Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods

By JEFFREY A. SMITH AND PETRA E. TODD*

There is a long-standing debate in the literature over whether social programs can be reliably evaluated without a randomized experiment. This paper summarizes results from a larger paper (Smith and Todd, 2001) that uses experimental data combined with nonexperimental data to evaluate the performance of alternative nonexperimental estimators. The impact estimates based on experimental data provide a benchmark against which to judge the performance of nonexperimental estimators. Our experimental data come from the National Supported Work (NSW) Demonstration and the nonexperimental data from the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). These same data were used in influential papers by Robert LaLonde (1986), James Heckman and Joseph Hotz (1989), and Rajeev Dehejia and Sadek Wahba (1998, 1999).

We focus on a class of estimators called *propensity-score matching estimators,* which were introduced in the statistics literature by Paul Rosenbaum and Donald Rubin (1983). Traditional propensity-score matching methods pair each program participant with a single nonparticipant, where pairs are chosen based on the degree of similarity in the estimated probabilities of participating in the program (the propensity scores). More recently developed nonparametric matching estimators described in Heckman et al. (1997, 1998a, b) use weighted averages over multiple observations to construct matches. We apply both kinds of estimators in this paper.

Heckman et al. (1997, 1998a, b) evaluate the performance of matching estimators using experimental data from the U.S. National Job Training Partnership Act (JTPA) Study combined with comparison group samples drawn from three sources. They show that data quality is a crucial ingredient to any reliable estimation strategy. Specifically, the estimators they examine are only found to perform well in replicating the results of the experiment when they are applied to comparison group data satisfying the following criteria: (i) the same data sources (i.e., the same surveys or the same type of administrative data or both) are used for participants and nonparticipants, so that earnings and other characteristics are measured in an analogous way, (ii) participants and nonparticipants reside in the same local labor markets, and (iii) the data contain a rich set of variables relevant to modeling the program-participation decision. If the comparison group data fails to satisfy these criteria, the performance of the estimators diminishes greatly.

More recently, Dehejia and Wahba (1998, 1999) have used the NSW data (also used by LaLonde) to evaluate the performance of propensity-score matching methods, including pairwise matching and caliper matching. They find that these simple matching estimators succeed in closely replicating the experimental NSW results, even though the comparison group data do not satisfy any of the criteria found to be important in Heckman et al. (1997, 1998a). From this evidence, they conclude that matching approaches are generally more reliable than traditional econometric estimators.

In this paper, we reanalyze the NSW data in an attempt to reconcile the conflicting findings

* Smith: Department of Economics, University of Western Ontario, Social Science Centre, London, ON, N6A 5C2, Canada, and NBER; Todd: Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, and NBER. Versions of the paper that is the source for this paper have been presented at a 2000 meeting of the Institute for Research on Poverty in Madison, Wisconsin, at the Western Research Network on Employment and Training summer workshop (August 2000), at the Canadian International Labour Network meetings (September 2000), and at the University of North Carolina. We thank James Heckman for comments on this paper, and we thank Robert LaLonde for comments and for providing us with the data from his 1986 study. We thank Rajeev Dehejia for providing us with information helpful in reconstructing the samples used in the Dehejia and Wahba (1998, 1999) studies. Jingjing Hsee and Miana Plesca provided excellent research assistance.

on the performance of matching estimators. In particular, we examine the sensitivity of the Dehejia and Wahba (1998, 1999) results to their choice of sample, the choice of variables for the propensity-score model, and the choice of matching method. We find that the exclusion of about 40 percent of LaLonde's observations by Dehejia and Wahba in order to incorporate one additional variable into their propensity-score model has a strong effect on their results. In fact, most conventional econometric estimators applied to their particular subsample exhibit lower bias values. The finding of low bias in Dehejia and Wahba (1998, 1999) is also sensitive to which set of variables is included in the propensity-score model. In contrast, their results (and ours) are not particularly sensitive to the details of the matching procedure used, as long as a common support condition (described below) is imposed.

Overall, our findings here, and those in Smith and Todd (2001), serve to reconcile the seemingly conflicting findings in the existing literature. Except in the special case of Dehejia and Wahba's sample and their propensity-score specification, the matching estimators applied to the NSW data often exhibit substantial biases. This finding is consistent with the fact that the NSW data combined with LaLonde's nonexperimental comparison groups do not place nonparticipants in the same local labor markets as participants, do not measure the dependent variable (earnings) in the same way across samples, and do not include a particularly rich set of covariates for matching.

I. The Evaluation Problem

Assessing the impact of any intervention requires making an inference about the outcomes that would have been observed for program participants had they not participated. Denote by Y_1 the outcome conditional on participation and by Y_0 the outcome conditional on nonparticipation, so that the impact of participating in the program is $\Delta = Y_1 - Y_0$. For each person, only Y_1 or Y_0 is observed, which leads to a missing-data problem. Let $D = 1$ for individuals who applied and got accepted into the program, for whom Y_1 is observed, and let $D = 0$ for persons who do not enter the program, for whom Y_0 is observed. Let $\mathbf{Z}$ denote a vector of observed individual characteristics used as conditioning variables. The most common evaluation parameter of interest is the *mean impact of treatment on the treated* (TT):

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

This parameter estimates the average impact among participants. It is the parameter on which LaLonde (1986) and Dehejia and Wahba (1998, 1999) focus, and it is a central parameter in many evaluations.[1] Data on program participants identify the mean outcome in the treated state, $E(Y_1|\mathbf{Z}, D=1)$. In a social experiment, where persons who would otherwise participate are randomly denied access to the program, the randomized-out control group provides a direct estimate of $E(Y_0|\mathbf{Z}, D=1)$. However, in nonexperimental (or observational) studies, no direct estimate of this counterfactual mean is available.

II. Matching Estimators

For simplicity, let $P = \Pr(D = 1|\mathbf{Z})$ denote the probability of participating in the program (the propensity score). Cross-sectional matching estimators assume

$$E(Y_0|P, D = 1) = E(Y_0|P, D = 0)$$

along with the condition $P < 1$. The latter condition ensures that matches can be found for each program participant. Under these assumptions, the mean impact of the program is given by

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

The second term can be estimated from the mean outcomes of the matched (on P) comparison

[1] See Heckman et al. (1999) for discussions of other parameters of interest.

group.[2] A typical matching estimator takes the form

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

where $\hat{E}(Y_{0i}|D = 1, P_i) = \sum_{j \in I_0} W(i, j)Y_{0j}$ is the matched outcome. I_1 denotes the set of program participants, I_0 the set of nonparticipants, S_P the region of common support (see below), and n_1 denotes the number of persons in the set $I_1 \cap S_P$. The match for each participant $i \in I_1 \cap S_P$ is constructed as a weighted average over the outcomes of nonparticipants, where the weights $W(i, j)$ depend on the distance between P_i and P_j. Define a neighborhood $C(P_i)$ for each i in the participant sample. Neighbors for i are nonparticipants $j \in I_0$ for whom $P_j \in C(P_i)$. The persons matched to i are those people in set $\mathcal{A}_i$ where $\mathcal{A}_i = \{j \in I_0 | P_j \in C(P_i)\}$. Alternative matching estimators differ in how the neighborhood is defined and in how the weights $W(i, j)$ are constructed. For example, *nearest-neighbor matching* sets $C(P_i) = \min_j \|P_i - P_j\|, j \in I_0$. That is, the nonparticipant with the value of P_j that is closest to P_i is selected as the match and $\mathcal{A}_i$ is a singleton set. This estimator is often used in practice due to its ease of implementation.

Recently developed nonparametric matching estimators construct a match for each program participant using a kernel-weighted average over multiple persons in the comparison group. Heckman et al. (1997, 1998a, b) describe in detail the *local linear matching estimator* that we use in this paper.[3] To implement the matching estimator given by equation (1), the region of common support, S_P, needs to be determined. To determine this region, we use a method described in Heckman et al. (1997) and Smith and Todd (2001), which is based on direct nonparametric estimation of the density within the $D = 0$ and $D = 1$ samples.

III. When Does Bias Arise in Matching?

The success of a matching estimator depends on the availability of observable data to construct the conditioning set $\mathbf{Z}$, such that the identifying conditions are satisfied. Suppose only a subset $\mathbf{Z}_0$ of the variables required for matching is observed. The propensity score matching estimator based on $\mathbf{Z}_0$ then converges to $E_{P(\mathbf{Z}_0)|D=1}(E(Y_1|P(\mathbf{Z}_0), D = 1) - E(Y_0|P(\mathbf{Z}_0), D = 0))$. The bias for the parameter of interest (TT) is $E(Y_0|D = 1) - E_{P(\mathbf{Z}_0)|D=1}\{E(Y_0|P(\mathbf{Z}_0), D = 0)\}$. Heckman et al. (1998a) show that which variables are included in the propensity score matters in practice for the estimated bias. They found that the lowest bias values were obtained when the $\mathbf{Z}$ data included a rich set of variables relevant to modeling the program-participation decision. Higher bias values were obtained for a cruder set of $\mathbf{Z}$ variables.

IV. Using Data on Randomized-Out Controls and Nonparticipants to Estimate Bias

With only nonexperimental data, it is impossible to disentangle the treatment effect from the evaluation bias associated with any particular estimator. Data on a randomized-out control group make it possible to separate out the bias. First, randomization ensures that the control group is identical to the treatment group in terms of the pattern of self-selection. Second, the randomized-out control group does not participate in the program, so the impact of the program on them is known to be zero. Thus, a nonexperimental estimator applied to the control-group data combined with the nonexperimental comparison group data should, if consistent, produce an estimated impact equal to zero. Deviations from zero are interpretable as evaluation bias. Therefore, the performance of nonexperimental estimators can be evaluated by applying them to data from a randomized-out control group and a nonexperimental comparison group and then checking whether the resulting estimates equal zero.

V. The Data

The NSW Demonstration was a transitional, subsidized work experience program that provided work in a sheltered training environment

[2] For in-depth discussion of these assumptions, see Smith and Todd (2001).

[3] See Heckman et al. (1998b) for conditions required for consistency and asymptotic normality of the local linear matching estimator.

and assisted in job placement. Robinson Hollister et al. (1984) provide a detailed description of the NSW Demonstration, which operated from April 1975 to August 1977 in 10 cities as a randomized experiment, with some applicants randomly assigned to a control group that was not allowed to participate in the program. LaLonde (1986) shows that male NSW participants were almost all minorities (mostly black), high-school dropouts, and unmarried. The earnings variables for the NSW experimental samples all derive from self-reported earnings measures from surveys. Following LaLonde (1986), all of the earnings variables (for all of the samples) are expressed in 1982 dollars. Mean earnings in the male NSW sample prior to random assignment were quite low. They also fall from 1974 to 1975, another example of the common pattern denoted "Ashenfelter's dip" in the literature (see e.g., Heckman and Smith, 1999).

In this study, we consider three experimental samples and two nonexperimental comparison groups. All of the samples are based on the male samples from LaLonde (1986). The first experimental sample, which consists of 297 treatment-group observations and 425 control-group observations, is the same as that employed by LaLonde (1986). The experimental impact estimate for this group is $886, which is statistically significant at the 10-percent level. The second experimental sample is that used by Dehejia and Wahba (1998, 1999). In order to include two years of pre-program earnings in their model for program participation, Dehejia and Wahba omit about 40 percent of LaLonde's original sample for which that information was missing. In particular, they include all persons randomly assigned in January through April of 1976. For these persons, the variable in LaLonde's data set that gives earnings in months 13–24 before random assignment roughly corresponds to earnings in 1974 (the variable is labeled "real earnings in 1974" in their papers). They also include persons randomly assigned after this date, but only if they have zero earnings in months 13–24 before random assignment. The justification for including this latter group is somewhat unclear. The effect of differentially including persons with zero earnings in months 13–24 prior to random assignment is to reduce the extent of Ashenfelter's dip in the experimental sample. Smith and Todd (2001) show that for this particular subsample, most conventional regression-based nonexperimental estimators have low bias values. The experimental impact estimate for the Dehejia and Wahba sample is $1,794, almost twice as large as that for the LaLonde sample.

The third experimental sample we examine is not used in the previous studies. It excludes all persons who were randomized after April 1976, so as to include earnings in months 13–24 but not just include those with positive earnings in later months (the Dehejia and Wahba restriction). This third experimental sample includes 108 treatment-group members and 142 control-group members and is a proper subset of the Dehejia and Wahba sample. Ashenfelter's dip is stronger for this sample (a drop of about $1,200 rather than one of about $700) than for the Dehejia and Wahba sample, as is to be expected given that it drops the large contingent of persons with zero earnings in months 13–24 prior to random assignment (RA). The $2,748 experimental impact for this last ("early RA") sample is the largest among the three experimental samples.

The comparison-group sample we use in this paper is the CPS sample also used by LaLonde (1986) and Dehejia and Wahba (1998, 1999). It is based on Westat's matched Current Population Survey–Social Security Administration file. This file contains male respondents from the March 1976 Current Population Survey (CPS) with matched Social Security earnings data. The sample excludes persons with nominal own incomes greater than $20,000 and nominal family incomes greater than $30,000 in 1975. Men over age 55 are also excluded. LaLonde (1986) shows that the full CPS comparison-group sample is much older, much less likely to be a minority, better educated (70 percent completed high school), and much more likely to be married than any of the NSW experimental samples.

The earnings measures for the CPS sample are individual-level administrative annual earnings totals from the U.S. Social Security system. The CPS comparison-group sample had, on average, much higher earnings than the NSW experimental sample in every year. (The "Real Earnings in 1974" variable for the CPS comparison group corresponds to calendar-year 1974.) There is a slight dip in the mean earnings of the CPS comparison group from 1974 to 1975, which is consistent with the imposition of maximum individual and family income criteria in 1975 for inclusion in the sample, along with some level of mean-reversion in earnings. The very substantial differences between this comparison group and the NSW experimental group poses a tough problem for any nonexperimental estimator to solve. In addition, because the CPS sample is a national one, only a very small number of comparison-group members will be in the same local labor markets as the NSW participants.

TABLE 1—BIAS ASSOCIATED WITH ALTERNATIVE MATCHING ESTIMATORS

	Sample and propensity-score model[a]			
Estimator	(i)	(ii)	(iii)	(iv)
Mean difference	−9,757 (255) [−1,101]	−10,291 (306) [−574]	−11,101 (461) [−404]	−10,227 (296) [−1,154]
1 NN, no CS	−555 (596) [−63]	407 (698) [23]	−7,781 (1,245) [−283]	−3,602 (1,459) [−406]
10 NN, no CS	−270 (493) [−30]	−5 (672) [−0.3]	−3,632 (1,354) [−132]	−2,122 (1,299) [−240]
1 NN, with CS	−838 (628) [−95]	−27 (723) [−1.5]	−5,417 (1,407) [−197]	−3,586 (1,407) [405]
10 NN, with CS	−1,299 (529) [−147]	−261 (593) [−15]	−2,396 (1,152) [−87]	−2,342 (1,165) [264]
Local linear (bw = 1.0)	−1,380 (437) [−156]	−88 (630) [−5]	−3,427 (1,927) [−125]	−3,562 (3,969) [402]
Local linear (bw = 4.0)	−1,431 (441) [−162]	−67 (611) [−4]	−2,191 (1,069) [−80]	−2,708 (1,174) [306]

Notes: The comparison group is the CPS adult male sample. The dependent variable is real earnings in 1978. NN = nearest neighbor; CS = common support; bw = bandwidth. Numbers in parentheses show standard errors. Numbers in square brackets are the percentages of the experimental impact estimate for that sample that the bias estimate represents.

[a] The samples and propensity-score models in the four columns are: (i) LaLonde sample with Dehejia and Wahba (DW) propensity scores (impact estimate = $886); (ii) DW sample with DW propensity scores (impact estimate = $1,794); (iii) early RA sample with DW propensity scores (impact estimate = $2,748); (iv) LaLonde sample with LaLonde propensity scores (impact estimate = $886).

VI. Empirical Results

We now present our estimates of the bias obtained when we apply matching estimators to the experimental NSW data and the CPS comparison groups. Our estimation strategy differs somewhat from that of LaLonde (1986) and Dehejia and Wahba (1998, 1999) in that we obtain direct estimates of the bias by applying matching to the randomized-out control group and nonexperimental comparison group, whereas the other papers obtain the bias indirectly by applying matching to the treatment and comparison groups and comparing the experimental and the nonexperimental estimates. Another difference is that we match on the log-odds ratio rather than on the propensity score itself, so that our estimates are robust to choice-based sampling (see Smith and Todd, 2001). We present matching estimates based on two alternative specifications of the propensity score. The first specification is that of Dehejia and Wahba (1998, 1999), and the second specification is that of LaLonde (1986). Estimated coefficients and standard errors for the propensity-score models are reported in Smith and Todd (2001).

Estimates of the bias associated with cross-sectional matching appear in Table 1. The outcome variable is earnings in calendar-year 1978, where January 1978 is at least five months after random assignment for all of the controls. The first row of Table 1 gives the simple mean difference in 1978 earnings between each experimental control group and the CPS comparison group. The remaining rows present estimates of the bias associated with different matching estimators. The first three columns of the table refer to estimates using the Dehejia and Wahba propensity-score specification, while the final column refers to the LaLonde propensity-score specification. We present bias estimates for each experimental sample, along with the percentage of the experimental impact estimate for that sample that the bias estimate represents. These percentages are useful for comparisons of different estimators within each column, but are not useful for comparisons across columns, given the large differences in experimental impact estimates among the three experimental samples.

The second through the fifth rows give various estimates based on nearest-neighbor matching, using either the one nearest or ten nearest neighbors, with and without imposing the common-support condition. Five important patterns characterize the nearest-neighbor estimates. First, using Dehejia and Wahba's experimental sample and their propensity-score model, we replicate the low biases that were reported in Dehejia and Wahba (1998, 1999). Second, when Dehejia and Wahba's propensity-score model is applied to the LaLonde sample or to the early RA sample, the bias estimates are substantially higher. Indeed, the bias estimates for the Dehejia and Wahba scores as applied to the early RA sample are among the largest in the table. Third, the imposition of the common-support condition has little effect on the estimates for LaLonde and Dehejia-Wahba samples, but it does result in a substantial bias reduction for the early RA sample. Fourth, increasing the number of nearest neighbors reduces bias in the relatively small early RA sample but does little to change the bias estimates for the other two experimental samples. Fifth, when the LaLonde propensity-score model is applied to the LaLonde sample, it does quite poorly in terms of bias, though not as poorly as the Dehejia-Wahba scores in the early RA sample. Thus, the results obtained by Dehejia and Wahba (1998, 1999) using simple nearest-neighbor matching on their sample are highly sensitive both to changes in the sample composition and to changes in the variables included in the propensity-score model.

The last two rows present estimates obtained using local linear matching methods with two different bandwidths. In general, increasing the bandwidth will increase the bias and reduce the variance associated with the estimator by putting a heavier weight on the information provided by more distant observations in constructing the counterfactual for each $D = 1$ observation. Interestingly, both the variance and the overall average bias usually decrease when we increase the bandwidth.

In Smith and Todd (2001), we present additional results obtained using comparison-group samples drawn from the PSID and using difference-in-difference matching estimators applied to both the CPS and PSID data. Inference from the PSID is generally similar to that obtained from the CPS samples, revealing extreme sensitivity of the impact estimates to changes in the sample and to changes in the specification of the propensity-score model. Again, we only find low bias estimates for the Dehejia-Wahba subsample. In comparing the performance of the cross-sectional matching methods to the difference-in-difference methods, we find that the difference-in-difference matching estimators generally exhibit better overall performance. The bias estimates are lower in most cases for the Dehejia-Wahba scores and the early RA sample and in all cases with the LaLonde scores applied to the LaLonde sample.

The implications of our findings for evaluation research are clear. When matching methods are applied to high-quality data, they have been found to perform reasonably well. In Heckman et al. (1997, 1998a), matching methods applied to the data from the JTPA experiment yield biases of roughly the same order of magnitude as the experimental impact estimates. However, the CPS and PSID comparison groups in LaLonde's (1986) and Dehejia and Wahba's (1998, 1999) studies using the NSW data suffer from the problem of geographic mismatch and from the dependent variable being measured in different ways in the participant and nonparticipant samples. Our findings using the NSW data show that the finding of low bias for matching estimators using the NSW data is very sensitive to both the experimental sample employed and to the set of variables included in the matching. A more general analysis finds biases substantially larger, relative to the experimental impact estimates, in the NSW data than in the JTPA data. This finding is consistent with the conclusions regarding the importance of data quality drawn by Heckman et al. (1997, 1998a).

REFERENCES

Dehejia, Rajeev and Wahba, Sadek. "Propensity Score Matching Methods for Nonexperimental Causal Studies." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 6829, 1998.

______ . "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*, December 1999, *94*(448), pp. 1053–62.

Heckman, James and Hotz, Joseph. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs:

The Case of Manpower Training." *Journal of the American Statistical Association*, December 1989, *84*(408), pp. 862–74.

Heckman, James; Ichimura, Hidehiko; Smith, Jeffrey and Todd, Petra. "Characterizing Selection Bias Using Experimental Data." *Econometrica*, September 1998a, *66*(5), pp. 1017–98.

Heckman, James; Ichimura, Hidehiko and Todd, Petra. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies*, October 1997, *64*(4), pp. 605–54.

______ . "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies*, April 1998b, *65*(2), pp. 261–94.

Heckman, James; LaLonde, Robert and Smith, Jeffrey. "The Economics and Econometrics of Active Labor Market Programs," in Orley Ashenfelter and David Card, eds., *Handbook of labor economics*, Vol. 3A. Amsterdam: North-Holland, 1999, pp. 1865–2097.

Heckman, James and Smith, Jeffrey. "The Pre-programme Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies." *Economic Journal*, July 1999, *109*(457), pp. 313–48.

Hollister, Robinson; Kemper, Peter and Maynard, Rebecca. *The National Supported Work Demonstration*. Madison: University of Wisconsin Press, 1984.

LaLonde, Robert. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, September 1986, *76*(4), pp. 604–20.

Rosenbaum, Paul and Rubin, Donald. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, April 1983, *70*(1), pp. 41–55.

Smith, Jeffrey and Todd, Petra. "Does Matching Address LaLonde's Critique of Nonexperimental Estimators." Unpublished manuscript, University of Western Ontario, 2001.