Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs

Rajeev H. Dehejia; Sadek Wahba

Journal of the American Statistical Association, Vol. 94, No. 448 (Dec., 1999), 1053-1062.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199912%2994%3A448%3C1053%3ACEINSR%3E2.0.CO%3B2-K

Journal of the American Statistical Association is currently published by American Statistical Association.

Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs

Rajeev H. DEHEJIA and Sadek WAHBA

This article uses propensity score methods to estimate the treatment impact of the National Supported Work (NSW) Demonstration, a labor training program, on postintervention earnings. We use data from Lalonde's evaluation of nonexperimental methods that combine the treated units from a randomized evaluation of the NSW with nonexperimental comparison units drawn from survey datasets. We apply propensity score methods to this composite dataset and demonstrate that, relative to the estimators that Lalonde evaluates, propensity score estimates of the treatment impact are much closer to the experimental benchmark estimate. Propensity score methods assume that the variables associated with assignment to treatment are observed (referred to as ignorable treatment assignment, or selection on observables). Even under this assumption, it is difficult to control for differences between the treatment and comparison groups when they are dissimilar and when there are many preintervention variables. The estimated propensity score (the probability of assignment to treatment, conditional on preintervention variables) summarizes the preintervention variables. This offers a diagnostic on the comparability of the treatment and comparison groups, because one has only to compare the estimated propensity score across the two groups. We discuss several methods (such as stratification and matching) that use the propensity score to estimate the treatment impact. When the range of estimated propensity scores of the treatment and comparison groups overlap, these methods can estimate the treatment impact for the treatment group. A sensitivity analysis shows that our estimates are not sensitive to the specification of the estimated propensity score, but are sensitive to the assumption of selection on observables. We conclude that when the treatment and comparison groups overlap, and when the variables determining assignment to treatment are observed, these methods provide a means to estimate the treatment impact. Even though propensity score methods are not always applicable, they offer a diagnostic on the quality of nonexperimental comparison groups in terms of observable preintervention variables.

KEY WORDS: Matching; Program evaluation; Propensity score.

1. INTRODUCTION

This article discusses the estimation of treatment effects in observational studies. This issue has been the focus of much attention because randomized experiments cannot always be implemented and has been addressed inter alia by Lalonde (1986), whose data we use herein. Lalonde estimated the impact of the National Supported Work (NSW) Demonstration, a labor training program, on postintervention income levels. He used data from a randomized evaluation of the program and examined the extent to which nonexperimental estimators can replicate the unbiased experimental estimate of the treatment impact when applied to a composite dataset of experimental treatment units and nonexperimental comparison units. He concluded that standard nonexperimental estimators such as regression, fixed-effects, 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 nonexperimental evaluations (see Manski and Garfinkel 1992) and in spurring a search for alternative estimators and specification tests (see, e.g., Heckman and Hotz 1989; Manski, Sandefur, McLanahan, and Powers 1992).

In this article we apply propensity score methods (Rosenbaum and Rubin 1983) to Lalonde's dataset. The propensity score is defined as the probability of assignment to treatment, conditional on covariates. Propensity score methods focus on the comparability of the treatment and nonexperimental comparison groups in terms of preintervention variables. Controlling for differences in preintervention variables is difficult when the treatment and comparison groups are dissimilar and when there are many preintervention variables. The estimated propensity score, a single variable on the unit interval that summarizes the preintervention variables, can control for differences between the treatment and nonexperimental comparison groups. When we apply these methods to Lalonde's nonexperimental data for a range of propensity score specifications and estimators, we obtain estimates of the treatment impact that are much closer to the experimental treatment effect than Lalonde's nonexperimental estimates.

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

Rajeev Dehejia is Assistant Professor, Department of Economics and SIPA, Columbia University, New York, NY 10027 (E-mail: *dehejia@columbia.edu*). Sadek Wahba is Vice President, Morgan Stanley & Co. Incorporated, New York, NY 10036 (E-mail: *wahbas@ms.com*). 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 an associate editor, two anonymous referees, Gary Chamberlain, Guido Imbens, and Donald Rubin, whose detailed comments and suggestions greatly improved the article. They thank Robert Lalonde for providing the data from his 1986 study and substantial help in recreating the original dataset, and also thank Joshua Angrist, George Cave, David Cutler, Lawrence Katz, Caroline Minter-Hoxby, and Jeffrey Smith.

Journal of the American Statistical Association
December 1999, Vol. 94, No. 448, Applications and Case Studies

The article is organized as follows. Section 2 reviews Lalonde's data and reproduces his results. Section 3 identifies the treatment effect under the potential outcomes causal model and discusses estimation strategies for the treatment effect. Section 4 applies our methods to Lalonde's dataset, and Section 5 discusses the sensitivity of the results to the methodology. Section 6 concludes the article.

2. LALONDE'S RESULTS

2.1 The Data

The NSW Demonstration [Manpower Demonstration Research Corporation (MDRC) 1983] was a federally and privately funded program implemented in the mid-1970s to provide work experience for a period of 6–18 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, such as restaurant and construction work. Information on preintervention variables (preintervention earnings as well as education, age, ethnicity, and marital status) was obtained from initial surveys and Social Security Administration records. Both the treatment and control groups participated in follow-up interviews at specific intervals. Lalonde (1986) offered a separate analysis of the male and female participants. In this article we focus on the male participants, as estimates for this group were the most sensitive to functional-form specification, as indicated by Lalonde.

Candidates eligible for the NSW were randomized into the program between March 1975 and July 1977. One consequence of randomization over a 2-year period was that individuals who joined early in the program had different characteristics than those who entered later; this is referred to as the "cohort phenomenon" (MDRC 1983, p. 48). Another consequence is that data from the NSW are delineated in terms of experimental time. Lalonde annualized earnings data from the experiment because the nonexperimental comparison groups that he used (discussed later) are delineated in calendar time. By limiting himself to those assigned to treatment after December 1975, Lalonde ensured that retrospective earnings information from the experiment included calendar 1975 earnings, which he then used as preintervention earnings. By likewise limiting himself to those who were no longer participating in the program by January 1978, he ensured that the postintervention data included calendar 1978 earnings, which he took to be the outcome of interest. Earnings data for both these years are available for both nonexperimental comparison groups. This reduces the NSW sample to 297 treated observations and 425 control observations for male participants.

However, it is important to look at several years of preintervention earnings in determining the effect of job training programs (Angrist 1990, 1998; Ashenfelter 1978; Ashenfelter and Card 1985; Card and Sullivan 1988). Thus we further limit ourselves to the subset of Lalonde's NSW data for which 1974 earnings can be obtained: those individuals who joined the program early enough for the retrospective earnings information to include 1974, as well as those individuals who joined later but were known to have been unemployed prior to randomization. Selection of this subset is based only on preintervention variables (month of assignment and employment history). Assuming that the initial randomization was independent of preintervention covariates, the subset retains a key property of the full experimental data: The treatment and control groups have the same distribution of preintervention variables, although this distribution could differ from the distribution of covariates for the larger sample. A difference in means remains an unbiased estimator of the average treatment impact for the reduced sample. The subset includes 185 treated and 260 control observations.

We present the preintervention characteristics of the original sample and of our subset in the first four rows of Table 1. Our subset differs from Lalonde's original sample, especially in terms of 1975 earnings; this is a consequence both of the cohort phenomenon and of the fact that our subsample contains more individuals who were unemployed prior to program participation. The distribution of preintervention variables is very similar across the treatment and control groups for each sample; none of the differences is significantly different from 0 at a 5% level of significance, with the exception of the indicator for "no degree".

Lalonde's nonexperimental estimates of the treatment effect are based on two distinct comparison groups: the Panel Study of Income Dynamics (PSID-1) and Westat's Matched Current Population Survey-Social Security Administration File (CPS-1). Table 1 presents the preintervention characteristics of the comparison groups. It is evident that both PSID-1 and CPS-1 differ dramatically from the treatment group in terms of age, marital status, ethnicity, and preintervention earnings; all of the mean differences are significantly different from 0 well beyond a 1% level of significance, except the indicator for "Hispanic". To bridge the gap between the treatment and comparison groups in terms of preintervention characteristics, Lalonde extracted subsets from PSID-1 and CPS-1 (denoted PSID-2 and -3 and CPS-2 and -3) that resemble the treatment group in terms of single preintervention characteristics (such as age or employment status; see Table 1). Table 1 reveals that the subsets remain statistically substantially different from the treatment group; the mean differences in age, ethnicity, marital status, and earnings are smaller but remain statistically significant at a 1% level.

2.2 Lalonde's Results

Because our analysis in Section 4 uses a subset of Lalonde's original data and an additional variable (1974 earnings), in Table 2 we reproduce Lalonde's results using his original data and variables (Table 2, panel A), and then apply the same estimators to our subset of his data both without and with the additional variable (Table 2, panels B and C). We show that when his analysis is applied to the data and variables that we use, his basic conclusions remain unchanged. In Section 5 we discuss the sensitivity of our propensity score results to dropping the additional earnings data. In his article, Lalonde considered linear regression, fixed-effects, and latent variable selection models

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

	No. of observations	*Age*	*Education*	*Black*	*Hispanic*	*No degree*	*Married*	*RE74 (U.S. $)*	*RE75 (U.S. $)*
NSW/Lalonde:[a]									
Treated	297	24.63	10.38	.80	.09	.73	.17		3,066
		(.32)	(.09)	(.02)	(.01)	(.02)	(.02)		(236)
Control	425	24.45	10.19	.80	.11	.81	.16		3,026
		(.32)	(.08)	(.02)	(.02)	(.02)	(.02)		(252)
RE74 subset:[b]									
Treated	185	25.81	10.35	.84	.059	.71	.19	2,096	1,532
		(.35)	(.10)	(.02)	(.01)	(.02)	(.02)	(237)	(156)
Control	260	25.05	10.09	.83	.1	.83	.15	2,107	1,267
		(.34)	(.08)	(.02)	(.02)	(.02)	(.02)	(276)	(151)
Comparison groups:[c]									
PSID-1	2,490	34.85	12.11	.25	.032	.31	.87	19,429	19,063
		[.78]	[.23]	[.03]	[.01]	[.04]	[.03]	[991]	[1,002]
PSID-2	253	36.10	10.77	.39	.067	.49	.74	11,027	7,569
		[1.00]	[.27]	[.04]	[.02]	[.05]	[.04]	[853]	[695]
PSID-3	128	38.25	10.30	.45	.18	.51	.70	5,566	2,611
		[1.17]	[.29]	[.05]	[.03]	[.05]	[.05]	(686)	[499]
CPS-1	15,992	33.22	12.02	.07	.07	.29	.71	14,016	13,650
		[.81]	[.21]	[.02]	[.02]	[.03]	[.03]	[705]	[682]
CPS-2	2,369	28.25	11.24	.11	.08	.45	.46	8,728	7,397
		[.87]	[.19]	[.02]	[.02]	[.04]	[.04]	[667]	[600]
CPS-3	429	28.03	10.23	.21	.14	.60	.51	5,619	2,467
		[.87]	[.23]	[.03]	[.03]	[.04]	[.04]	[552]	[288]

NOTE: Standard errors are in parentheses. Standard error on difference in means with RE74 subset/treated is given in brackets. Age = age in years; Education = number of years of schooling; Black = 1 if black, 0 otherwise; Hispanic = 1 if Hispanic, 0 otherwise; No degree = 1 if no high school degree, 0 otherwise; Married = 1 if married, 0 otherwise; REx = earnings in calendar year 19x.

[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 under age 55 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 under age 55.

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 by Lalonde. CPS2–3 are similar to those used by Lalonde, but Lalonde's original subset could not be recreated.

of the treatment impact. Because our analysis focuses on the importance of preintervention variables, we focus on the first of these.

Table 2, panel A, reproduces the results of Lalonde (1986, Table 5). Comparing Panels A and B, we note that the treatment effect, as estimated from the randomized experiment, is higher in the latter ($1,794 compared to $886). This reflects differences in the composition of the two samples, as discussed in the previous section: A higher treatment effect is obtained for those who joined the program earlier or who were unemployed prior to program participation. The results in terms of the success of nonexperimental estimates are qualitatively similar across the two samples. The simple difference in means, reported in column (1), yields negative treatment effects for the CPS and PSID comparison groups in both samples (except PSID-3). The fixed-effects-type differencing estimator in the third column fares somewhat better, although many estimates are still negative or deteriorate when we control for covariates in both panels. The estimates in the fifth column are closest to the experimental estimate, consistently closer than those in the second column, which do not control for earnings in 1975. The treatment effect is underestimated by about $1,000 for the CPS comparison groups and by $1,500 for the PSID groups. Lalonde's conclusion from panel A, which also holds in panel B, is that the regression specifications and comparison groups fail to replicate the treatment impact.

Including 1974 earnings as an additional variable in the regressions in Table 2, panel C does not alter Lalonde's basic message, although the estimates improve compared to those in panel B. In columns (1) and (3), many estimates remain negative, but less so than in panel B. In column (2) the estimates for PSID-1 and CPS-1 are negative, but the estimates for the subsets improve. In columns (4) and (5) the estimates are closer to the experimental benchmark than in panel B, off by about $1,000 for PSID1–3 and CPS1–2 and by $400 for CPS-3. Overall, the results closest to the experimental benchmark in Table 2 are for CPS-3, panel C. This raises a number of issues. The strategy of considering subsets of the comparison group improves estimates of the treatment effect relative to the benchmark. However, Table 1 reveals that significant differences remain between the comparison groups and the treatment group. These subsets are created based on one or two preintervention variables. In Sections 3 and 4 we show that propensity score methods provide a systematic means of creating such subsets.

3. IDENTIFYING AND ESTIMATING THE AVERAGE TREATMENT EFFECT

3.1 Identification

Let Y_{i1} represent the value of the outcome when unit i is exposed to regime 1 (called treatment), and let Y_{i0} represent

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

	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		Unrestricted differences in differences: Quasi-difference in earnings growth 1975−1978		Controlling for all variables[f]	NSW treatment earnings less comparison group earnings 1978		Unrestricted differences in differences: Quasi-difference in earnings growth 1975−1978		Controlling for all variables[f]	NSW treatment earnings less comparison group earnings 1978		Unrestricted differences in differences: Quasi-difference in earnings growth 1975−1978		Controlling for all variables[f]
Comparison group	Unadjusted[b] (1)	Adjusted[c] (2)	Unadjusted[d] (3)	Adjusted[e] (4)	(5)	Unadjusted[b] (1)	Adjusted[c] (2)	Unadjusted[d] (3)	Adjusted[e] (4)	(5)	Unadjusted[b] (1)	Adjusted[c] (2)	Unadjusted[d] (3)	Adjusted[e] (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 the sample of Lalonde (1986, table 5). The estimates for columns (1)–(4) for NSW, PSID1–3, and CPS-1 are identical to Lalonde's. CPS-2 and CPS-3 are similar but not identical, because we could not exactly recreate his subset. Column (5) differs because the data file that we obtained did not contain all of the covariates used in column (10) of Lalonde's Table 5.

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

[b] The estimates based on the NSW control groups are unbiased estimates of the treatment impacts for the original sample ($886) and for the RE74 sample ($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] Regresses RE78 on a treatment indicator and RE75.

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

[f] Controls for all pretreatment covariates.

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, because one cannot 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 experimental setting where assignment to treatment is randomized, the treatment and control groups are drawn from the same population. The average treatment effect for this population is $\tau = E(Y_{i1}) - E(Y_{i0})$. But randomization implies that $\{Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i\}$ [using Dawid's (1979) notation, $\perp\!\!\!\perp$ represents independence], so that for $j = 0, 1$,

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

and

$$\begin{aligned}\tau &= E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0)\\ &= E(Y_i|T_i = 1) - E(Y_i|T_i = 0),\end{aligned}$$

which is readily estimated.

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

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

This expression cannot be estimated directly, because Y_{i0} is not observed for treated units. Assuming selection on observable covariates, $\mathbf{X}_i$—namely, $\{Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i\}|\mathbf{X}_i$ (Rubin 1974, 1977)—we obtain

$$E(Y_{ij}|\mathbf{X}_i, T_i = 1) = E(Y_{ij}|\mathbf{X}_i, T_i = 0) = E(Y_i|\mathbf{X}_i, T_i = j)$$

for $j = 0, 1$. Conditional on the observables, $\mathbf{X}_i$, there is no systematic pretreatment difference between the groups assigned to treatment and control. This allows us to identify the treatment effect for the treated,

$$\begin{aligned}&\tau|_{T=1}\\ &\quad = E\{E(Y_i|\mathbf{X}_i, T_i = 1) - E(Y_i|\mathbf{X}_i, T_i = 0)|T_i = 1\}, \quad (1)\end{aligned}$$

where the outer expectation is over the distribution of $\mathbf{X}_i|T_i = 1$, the distribution of preintervention variables in the treated population.

In our application we have both an experimental control group and a nonexperimental comparison group. Because the former is drawn from the population of interest along with the treated group, we economize on notation and use $T_i = 1$ to represent the entire group of interest and use $T_i = 0$ to represent the nonexperimental group. Thus in (1) the expectation is over the distribution of X_i for the NSW population.

One method for estimating the treatment effect that stems from (1) is estimating $E(Y_i|\mathbf{X}_i, T_i = 1)$ and $E(Y_i|\mathbf{X}_i, T_i = 0)$ as two nonparametric equations. This estimation strategy becomes difficult, however, if the covariates, $\mathbf{X}_i$, are high dimensional. The propensity score theorem provides an intermediate step.

Proposition 1 (Rosenbaum and Rubin 1983). Let $p(\mathbf{X}_i)$ be the probability of unit i having been assigned to treatment, defined as $p(\mathbf{X}_i) \equiv \Pr(T_i = 1|\mathbf{X}_i) = E(T_i|\mathbf{X}_i)$. Assume that $0 < p(\mathbf{X}_i) < 1$, for all $\mathbf{X}_i$, and $\Pr(T_1, T_2, \ldots T_N|\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_N) = \prod_{i=1,\ldots,N} p(\mathbf{X}_i)^{T_i}(1 - p(\mathbf{X}_i))^{(1-T_i)}$ for the N units in the sample. Then

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

Corollary. If $\{(Y_{i1}, Y_{i0}) \perp\!\!\!\perp T_i\}|\mathbf{X}_i$ and the assumptions of Proposition 1 hold, then

$$\begin{aligned}\tau|_{T=1} = E\{&E(Y_i|T_i = 1, p(\mathbf{X}_i))\\ &- E(Y_i|T_i = 0, p(\mathbf{X}_i))|T_i = 1\}, \quad (2)\end{aligned}$$

assuming that the expectations are defined. The outer expectation is over the distribution of $p(\mathbf{X}_i)|T_i = 1$.

One intuition for the propensity score is that whereas in (1) we are trying to condition on $\mathbf{X}_i$ (intuitively, to find observations with similar covariates), in (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, $\mathbf{X}_i$.

3.2 The Estimation Strategy

Estimation is done in two steps. First, we estimate the propensity score separately for each nonexperimental sample consisting of the experimental treatment units and the specified set of comparison units (PSID1–3 or CPS1–3). We use a logistic probability model, but other standard models yield similar results. One issue is what functional form of the preintervention variables to include in the logit. We rely on the following proposition.

Proposition 2 (Rosenbaum and Rubin 1983). If $p(\mathbf{X}_i)$ is the propensity score, then

$$\mathbf{X}_i \perp\!\!\!\perp T_i|p(\mathbf{X}_i).$$

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 (see Rosenbaum and Rubin 1984). If there are no significant differences between the two groups within each stratum, then we accept the specification. If there are significant differences, then we add higher-order terms and interactions of the covariates until this condition is satisfied. In Section 5 we demonstrate 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 nonparametric regression, $E(Y_i|T_i = j, p(\mathbf{X}_i))$, 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 of the standard array of nonparametric techniques (see, e.g., Härdle and Linton 1994; Heckman, Ichimura, and Todd 1997).

With stratification, observations are sorted from lowest to highest estimated propensity score. We discard the comparison units with an estimated propensity score less than the minimum (or greater than the maximum) estimated propensity score for treated units. 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 that such strata exist from step 1.) Based on (2), within each stratum we take a difference in means of the outcome between the treatment and comparison groups, then 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 1998 for more details; also Heckman, Ichimura, Smith, and Todd 1998; Heckman, Ichimura, and Todd 1998; Rubin 1979).

There are a number of reasons for preferring this two-step approach to direct estimation of (1). First, tackling (1) directly with a nonparametric regression would encounter the curse of dimensionality as a problem in many datasets such as ours that have a large number of covariates. This would also occur when estimating the propensity score using nonparametric techniques. Hence we use a parametric model for the propensity score. This is preferable to applying a parametric model directly to (1) because, as we will see, the results are less sensitive to the logit specification than regression models, such as those in Table 2. Finally, depending on the estimator that one adopts (e.g., stratification), a precise estimate of the propensity score is not required. The process of validating the propensity score estimate produces at least one partition structure that balances preintervention covariates across the treatment and comparison groups within each stratum, which, by (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 separately estimate the propensity score for each sample of comparison units and treatment units. Figures 1 and 2 present histograms of the estimated propensity scores for the treatment and PSID-1 and CPS-1 comparison groups. Most of the comparison units (1,333 of a total of 2,490 PSID-1 units and 12,611 of 15,992 CPS-1 units) are discarded because their estimated propensity scores are less than the minimum for the treatment units. Even then, the first bin (units with an estimated propensity score of 0–.05) contains most of the remaining comparison units and few treatment units. An important difference between the figures is that Figure 1 has many bins in which the treatment units greatly outnumber the comparison units. (Indeed, for three bins there are no comparison units.) In contrast, in Figure 2 for CPS-1, each bin contains at least a few comparison units. Overall, for PSID-1 there are 98 (more than half the total number) treated units with an estimated propensity score in excess of .8, and only 7 comparison units, compared to 35 treated and 7 comparison units for CPS-1.

Figures 1 and 2 illustrate the diagnostic value of the propensity score. They reveal that although the comparison groups are large relative to the treatment group, there is limited overlap in terms of preintervention characteristics. Had there been no comparison units overlapping with a broad range of the treatment units, then it would not have been possible to estimate the average treatment effect on the

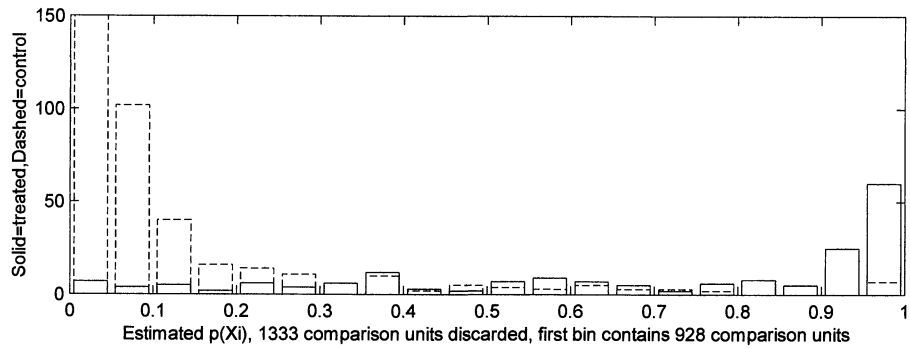

Figure 1. Histogram of the Estimated Propensity Score for NSW Treated Units and PSID Comparison Units. The 1,333 PSID units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 928 PSID units. There is minimal overlap between the two groups. Three bins (.8–.85, .85–.9, and .9–.95) contain no comparison units. There are 97 treated units with an estimated propensity score greater than .8 and only 7 comparison units.

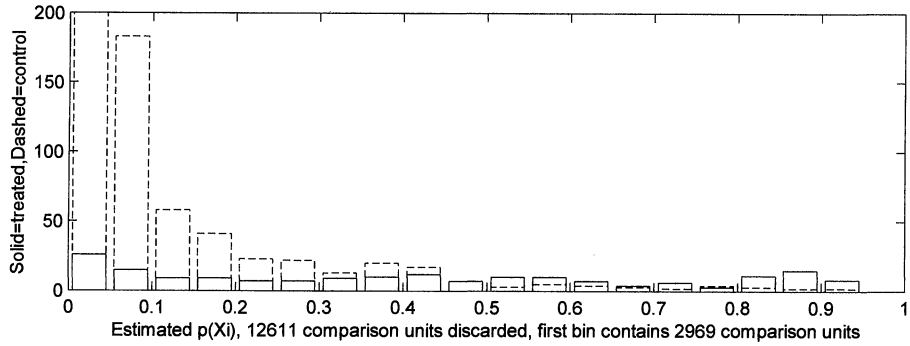

Figure 2. Histogram of the Estimated Propensity Score for NSW Treated Units and CPS Comparison Units. The 12,611 CPS units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 2,969 CPS units. There is minimal overlap between the two groups, but the overlap is greater than in Figure 1; only one bin (.45–.5) contains no comparison units, and there are 35 treated and 7 comparison units with an estimated propensity score greater than .8.

treatment group (although the treatment impact still could be estimated in the range of overlap). With limited overlap, we can proceed cautiously with estimation. Because in our application we have the benchmark experimental estimate, we are able to evaluate the accuracy of the estimates. Even in the absence of an experimental estimate, we show in Section 5 that the use of multiple comparison groups provides another means of evaluating the estimates.

We use stratification and matching on the propensity score to group the treatment units with the small number of comparison units whose estimated propensity scores are greater than the minimum—or less than the maximum—propensity score for treatment units. We estimate the treatment effect 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 eliminate the remaining within-block differences. Likewise for matching, we can estimate a difference in means between the treatment and matched comparison groups 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, compared to the benchmark randomized-experiment estimate of \$1,794. The estimates from a difference in means and regression on the full sample are −\$15,205 and \$731. In columns (5) and (8), controlling for covariates has little impact on the stratification and

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

	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) Unadjusted*	*(5) Adjusted*	*(6) Observations*[c]	*(7) Unadjusted*	*(8) Adjusted*[d]
NSW	1,794	1,672						
	(633)	(638)						
PSID-1[e]	−15,205	731	294	1,608	1,494	1,255	1,691	1,473
	(1,154)	(886)	(1,389)	(1,571)	(1,581)		(2,209)	(809)
PSID-2[f]	−3,647	683	496	2,220	2,235	389	1,455	1,480
	(959)	(1,028)	(1,193)	(1,768)	(1,793)		(2,303)	(808)
PSID-3[f]	1,069	825	647	2,321	1,870	247	2,120	1,549
	(899)	(1,104)	(1,383)	(1,994)	(2,002)		(2,335)	(826)
CPS-1[g]	−8,498	972	1,117	1,713	1,774	4,117	1,582	1,616
	(712)	(550)	(747)	(1,115)	(1,152)		(1,069)	(751)
CPS-2[g]	−3,822	790	505	1,543	1,622	1,493	1,788	1,563
	(670)	(658)	(847)	(1,461)	(1,346)		(1,205)	(753)
CPS-3[g]	−635	1,326	556	1,252	2,219	514	587	662
	(657)	(798)	(951)	(1,617)	(2,082)		(1,496)	(776)

[a] Least squares regression: RE78 on a constant, a treatment indicator, age, age^2, education, no degree, black, Hispanic, 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] Number of observations refers to the actual number of comparison and treatment units used for (3)–(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.

[d] 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)]. Propensity scores are estimated using the logistic model, with specifications as follows:

[e] PSID-1: Prob (T_i = 1) = F(age, age^2, education, education^2, married, no degree, black, Hispanic, RE74, RE75, RE74^2, RE75^2, u74*black).

[f] PSID-2 and PSID-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE74^2, RE75, RE75^2, u74, u75).

[g] CPS-1, CPS-2, and CPS-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE75, u74, u75, education*RE74, age^3).

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.

We also consider estimates from the subsets of the PSID and CPS. In Table 2 the estimates tend to improve when applied to narrower subsets. However, the estimates still range from −$3,822 to $1,326. In Table 3 the estimates do not improve for the subsets, although the range of fluctuation is narrower, from $587 to $2,321. Tables 1 and 4 shed light on this.

Table 1 presents the preintervention characteristics of the various comparison groups. We note that the subsets—PSID-2 and -3 and CPS-2 and -3—although more closely resembling the treatment group are still considerably different in 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 correspond closely to the treatment group; none of the differences is statistically significant. But as we create subsets of the comparison groups, the quality of the matches declines, most dramatically for the PSID. PSID-2 and -3 earnings now increase from 1974 to 1975, whereas they decline for the treatment group. The training literature has identified the "dip" in earnings as an important characteristic of participants in training programs (see Ashenfelter 1974, 1978). The CPS subsamples retain the dip, but 1974 earnings are substantially higher for the matched subset of CPS-3 than for the treatment group.

This illustrates one of the important features of propensity score methods, namely that creation of ad hoc subsamples from the nonexperimental comparison group is neither necessary nor desirable; subsamples based on single preintervention characteristics may dispose of comparison units that still provide good overall comparisons with treatment units. The propensity score sorts out which comparison units are most relevant, considering all preintervention characteristics simultaneously, not just one characteristic at a time.

Column (3) in Table 3 illustrates the value of allowing both for a heterogeneous treatment effect and for a nonlinear functional form in the propensity score. The estimators in columns (4)–(8) have both of these characteristics, whereas column (3) regresses 1978 earnings on a less nonlinear function [quadratic, as opposed to the step function in columns (4) and (5)] 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 preintervention characteristics, and are farther from the experimental benchmark than the estimates in columns (4)–(8). This demonstrates the ability of the propensity score to summarize all preintervention variables, but underlines the importance of using the propensity score in a sufficiently nonlinear functional form.

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 to 550 and 886 in Table 2, Panel C, column (5). This is because the propensity score estimators use fewer observations. When stratifying on the propensity score, we discard irrelevant controls, so that 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

The upper half of Table 5 demonstrates that the estimates of the treatment impact are not particularly sensitive to the specification used for the propensity score. Specifications 1 and 4 are the same as those in Table 3 (and hence they bal-

Table 4. Sample Means of Characteristics for Matched Control Samples

Matched samples	*No. of observations*	*Age*	*Education*	*Black*	*Hispanic*	*No degree*	*Married*	*RE74 (U.S. $)*	*RE75 (U.S. $)*
NSW	185	25.81	10.35	.84	.06	.71	.19	2,096	1,532
MPSID-1	56	26.39	10.62	.86	.02	.55	.15	1,794	1,126
		[2.56]	[.63]	[.13]	[.06]	[.13]	[.12]	[1,406]	[1,146]
MPSID-2	49	25.32	11.10	.89	.02	.57	.19	1,599	2,225
		[2.63]	[.83]	[.14]	[.08]	[.16]	[.16]	[1,905]	[1,228]
MPSID-3	30	26.86	10.96	.91	.01	.52	.25	1,386	1,863
		[2.97]	[.84]	[.13]	[.08]	[.16]	[.16]	[1,680]	[1,494]
MCPS-1	119	26.91	10.52	.86	.04	.64	.19	2,110	1,396
		[1.25]	[.32]	[.06]	[.04]	[.07]	[.06]	[841]	[563]
MCPS-2	87	26.21	10.21	.85	.04	.68	.20	1,758	1,204
		[1.43]	[.37]	[.08]	[.05]	[.09]	.08	[896]	[661]
MCPS-3	63	25.94	10.69	.87	.06	.53	.13	2,709	1,587
		[1.68]	[.48]	[.09]	[.06]	[.10]	[.09]	[1,285]	[760]

NOTE: Standard error on the difference in means with NSW sample is given in brackets.
MPSID1–3 and MCPS1–3 are the subsamples of PSID1–3 and CPS1–3 that are matched to the treatment group.

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

	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*	
Comparison group	*(1) Unadjusted*	*(2) Adjusted*[a]	*(3)*	*(4) Unadjusted*	*(5) Adjusted*	*(6) Observations*[d]	*(7) Unadjusted*	*(8) Adjusted*[b]
NSW	1,794	1,672						
	(633)	(638)						
Dropping higher-order terms								
PSID-1:	−15,205	218	294	1,608	1,254	1,255	1,691	1,054
Specification 1	(1,154)	(866)	(1,389)	(1,571)	(1,616)		(2,209)	(831)
PSID-1:	−15,205	105	539	1,524	1,775	1,533	2,281	2,291
Specification 2	(1,154)	(863)	(1,344)	(1,527)	(1,538)		(1,732)	(796)
PSID-1:	−15,205	105	1,185	1,237	1,155	1,373	1,140	855
Specification 3	(1,154)	(863)	(1,233)	(1,144)	(1,280)		(1,720)	(906)
CPS-1:	−8,498	738	1,117	1,713	1,774	4,117	1,582	1,616
Specification 4	(712)	(547)	(747)	(1,115)	(1,152)		(1,069)	(751)
CPS-1:	−8,498	684	1,248	1,452	1,454	6,365	835	904
Specification 5	(712)	(546)	(731)	(632)	(2,713)		(1,007)	(769)
CPS-1:	−8,498	684	1,241	1,299	1,095	6,017	1,103	1,471
Specification 6	(712)	(546)	(671)	(547)	(925)		(877)	(787)
Dropping RE74								
PSID-1:	−15,205	−265	−697	−869	−1,023	1,284	1,727	1,340
Specification 7	(1,154)	(880)	(1,279)	(1,410)	(1,493)		(1,447)	(845)
PSID-2:	−3,647	297	521	405	304	356	530	276
Specification 8	(959)	(1,004)	(1,154)	(1,472)	(1,495)		(1,848)	(902)
PSID-3:	1,069	243	1,195	482	−53	248	87	11
Specification 8	(899)	(1,100)	(1,261)	(1,449)	(1,493)		(1,508)	(938)
CPS-1:	−8,498	525	1,181	1,234	1,347	4,558	1,402	861
Specification 9	(712)	(557)	(698)	(695)	(683)		(1,067)	(786)
CPS-2:	−3,822	371	482	1,473	1,588	1,222	1,941	1,668
Specification 9	(670)	(662)	(731)	(1,313)	(1,309)		(1,500)	(755)
CPS-3:	−635	844	722	1,348	1,262	504	1,097	1,120
Specification 9	(657)	(807)	(942)	(1,601)	(1,600)		(1,366)	(783)

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

[a] Least squares regression: RE78 on a constant, a treatment indicator, age, education, no degree, black, Hispanic, 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)–(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.

ance the preintervention characteristics). In specifications 2–3 and 5–6, we drop the squares and cubes of the covariates, and then the interactions and dummy variables. In specifications 3 and 6, the logits simply use the covariates linearly. These estimates are farther from the experimental benchmark than those in Table 3, ranging from $835 to $2,291, but they remain concentrated compared to the range of estimates from Table 2. Furthermore, for the alternative specifications, we are unable to find a partition structure such that the preintervention characteristics are balanced within each stratum, which then constitutes a well-defined criterion for rejecting these alternative specifications. Indeed, the specification search begins with a linear specification, then 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 influence assignment to treatment and that 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 2 years (rather than 1 year) of preintervention earnings data is available. In this section we consider how our estimators would fare in the absence of 2 years of preintervention earnings data. In the bottom part of Table 5, we reestimate the treatment impact without using 1974 earnings. For PSID1–3, the stratification estimates (ranging from −$1,023 to $482) are more variable than the regression estimates in column (2) (ranging from −$265 to $297) and the estimates in Table 3, which use 1974 earnings (ranging from $1,494 to $2,321). The estimates from matching vary less than those from stratification. Compared to the PSID estimates, the estimates from the CPS are closer to the experimental benchmark (ranging from $1,234 to $1,588 for stratification and from $861 to $1,941 for matching). They are also closer than the regression estimates in column (2).

The results clearly are sensitive to the set of preintervention variables used, but the degree of sensitivity varies with the comparison group. This illustrates the importance of a sufficiently lengthy preintervention earnings history

for training programs. Table 5 also demonstrates the value of using multiple comparison groups. Even if we did not know the experimental estimate, the variation in estimates between the CPS and PSID would raise the concern 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. 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. Rosenbaum (1987) has developed this idea in more detail.

6. CONCLUSIONS

In this article we have demonstrated how to estimate the treatment impact in an observational study using propensity score methods. Our contribution is to demonstrate the use of propensity score methods and to apply them in a context that allows us to assess their efficacy. Our results show that the estimates of the training effect for Lalonde's hybrid of an experimental and nonexperimental dataset are close to the benchmark experimental estimate and are robust to the specification of the comparison group and to the functional form used to estimate the propensity score. A researcher using this method would arrive at estimates of the treatment impact ranging from $1,473 to $1,774, 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 attempted to follow this strategy in his construction of other comparison groups, his method relies on an informal selection based on preintervention variables. Our application illustrates that even among a large set of potential comparison units, very few may be relevant, and that even a few comparison units may be sufficient 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. However, rather than giving up, or relying on assumptions about the unobserved variables, there is substantial reward in exploring first the information contained in the variables that *are* observed. In this regard, propensity score methods can offer both a diagnostic on the quality of the comparison group and a means to estimate the treatment impact.

[Received October 1998. Revised May 1999.]

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, WI: Industrial Relations Research Association.

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

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

Card, D., and Sullivan, D. (1988), "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment," *Econometrica*, 56, 497–530.

Dawid, A. P. (1979), "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society*, Ser. B, 41, 1–31.

Dehejia, R. H., and Wahba, S. (1998), "Matching Methods for Estimating Causal Effects in Non-Experimental Studies," Working Paper 6829, National Bureau of Economic Research.

Härdle, W., and Linton, O. (1994), "Applied Nonparametric Regression," in *Handbook of Econometrics*, Vol. 4, eds. R. Engle and D. L. McFadden, Amsterdam: Elsevier, pp. 2295–2339.

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

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

Heckman, J., Ichimura, H., and Todd, P. (1997), "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies*, 64, 605–654.

——— (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Heckman, J., and Robb, R. (1985), "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, Econometric Society Monograph No. 10, eds. J. Heckman and B. Singer, Cambridge, U.K.: Cambridge University Press, pp. 63–113.

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

Lalonde, R. (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, U.K.: Ballinger.

Manski, C. F., and Garfinkel, I. (1992), "Introduction," in *Evaluating Welfare and Training Programs*, eds. C. Manski and I. Garfinkel, Cambridge, U.K.: Harvard University Press, pp. 1–22.

Manski, C. F., Sandefur, G., McLanahan, S., and Powers, D. (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, 292–316.

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

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

Rubin, D. (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.