Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Econometrics 125 (2005) 365–375
JOURNAL OF Econometrics
www.elsevier.com/locate/econbase

Rejoinder

Jeffrey Smith[a,*], Petra Todd[b]

[a] Department of Economics, University of Maryland, 3105 Tydings Hall, College Park, MD, 20742-7211, USA
[b] Department of Economics, University of Pennsylvania, USA

Available online 17 July 2004

1. Introduction

In this rejoinder, we address the points made in the response by Prof. Dehejia. As his response argues that propensity score matching provides a robust solution to the evaluation problem only for the sample employed in the Dehejia and Wahba (here after DW) (1999, 2002) papers, we focus our attention on that sample in the rejoinder. We argue that, in fact, the evidence from the DW sample clearly shows that propensity score matching does not solve the evaluation problem in this context.

In Section 2, we clarify the purpose of some of the remarks in the introduction and conclusion of our original paper. In Section 3, we show the sensitivity of bias estimates from single nearest neighbor matching to changes in the experimental sample (treatment or control) used to estimate the bias, the experimental sample (treatment, control, or both) used to estimate the propensity score, and the choice of seed for the random number generator used to break ties in the matching. We also show the sensitivity of estimates from kernel matching to the first two factors. In Section 4 we apply three alternative balancing tests to the data, and show that we can reject the balancing condition for the scores employed in the response using these alternative tests. Section 5 concludes.

2. Dehejia and Wahba (1999, 2002) in the literature

Before turning to the substantive issues raised by the response, a few remarks about the statements in the introduction and conclusion to our paper regarding the

*Corresponding author. Tel.: +1-301-405-3532; fax: +1-301-405-3542.
E-mail addresses: smith@econ.umd.edu (J. Smith), petra@athena.sas.upenn.edu (P. Todd).

0304-4076/$ - see front matter
doi:10.1016/j.jeconom.2004.04.013

perception in the literature of matching methods in general, and of the DW (1999, 2002) papers in particular, are in order. These statements are questioned in Section 2 of the response.

These remarks are aimed at the general readership of our paper, and not specifically at Prof. Dehejia, who makes it clear in his response that he is aware of the limitations of propensity score matching. Moreover, Prof. Dehejia is correct that neither of the DW (1999, 2002) papers claim that matching is a magic bullet or that it always solves the evaluation problem. What they do claim is that simple cross-sectional matching methods solve the evaluation problem when applied to the NSW data, even though the data are weak along several dimensions identified in the earlier literature (and discussed at length in our paper). Judging by the numerous citations to the DW (1999, 2002) papers (and their working paper predecessors) in both the published and unpublished literature, their findings have contributed strongly to the recent popularity of matching approaches even in weak data contexts, where we believe such methods are ineffective. Our analysis focused on the DW (1999, 2002) papers, rather than on other papers, precisely because they have been particularly influential in the literature.

3. Does matching solve the selection problem in the NSW data?

Dehejia asserts in his response that matching produces "reliable and robust" estimates in the subsample of the Supported Work data utilized in the DW (1999, 2002) papers. The evidence reported in this section disputes this claim.

Before presenting the evidence, however, we note that his response still provides no justification for the data subsample used in these two papers. Why were individuals with zero earnings in months 13–24 before random assignment who were randomly assigned after April 1976 included, while individuals with non-zero earnings in months 13–24 before random assignment who were randomly assigned after April 1976 were excluded? The response does not address this important question, which we raise in our paper. It seems to us that a researcher starting afresh with the NSW data would likely choose to use either the whole of LaLonde's sample, or what we term the early random assignment sample, but would be unlikely to set upon DW's sample.

We also remark that while we in general agree that different applications will require different propensity score specifications, it seems to us worrisome that small changes, such as restricting the dates of random assignment in the treated sample or including and excluding some observations with zero earnings, should require major changes in the propensity score. After all, the same general problem is being solved for all three NSW samples—namely, the problem of selection into NSW from either the CPS population or the PSID population.

To examine the sensitivity of propensity score matching applied to the DW sample, we next present 36 alternative estimates of the bias constructed using single nearest neighbor matching with replacement. Half of the estimates use the CPS comparison group and half of the estimates use the PSID comparison group. For

each comparison group, we use the propensity score specification recommended in Dehejia's response. Within each comparison group, the estimates differ depending on whether the propensity score model was estimated using the treatment group, the control group, or both; whether the bias was estimated using the treatment group or the control group; and the choice of random number seed.[1]

As a result of random assignment, for each comparison group, all of these estimates are consistent estimates of the same parameter—the bias associated with using a non-experimental estimator and comparison group rather than the experimental control group to estimate the impact of the NSW treatment. The point of looking at all of them is to show that, due to the small sample sizes, as well as to the existence of multiple observations with identical Z (the variables included in the propensity score) but different values of the dependent variable, nearest neighbor bias estimates are quite sensitive to choices that would be expected to make little difference in larger samples.

The bias estimates from single nearest neighbor matching with replacement appear in Table 1. The first two columns of estimates refer to the CPS sample and the second two columns refer to the PSID sample. For each comparison sample, the first column shows bias estimates obtained by estimating the mean impact of treatment on the treated using the experimental treatment group and the comparison sample.[2] The second column for each comparison sample presents direct estimates of the bias obtained by applying the matching estimator to the experimental control group and the comparison group. Thus, in every case, the estimate should be compared to zero.

Table 1 contains three groups of three rows. The first group of three rows reports estimates based on propensity scores estimated using the full experimental sample and the relevant comparison group. The second group reports estimates based on propensity scores estimated using only the experimental treatment group and the third group reports estimates based on propensity scores estimated using only the experimental control group. Within each group of three, each row presents estimates obtained using a different value for the seed for the random number generator (which will affect how ties are broken among comparison observations with identical scores).

In all cases, given our aim of replication and sensitivity analysis, we match on the level of the propensity score rather than the log odds ratio and we do not impose the support condition. The estimates in Table 1 that use propensity scores based on the experimental treatment group and which report a bias estimate obtained using the experimental treatment group correspond to those in the response in the sense that they are constructed using the same samples. Bootstrap standard errors appear in parentheses below each estimate.

[1] All of the estimates were obtained using Leuven and Sianesi's (2003) psmatch2 program for Stata (version 1.1.1 of 13 August 2003). Except for issues relating to the random number generator and rounding, their program should yield results equivalent to the S+program used in our original paper.

[2] The bias is obtained by subtracting the experimental impact estimate of $1794 for the DW sample from the non-experimental impact estimate.

Table 1
Bias estimates from single nearest neighbor matching with replacment

	CPS		PSID	
	Treatment group bias	Control group bias	Treatment group bias	Control group bias
All experimental observations used for scores				
Seed 1	−338.61 (780.88)	−665.39 (647.16)	258.51 (1017.54)	708.63 (761.98)
Seed 2	−141.61 (842.02)	−679.93 (636.26)	258.51 (894.95)	708.63 (750.19)
Seed 3	−264.16 (814.27)	−453.92 (619.02)	258.51 (1007.31)	708.63 (821.09)
Experimental treatment group observations used for scores				
Seed 1	−251.84 (734.49)	−663.23 (599.36)	74.73 (1053.91)	834.62 (785.86)
Seed 2	38.67 (810.51)	−663.15 (633.24)	74.73 (953.48)	834.62 (732.32)
Seed 3	−159.52 (780.72)	−586.25 (660.96)	74.73 (1032.67)	834.62 (883.65)
Experimental control group observations used for scores				
Seed 1	−1078.74 (810.26)	−956.02 (618.53)	−53.37 (1139.99)	712.56 (743.24)
Seed 2	−840.99 (810.94)	−964.58 (670.01)	−53.37 (1054.68)	712.56 (763.20)
Seed 3	−1042.21 (870.13)	−777.96 (619.07)	−53.57 (1011.56)	712.56 (829.62)

Source: Authors' calculations using NSW data.
Notes: Seed 1 = 123456789, Seed 2 = 987654321 and Seed 3 = 456789123. All of the estimates are based on single nearest neighbor matching with replacement using the DW experimental sample. We sort the observations by their identification number from the LaLonde dataset. No common support condition is imposed. The experimental impact estimate is $1794. Bootstrap standard errors based on 250 replications appear in parentheses; these standard errors do not account for the variance component due to estimation of the scores and, for the treatment group bias estimates, treat the experimental impact as a constant.

Four major findings emerge from Table 1. First, the numbers in the table bounce around a lot. The bias estimates for the CPS comparison group vary from $38.67 to $1078.74 in absolute value and those for the PSID comparison group vary from $53.37 to $834.62 in absolute value. This variation likely results from small sample sizes. The DW subset of LaLonde's sample includes only 185 treatment observations and 260 control observations. Both the CPS and the PSID comparison groups, despite their large nominal sample sizes, contain only small numbers of observations for which the propensity scores are comparable to NSW participants. Furthermore, the dependent variable—earnings in 1978—has a high variance in this population.

The second finding is that using the treatment group to estimate the propensity scores and to estimate the bias, which corresponds to the procedure used to generate the estimates in Table 2 of the response, yields relatively low bias estimates. For the CPS comparison group, these estimates comprise three of the four smallest (in absolute value) out of 18. For the PSID comparison group, they are the second smallest (in absolute value) out of six.

The third finding is that the bias estimates based on the largest sample sizes—those that rely on propensity scores estimated using the full experimental sample and biases estimated using the experimental control group—are substantial for both

Table 2
Bias estimates from kernel matching

Bandwidth	CPS		PSID	
	Treatment group bias	Control group bias	Treatment group bias	Control group bias
All experimental observations used for scores				
0.09	−1814.69 (588.13)	−1506.26 (428.22)	−645.31 (866.86)	49.10 (709.60)
0.06	−1813.82 (580.33)	−1329.70 (429.25)	−568.79 (918.43)	183.99 (721.02)
0.03	−1724.49 (578.66)	−1083.50 (415.72)	−485.02 (1029.74)	260.38 (814.67)
0.01	−1158.32 (615.84)	−680.95 (437.99)	−982.91 (1055.17)	296.98 (770.12)
0.005	−898.97 (594.46)	−597.62 (476.34)	−1345.87 (1013.79)	−54.23 (850.13)
Experimental treatment group observations used for scores				
0.09	−2134.54 (619.99)	−2195.39 (460.86)	−424.13 (923.26)	229.11 (727.65)
0.06	−1853.07 (595.63)	−1719.77 (428.52)	−395.35 (894.16)	428.81 (755.84)
0.03	−1747.12 (596.44)	−1410.06 (438.40)	−99.47 (929.98)	620.72 (756.41)
0.01	−1354.92 (575.63)	−775.73 (423.22)	−205.51 (1034.45)	924.28 (864.07)
0.005	−837.55 (573.83)	−456.63 (420.71)	−636.86 (1171.63)	608.80 (895.62)
Experimental control group observations used for scores				
0.09	−2018.00 (584.44)	−1664.82 (433.38)	−1080.86 (1069.11)	−234.00 (751.56)
0.06	−1855.71 (593.04)	−1540.26 (423.36)	−1026.96 (1070.98)	−167.37 (781.50)
0.03	−1824.72 (587.81)	−1290.92 (422.10)	−764.99 (1106.57)	87.66 (775.14)
0.01	−1830.18 (586.22)	−886.22 (424.18)	−909.08 (1133.66)	−172.10 (854.94)
0.005	−1353.11 (619.62)	−732.31 (458.59)	−1147.93 (1308.36)	−838.91 (1007.38)

Source: Authors' calculations using NSW data.
Notes: All of the estimates are based on kernel matching using the Epanechnikov kernel and the DW experimental sample. The experimental impact estimate is $1794. Bootstrap standard errors based on 100 replications appear in parentheses; these standard errors do not account for the variance component due to estimation of the scores and, for the treatment group bias estimates, treat the experimental impact estimate as a constant.

comparison groups. They range between –679.93 and –453.92 for the CPS comparison group and equal 708.63 for the PSID comparison group. In addition, the biases estimated using the experimental control group have smaller estimated standard errors due to its larger sample size. These estimates suggest that the very low bias estimates implicit in Table 2 of the response result from sampling variation in a small sample.

The fourth finding is that the CPS estimates, but not the PSID estimates, are also sensitive to the particular value selected for the random number seed. The reason for this is that there exist sets of observations in the combined CPS and experimental samples that have exactly the same values of the variables use to calculate the propensity score (and thus identical estimated propensity scores). For example, in the pooled treatment group and CPS comparison group sample, there are ten distinct groups that contain at least one treatment group member and two comparison group members—three of size three, three of size four, three of size six and one of size 10.

To see the trouble this can cause, consider one of the groups with six members. Three of the six are treatment group observations and three are CPS observations.

Although all have the same values for the variables included in the scores—in particular, they all have zero earnings in "1974" and 1975—the comparison group observations have very different earnings in 1978. One earns $12,760, one earns $4520 and one earns $1956. With single nearest neighbor matching with replacement, the combined estimated counterfactual earnings for the three treated observations in the group can thus range from $38,280 (if all three treatment group members get matched to the comparison observation earning $12,760) down to $5868 (if all three get matched to the comparison observation earning $1956). In a treatment group with only 185 observations, moving from one extreme to the other will change the estimated bias by over $500. Obviously, the presence of groups with tied propensity score values introduces a lot of variation into the estimates, and suggests the value of looking at bias estimates based on several different random sorts of the data, rather than just one, as in the response. These problems also arise when the CPS comparison group is combined with the experimental control group.

Similar problems with ties in the propensity score values do not arise for the PSID comparison group, presumably due to its smaller sample size. As a result, we match almost exactly the bias estimate implicit in Table 2 of Dehejia's response when we use only the experimental treatment group to estimate the propensity scores and the bias.

To show that the instability of the estimates evident in Table 1 does not result solely from the use of single nearest neighbor matching with replacement, we present estimates based on kernel matching in Table 2. Switching from single nearest neighbor matching to kernel matching trades bias for efficiency, because kernel matching uses information from less similar observations—those other than the nearest neighbor—but at the same time uses more observations overall in constructing the estimated counterfactual mean. The efficiency gain can be seen in the lower estimated standard errors in Table 2 relative to Table 1. The kernel matching estimates have the virtue that they do not depend on the random number seed for any of the samples, but the vice that they depend on the choice of bandwidth. We present estimates obtained using the Epanechnikov kernel and five widely spaced bandwidths; we find that, while the estimates have some sensitivity to the choice of bandwidth, the broader patterns do not.

Two key findings emerge. First, the kernel matching estimates exhibit similar, though not quite as dramatic, sensitivity to whether the treatment group or the control group produces the bias estimate, and to which subset of the experimental sample is used to estimate the propensity scores. Although the kernel matching estimates make use of more information from the comparison group sample, they do not change the overall fact that the DW subsample of the NSW data does not contain very many observations. Second, the kernel matching estimates reveal substantially larger biases, in general, than the single nearest neighbor matching with replacement estimates. This finding, which is more negative than the results obtained using local linear matching in our paper, provides further evidence in support of our conclusion that matching does not solve the evaluation problem in the NSW data, even for the DW sample.

Table 3
Balancing tests from single nearest neighbor matching with replacement

	CPS		PSID	
	Treatment group bias	Control group bias	Treatment group bias	Control group bias
All experimental observations used for scores				
Standardized differences	2.12/20.29	1.43/28.40	−1.69/13.12	3.49/16.74
Hotelling test	0.0000	0.0000	0.0014	0.0000
Regression test	3/6/6	4/4/5	6/8/8	9/10/10
Experimental treatment group observations used for scores				
Standardized differences	−2.08/20.77	−1.26/26.62	−4.29/11.74	−4.62/10.75
Hotelling test	0.0000	0.0000	0.0035	0.0000
Regression test	5/5/6	6/8/9	5/7/9	8/8/9
Experimental control group observations used for scores				
Standardized differences	1.48/19.80	−1.28/21.30	3.34/15.14	−0.90/10.39
Hotelling test	0.0000	0.0000	0.0012	0.0000
Regression test	3/4/5	4/7/8	7/8/8	9/9/10

Source: Authors' calculations using NSW data.
Notes: For the standardized differences, the first number is the 5th highest of the 10 standardized differences corresponding to the 10 variables age, education, no high school degree, black, Hispanic, married, earnings in "1974", earnings in 1975, no earnings in "1974" and no earnings in 1975. The second number is the maximum standardized difference. For the regression test, the first number is the number of p-values below 0.01 for the ten variables, the second the number of p-values below 0.05 and the third the number of p-values below 0.10. The estimates correspond to Seed 1 in Table 1, values for the other seeds are qualitatively similar.

4. Balancing tests

4.1. The balancing tests we use

As we make clear in our paper, we agree with the remarks in the response regarding the utility of balancing tests in choosing the specification of the propensity score model when a parametric model such as a logit or a probit is used to estimate the scores. At the same time, these tests have a number of limitations. The most obvious limitation at present is that multiple versions of the balancing test exist in the literature, with little known about the statistical properties of each one or of how they compare to one another given particular types of data. DW use a test based on balancing covariates within intervals but provide no objective way to choose the intervals, which is critical because the power of the test is low for narrow intervals. We consider three different balancing tests in this rejoinder, in part to make the point that different tests sometimes yield different answers (particularly in small samples).

The first test we consider comes from Rosenbaum and Rubin (1985) and relies on the examination of standardized differences. Using the notation in our paper, the

standardized difference is defined as

$$SDIFF(Z_k) = 100\frac{\frac{1}{n_1}\sum_{i\in I_1}\left[Z_{ki} - \sum_{j\in I_0} w(i,j)Z_{kj}\right]}{\sqrt{\frac{\mathrm{var}_{i\in I_1}(Z_{ki}) + \mathrm{var}_{j\in I_0}(Z_{kj})}{2}}}.$$

In words, the standardized difference for a variable Z_k is the difference in means between the treated sample and the matched (or, more generally, reweighted) comparison group sample, divided by the square root of the average of the variances of Z_k in the unweighted treatment and comparison groups. Intuitively, the standardized difference considers the size of the difference in means of a conditioning variable between the treated and matched comparison groups, scaled by the square root of the average of the variances in the original (unweighted) samples. This balancing test has been employed in recent work in economics by, among others, Lechner (1999, 2000) and Sianesi (2002).

A common practice in the literature is to compute the standardized difference for all of the variables included in the matching. Additional power can be obtained by considering other moments of these variables and interactions between them, although this is rarely done. We have not been able to find any formal criteria in the literature for when a standardized bias is too big, though Rosenbaum and Rubin (1985) suggest that a value of 20 is "large". In addition to the lack of formal criteria for evaluating the size of the standardized bias statistic, the standardized bias test has the disadvantage that the researcher can reduce the standardized bias by adding additional observations to the comparison group, so long as these additional observations increase the second variance term in the denominator.

The second balancing test we examine is a Hotelling T^2 test of the joint null of equal means of all of the variables included in the matching between the treatment group and the matched or reweighted comparison group.[3] As implemented here, using the command available in Stata, this test is not quite correct because it treats the matching weights as fixed rather than random. Whether the test is too conservative or the reverse depends on the (unknown) covariance between the component of the variance associated with the matching weights and the sampling variation conditional on the weights.

The third balancing test we employ tests the balancing condition in a regression framework. In particular, for each variable included in the propensity score, we estimate the following regression:

$$\begin{aligned} Z_k = {} & \beta_0 + \beta_1\hat{P}(Z) + \beta_2\hat{P}(Z)^2 \\ & + \beta_3\hat{P}(Z)^3 + \beta_4\hat{P}(Z)^4 + \beta_5 D + \beta_6 D\hat{P}(Z) \\ & + \beta_7 D\hat{P}(Z)^2 + \beta_8 D\hat{P}(Z)^3 + \beta_9 D\hat{P}(Z)^4 + \eta. \end{aligned}$$

We then test the joint null that the coefficients on all of the terms involving the treatment dummy equal zero. Essentially, we test whether D provides any

[3] This test was also used in Black and Smith (2004).

information about Z_k conditional on a quartic in the estimated propensity score. If the propensity score satisfies the balancing condition, it should not. The downside to this test is that it requires selection of the order of the polynomial, which may have implications for results of the test.

4.2. Evidence on balancing

The results obtained from the three balancing tests described in Section 4.1 appear in Table 3. The test results correspond to the estimates in Table 1 for the first random number seed. The balancing test results for the other two random number seeds with single nearest neighbor matching with replacement, and for the kernel matching, are qualitatively similar.

For each combination of comparison group, experimental sample used to estimate the bias, and experimental sample used to estimate the propensity score, the first row reports the 5th largest standardized bias out of the 10 variables used in the scores—age, education, no high school degree, black, Hispanic, married, earnings in "1974", zero earnings in "1974", earnings in 1975 and zero earnings in 1975. We also report the largest standardized bias among the 10 variables. The second row reports the p-value from the Hotelling T^2 test. The third row reports the number of p-values from the regression test applied to the 10 variables that were less than 0.01, less than 0.05 and less than 0.10.

We highlight three patterns in the test results in Table 3. First, the balancing tests do fairly poorly, even in the case where both the propensity scores and the bias are estimated using the experimental treatment group, as in the response. In both cases, the Hotelling test rejects balance at the 0.005 level. In both cases, the p-value from the regression test is below 0.001 for at least five of the ten variables. In the CPS comparison group, although the 5th largest bias is relatively small, the highest standardized bias lies above 20, the level characterized as large in Rosenbaum and Rubin (1985).

Second, the Hotelling and regression tests suggest imbalance in every case in the table. The maximum standardized biases also suggest concern in most cases, though the 5th largest of the 10 standardized biases often turns out fairly modest. These strong patterns suggest rejecting the propensity score specification that yields the very low biases in Table 2 of Dehejia and Wahba. From this, we conclude not that balancing tests are useless, but rather that, taking the results here in combination with those in Tables 1 and 2, the samples are small and the estimates highly variable, so that even poor scores will sometimes yield low biases and apparent balance.

Finally, we note that in several cases the standardized bias test and the other tests suggest different conclusions about balance. This indicates the value of looking at multiple balancing tests. It also indicates the value of additional research comparing the power of the different tests for particular data generating processes.

5. Conclusions

We believe that the conclusions in our paper continue to hold in light of the evidence presented in the response and in this rejoinder. We add one refinement and one new conclusion.

First, we concluded in our paper that propensity score matching does not solve the selection problem in the NSW data. We believe that the evidence in both the response and our rejoinder serves only to strengthen that conclusion. The low bias estimates presented in DW (1999, 2002) and in the response are quite sensitive, not only in regard to the sample and the propensity score specification as shown in our original paper, but also to whether the treatment group or the control group (or both) is used to estimate the propensity score, to whether the bias is estimated using the experimental treatment group or the experimental control group and, in the case of the CPS comparison group, to how ties are broken in single nearest neighbor matching. Moreover, we show in Section 4 that application of standard balancing tests from the literature does not eliminate this sensitivity.[4]

Second, we add a caution that should be obvious to anyone who looked closely at the estimated standard errors in either the original DW (1999, 2002) papers or our paper, but which we previously emphasized too little. The NSW experimental sample employed in LaLonde (1986) is very small, and gets even smaller when subsets such as the DW sample or the early random assignment sample are considered. The number of useful comparison group observations is also fairly small; the support analyses in both DW (1999, 2002) and our paper make this quite clear. The vast majority of the observations in the CPS comparison sample look nothing like participants in Supported Work. There are more comparable observations in the PSID comparison sample in a relative sense, but that is because it includes a minority over-sample that is not well documented and for which no weights are available. Small sample sizes mean sensitive results, which is exactly what we document in our paper and this response. This does not mean that nothing can be learned from the Supported Work data, only that the results found here should be discounted appropriately and supplemented with findings from larger samples.

Acknowledgements

We thank Dan Black, Michael Lechner, Miana Plesca and Barbara Sianesi for helpful discussions, Barbara Sianesi and Edwin Leuven for quick responses to questions about the psmatch2 program and Prof. Dehejia for information about the empirical analyses in his response. We thank the editors of the special issue, John Ham and Robert LaLonde, for the opportunity to provide this rejoinder and for their comments on an earlier version.

[4] Zhong (2003) finds similar sensitivity in his work using the NSW data.

References

Black, D., Smith, J., 2004. How robust is the evidence on the effects of college quality? evidence from matching. Journal of Econometrics 121, 99–124.

Sianesi, B., 2002. An Evaluation of the Swedish System of Active Labour Market Programmes in the 1990s. Institute for Fiscal Studies Working Paper W02/01.

Zhong, Z., 2003. Data Issues of Using Matching to Estimate Treatment Effects: An Illustration with NSW Data Set. Unpublished manuscript, Peking University.