted[a]	(3)	(4) Un-adjusted	(5) Adjusted[a]	(6) Obs.[g]	(7) Un-adjusted	(8) Adjusted[f]
NSW	1,794 (633)	1,672 (638)						
PSID-1[c]	-15,205 (1154)	731 (886)	294 (1389)	1,608 (1571)	1,494 (1581)	1,255	1,691 (2209)	1,473 (809)
PSID-2[d]	-3,647 (959)	683 (1028)	496 (1193)	2,220 (1768)	2,235 (1793)	389	1,455 (2303)	1,480 (808)
PSID-3[d]	1,069 (899)	825 (1104)	647 (1383)	2,321 (1994)	1,870 (2002)	247	2,120 (2335)	1,549 (826)
CPS-1[e]	-8,498 (712)	972 (550)	1117 (747)	1,713 (1115)	1,774 (1152)	4,117	1,582 (1069)	1,616 (751)
CPS-2[e]	-3,822 (670)	790 (658)	505 (847)	1,543 (1461)	1,622 (1346)	1493	1,788 (1205)	1,563 (753)
CPS-3[e]	-635 (657)	1,326 (798)	556 (951)	1,252 (1617)	2,219 (2082)	514	587 (1496)	662 (776)

NOTES:

[a] Least Squares Regression: RE78 on a constant, expstat, age, age^2, educ, nodegree, black, hisp, RE74, RE75.

[b] Least squares regression of RE78 on a quadratic on the estimated propensity score and a treatment indicator, for observations used under stratification; see note (g).

[c] Logit: Prob (expstat=1)=F(age, age^2, educ, educ^2, married, nodegree, black, hisp, RE74, RE75, RE 74^2, RE75^2, u74*black) (Expstat=1 if the unit was subject to treatment, =0 otherwise).

[d] Logit: Prob (expstat=1)=F(age, age^2, educ, educ^2, nodegree, married, black, hisp, RE74, RE 74^2, RE75, RE75^2, u74, u75)

[e] Logit: Prob (expstat=1)=F(age, age^2, educ, educ^2, nodegree, married, black, hisp, RE74, RE75, u74, u75, educ*RE74,age^3)

[f] Weighted Least Squares: treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation (same covariates as (a)).

[g] Number of observations refers to the actual number of comparison and treatment units used for (3) to (5), namely, all treatment units and those comparison units whose estimated propensity score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.

Table 4. Sample Means of Characteristics for Matched Control Samples

		Sample Characteristics							
Matched Samples	No. of Obs.	Age	School	Black	Hisp	Nodegree	Married	RE74 US$	RE75 US$
NSW	**185**	**25.81**	**10.35**	**0.84**	**0.06**	**0.71**	**0.19**	**2,096**	**1,532**
MPSID-1	56	26.39	10.62	0.86	0.02	0.55	0.15	1,794	1,126
MPSID-2	49	24.32	11.10	0.89	0.02	0.57	0.19	1,599	2,225
MPSID-3	30	26.86	10.96	0.91	0.01	0.52	0.25	1,386	1,863
MCPS-1	119	26.91	10.52	0.86	0.04	0.64	0.19	2,110	1,396
MCPS-2	87	26.21	10.21	0.85	0.04	0.68	0.20	1,758	1,204
MCPS-3	63	25.94	10.69	0.87	0.06	0.53	0.13	2,709	1,587

NOTES:
MPSID 1-3 and MCPS 1-3 are the subsamples of PSID 1-3 and CPS 1-3 that are matched to the treatment group.

Table 5. Sensitivity of Estimated Training Effects to Specification of the Propensity Score

Comparison Group	NSW Earnings Less Comparison Group Earnings		NSW Treatment Earnings Less Comparison Group Earnings Conditional on the Estimated Propensity Score					
			Quadratic in Score[c]	Stratifying on the Score			Matching on the Score	
	(1) Unadjusted	(2) Adjusted[a]	(3)	(4) Unadjusted	(5) Adjusted[a]	(6) Obs.[d]	(7) Unadjusted	(8) Adjusted[b]
NSW	1,794	1,672						
	(633)	(638)						
A. Dropping higher-order terms								
PSID-1:	-15,205	218	294	1,608	1,254	1,255	1,691	1,054
Spec. 1	(1154)	(866)	(1389)	(1571)	(1616)		(2209)	(831)
PSID-1:	-15,205	105	539	1,524	1,775	1,533	2,281	2,291
Spec. 2	(1154)	(863)	(1344)	(1527)	(1538)		(1732)	(796)
PSID-1:	-15,205	105	1,185	1,237	1,155	1,373	1140	855
Spec. 3	(1154)	(863)	(1233)	(1144)	(1280)		(1720)	(906)
CPS-1:	-8,498	738	1,117	1,713	1,774	4,117	1,582	1,616
Spec. 4	(712)	(547)	(747)	(1115)	(1152)		(1069)	(751)
CPS-1:	-8,498	684	1,248	1,452	1,454	6,365	835	904
Spec. 5	(712)	(546)	(731)	(632)	(2713)		(1007)	(769)
CPS-1:	-8,498	684	1,241	1,299	1,095	6,017	1,103	1,471
Spec. 6	(712)	(546)	(671)	(547)	(925)		(877)	(787)
B. Dropping RE74								
PSID-1:	-15,205	-265	-697	-869	-1,023	1,284	1,727	1,340
Spec. 7	(1154)	(880)	(1279)	(1410)	(1493)		(1447)	(845)
PSID-2:	-3,647	297	521	405	304	356	530	276
Spec. 8	(959)	(1004)	(1154)	(1472)	(1495)		(1848)	(902)
PSID-3:	1,069	243	1,195	482	-53	248	87	11
Spec. 8	(899)	(1100)	(1261)	(1449)	(1493)		(1508)	(938)
CPS-1:	-8,498	525	1,181	1,234	1,347	4,558	1,402	861
Spec. 9	(712)	(557)	(698)	(695)	(683)		(1067)	(786)
CPS-2:	-3,822	371	482	1,473	1,588	1,222	1,941	1,668
Spec. 9	(670)	(662)	(731)	(1313)	(1309)		(1500)	(755)
CPS-3:	-635	844	722	1,348	1,262	504	1,097	1,120
Spec. 9	(657)	(807)	(942)	(1601)	(1600)		(1366)	(783)

NOTES:
Spec. 1: Same as Table 3, note c. Spec. 2: Spec. 1 without higher powers. Spec. 3: Spec. 2 without higher-order terms. Spec. 4: Same as Table 3, note e. Spec. 5: Spec. 4 without higher powers. Spec. 6: Spec. 5 without higher-order terms. Spec. 7: Same as Table 3, note c, with RE74 removed. Spec. 8: Same as Table 3, note d, with RE74 removed. Spec. 9: Same as Table 3, note e, with RE74 removed.

[a] Least Squares Regression: RE78 on a constant, expstat, age, educ, nodegree, black, hisp, RE74, RE75.
[b] Weighted Least Squares: treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation (same covariates as (a)).
[c] Least squares regression of RE78 on a quadratic on the estimated propensity score and a treatment indicator, for observations used under stratification; see note (d).
[d] Number of observations refers to the actual number of comparison and treatment units used for (3) to (5), namely, all treatment units and those comparison units whose estimated propensity score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.

Figure 1: Histogram of Estimated Propensity Score, NSW and PSID

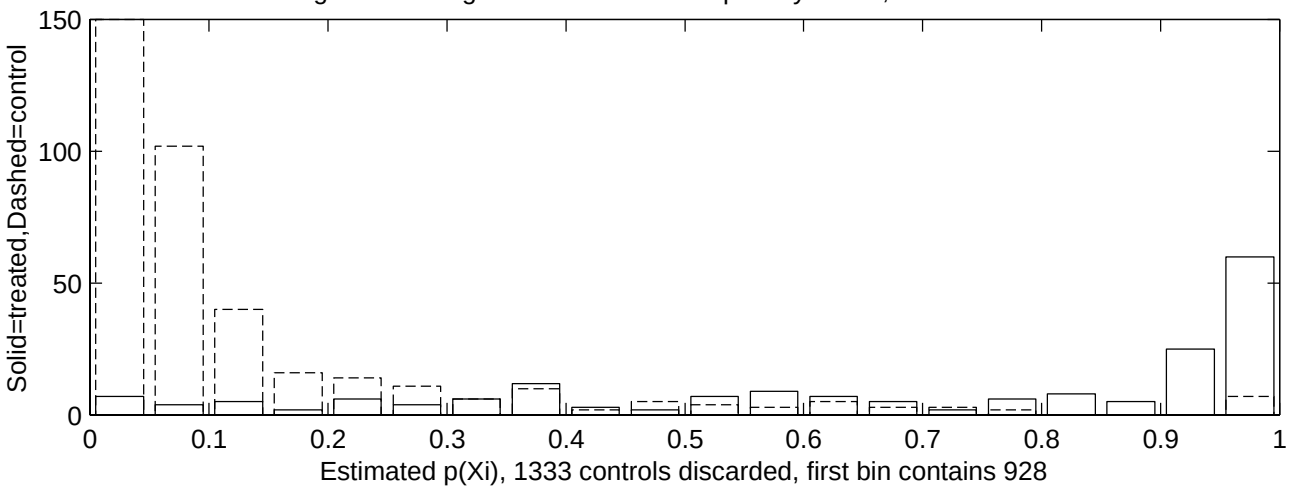

Figure 2: Histogram of Estimated Propensity Score, NSW and CPS