Empir Econ (2013) 44:47–80
DOI 10.1007/s00181-011-0481-0

Propensity score matching and variations on the balancing test

Wang-Sheng Lee

Received: 28 April 2008 / Accepted: 5 April 2011 / Published online: 27 May 2011

Abstract Balancing tests are diagnostics designed for use with propensity score methods, a widely used non-experimental approach in the evaluation literature. Such tests provide useful information on whether plausible counterfactuals have been created. Currently, multiple balancing tests exist in the literature but it is unclear which is the most useful. This article highlights the poor size properties of commonly employed balancing tests and attempts to shed some light on the link between the results of balancing tests and bias of the evaluation estimator. The simulation results suggest that in scenarios where the conditional independence assumption holds, a permutation version of the balancing test described in Dehejia and Wahba (Rev Econ Stat 84:151–161, 2002) can be useful in applied study. The proposed test has good size properties. In addition, the test appears to have good power for detecting a misspecification in the link function and some power for detecting an omission of relevant non-linear terms involving variables that are included at a lower order.

Keywords Matching · Propensity score · Balancing test · Permutation test · Monte Carlo simulation

JEL Classification C14 · C99

1 Introduction

This article focuses on the issue of how to appraise balance on observed covariates in matching. Matching is a method of sampling from a large reservoir of potential

W.-S. Lee (✉)
School of Economics, Finance and Marketing, RMIT University and IZA, Level 12, 239 Bourke Street, Melbourne, VIC 3000, Australia
e-mail: wangsheng.lee@rmit.edu.au

comparison group members in which the goal is to select a subset of the comparison sample that has covariate values similar to those in the treated group.[1] One can attempt to match on all covariates, but this may be difficult to implement when the set of covariates is large. In order to reduce the dimensionality of the matching problem, Rosenbaum and Rubin (1983) suggested an alternative method which is based on matching on the propensity score $p(X)$. This is defined for each subject as the probability of receiving treatment given the covariate values X and is thus a scalar function of X. When all relevant differences between treatment and comparison group members that affect outcomes are captured in the observed covariates (i.e., potential outcomes are independent of assignment to treatment, conditional on pre-treatment covariates) matching on the propensity score can yield a consistent estimate of the treatment impact. After matching is performed, an evaluation of the similarity in covariates between treatment and comparison group members achieved through matching is often conducted without using any information on the outcome of interest. This is an attempt to mimic the process in experimental studies where balance checks after randomization are often done to see if randomization has been conducted properly.

A recent exchange between Smith and Todd (2005b) and Dehejia (2005a,b) regarding Dehejia and Wahba (1999, 2002) highlights some of the currently unresolved issues regarding the use of propensity score matching estimators. Among them, it is agreed that there is a lack of consensus regarding the utility of balancing tests.

> "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 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." (Smith and Todd 2005b, p. 371)
>
> "… Smith and Todd's observation that there is no consensus on which balancing test to use is useful, and points to the value of ongoing research on this and related topics." (Dehejia 2005b, p. 4)

This article proposes the use of permutation methods for conducting balancing tests. Several authors have considered the use of permutation tests for the related issue of making inferences about the standard error of the treatment effect estimator (Rosenbaum 2002; Hill and Reiter 2006). Recently, Hansen and Bowers (2008) have independently suggested the idea of using permutation tests for appraising balance in covariates in randomized experiments and observational studies (for the case of matching without replacement). Kleyman (2009) also recently suggests interesting logistic regression approaches to balance testing based on using permutation-based likelihood ratios and Bayesian modeling techniques. We demonstrate in Monte Carlo simulations that balancing tests implemented using permutation tests have better

[1] Alternatively, many applied econometricians view matching as an application of non-parametric regression methods to the estimation of treatment effects. Viewed from this perspective, matching is closely related to weighted estimators such as inverse probability weighting (e.g., see Hirano et al. 2003).

size properties than traditional tests. We examine the power of permutation tests in additional simulations to help establish a link between the results of such balancing tests and bias of the evaluation estimator.

In Sect. 2, a broad overview of the utility of balancing tests in matching studies is provided. In Sect. 3, we discuss the difference between a before-matching and after-matching balancing test, a difference that has not been sufficiently highlighted in the literature. Section 4 provides a motivating example using the National Supported Work Demonstration (NSW) data, a well-known data set in the evaluation literature, showing how the use of these different balancing tests can give different results. Section 5 presents test sizes of some commonly used balancing tests. As the simulations in Sect. 5 shows that conventional balancing tests have poor size properties, we introduce non-parametric balancing tests (permutation tests) in Sect. 6 and demonstrate via simulation that they have much better test sizes. Section 7 discusses power properties of the permutation versions of balancing tests. Finally, Sect. 8 concludes.

2 Overview of the utility of balancing tests

An important property of balancing tests as a diagnostic for matching studies is that no information on any outcome is used. Proponents of matching argue that this approach to bias reduction in observational studies that is not available in parametric linear regression models is important for maintaining objectivity of the study. This is because using both the treated and untreated outcomes might lead one to choose specifications that tend to support the intended research hypothesis. However, it has often not been noted that this does not necessarily mean that using *any* information on the outcome is bad. For example, in semi- or non-parametric regression models, information on the untreated outcome is often used to help choose a specification via cross-validation.[2]

There is also a related but separate issue involving having bias or variance reduction as the main objective. If the focus is on variance reduction, one might argue that balancing tests are of limited use. For instance, some non-parametric econometricians who attempt to minimize in some sense the mean square error of the final estimator of interest, instead of just the bias, might be willing to allow some imbalance in the distribution of the covariates if this imbalance reduces variability of the impact estimates without increasing the bias of the impact estimates by too much in any finite sample size (so the estimator remains consistent). In observational studies, however, bias reduction in design is typically regarded as more important than variance reduction (Rubin 2001).

In general, there is basically a consensus that checking for balance is important in matching studies, as evidenced by a discussion of the issue of covariate balance in practically all articles in the applied matching literature. Despite the widespread practice of using hypothesis tests to evaluate balance in the econometric literature (e.g., see Smith and Todd 2005a); Imai et al. (2008) have an alternate view and describe this as a "balancing test fallacy" that should be avoided. They argue that balancing tests should not be used as a stopping rule or test of a formal hypothesis, and that

[2] Iam grateful to Jeff Smith for pointing this out.

balance should simply be maximized.[3] Hansen (2008), however, points out that balance assessments can simply be regarded as goodness-of-fit tests of a certain kind and can be useful if they tend to reject when bias due to inexact propensity matching is enough to undermine causal inferences. As Kleyman (2009) notes, one problem with maximizing balance and not testing its statistical significance is the lack of assurance that even the maximum accomplishable balance on observed covariates is actually good enough for unbiased inference.

An alternative to balance testing as a specification check for propensity score models is an interesting suggestion from Shaikh et al. (2009). They propose a specification test based on whether the unconditional densities of the propensity score for the unmatched treated and comparison group members differ in a certain way.

This article highlights the fact that conventional balancing tests often used in the literature do not necessarily provide the information they are supposed to (i.e., have poor size properties) and suggests the use of permutation tests as an improvement.

3 When should a balancing test be conducted?

The balancing property of propensity scores states that $f(X|D, p(X)) = f(X|p(X))$, where $D = 1$ denotes the treatment group and $D = 0$ the comparison group. Although the balancing property is a statistical statement regarding the conditional distribution, balancing tests in the literature typically focus on the first moment. Just as there are different ways of verifying balance in covariates when conducting an experiment to ensure that randomization was implemented well, there should also be different ways of verifying balance depending on the matching approach employed. For example, when a single-factor experiment is done, checking for balance involves checking for similarity in the covariates between the treatment and control group. In contrast, when a block experiment is done, checking for balance would involve checking for covariate balance within blocks. Likewise, when a matched pair experiment is done, balance involves checking for overall balance between the two groups of matched pairs.[4] The intuition of how balance should be verified from randomized experiments can be carried over to the case of observational studies where matching is used. The point is that different tests for balance are appropriate depending on the type of matching that is performed. In this section, an important distinction is made between a before-matching and an after-matching balancing test.

If stratification on the propensity score is to be performed, the check for balance within each stratum is done after the initial estimation of the propensity score, before examining any outcomes. Rosenbaum and Rubin (1984) and Rubin (1997) suggest a process of cycling between checking for balance on the covariates and reformulating

[3] This echoes a debate about the usefulness of balancing tests in randomized experiments. See, for example, Begg (1990) and Senn (1994).

[4] Single-factor experiments involve having one independent variable that is randomized. Blocking in experimental research involves the random assignment of units to treatment and control groups within strata (blocks) defined by a set of observed pre-treatment covariates. A matched pair experiment involves having participants first matched on variables which are considered to be relevant to the experiment in question before random assignment.

the propensity score. For example, when large mean differences in an important covariate are found to exist between the treatment and comparison groups, even after its inclusion in the model, then the square of the variable and interactions with other variables can be tried. This idea forms the basis of the balancing test algorithm (henceforth the DW test) given in more detail in the appendix of Dehejia and Wahba (2002).

The DW test shares some similarities with other tests in the statistical literature. These tests all involve some partitioning in the "x" or "y" space. For example, the Hosmer and Lemeshow (1980) test is a goodness-of-fit test for logistic regression based on regrouping the data by ordering on the predicted probabilities. Another example is a graphical method for assessing the fit of logistic regression models based on local mean deviance plots, suggested by Landwehr et al. (1984), where groups are created based on a clustering procedure, which partitions the "x" space into homogeneous clusters with approximately identical observations. A common weakness of these tests that involve creating groups is that the results of tests may depend heavily on the grouping strategy.

Another issue with the DW test is the issue of multiple comparisons, which affects the significance level of the test. When a series of tests is made so that a joint decision will be declared correct only if all of its parts are correct (as in the case of testing for mean differences in many covariates in the DW test), one may control the significance level for individual tests, but in so doing the probability of making one or more type 1 error for the entire set of m tests increases with m to $1-(1-\alpha)^m$ for independent tests and less than $1-(1-\alpha)^m$ otherwise. To the best of our knowledge, there have been no formal attempts to address this issue in the context of the DW test.

A careful reading and comparison of Dehejia and Wahba (1999) and Dehejia and Wahba (2002) reveals an important point not yet picked up by the literature—although the DW test is justifiable as a heuristic specification check when stratifying on propensity scores (because balance is checked for within the subclasses of the exact *same* sample in the region of common support to be used for estimating the average treatment effect), it is less appropriate as a specification check for the adequacy of the estimated propensity scores when matching approaches other than stratification are used. This is because the sample changes considerably when matching approaches other than stratification are used. For example, suppose there are n treatment units and nR comparison units (with $R \geq 1$) in the region of common support. Then under stratification, $(n+nR)$ units will be used in the estimation of the treatment effect. On the other hand, with other matching approaches, like nearest neighbor pair matching, for example, because the least similar comparison units are discarded, only $(n+n)$ units are used in the estimation. *It is important to realize that ensuring balance for the full sample in the common support region does not imply balance for the resulting matched sample.*

The point here is that a heuristic specification test that was originally designed for the specific case of stratification on propensity scores is now often inappropriately used as a universal balancing test (for example, it was not appropriately used in Dehejia and Wahba (2002) or Diaz and Handa (2006) because matching methods other than stratification were used). It is important to keep in mind that the propensity score is really a relative measure (it varies depending on the composition of the comparison group) and not some kind of a permanent identification tag for each observation.

Confusion in the literature has arisen because the term "balancing test" has been applied to both the DW test, and to checks for balance in matched samples. In the literature, balancing tests that were conducted before matching (or specification checks) were originally introduced by Rosenbaum and Rubin (1984), and applied, for example, in Rubin (1997) and Dehejia and Wahba (1999, 2002). Tests that were conducted after matching (for example, in Rosenbaum and Rubin 1985) were subsequently also labeled by Smith and Todd 2005a as balancing tests. In their Table 3, for example, they provide results of "balancing tests from single nearest neighbor matching with replacement." This is logically motivated by the fact that we should really be concerned with properties of the matched comparison group, and not necessarily the original or unweighted comparison group. These tests are also somewhat related to the pre-program alignment test suggested by Heckman and Hotz (1989), especially if the set of covariates includes pre-program outcomes, but yet different in nature. Balancing tests are specification tests that assume that the conditional independence assumption (CIA) underlying matching estimators holds.[5] They seek to help identify appropriate flexible ways of conditioning on a particular set of covariates. On the other hand, pre-program tests are tests of identifying restrictions, not specification tests conditional on a particular identifying assumption.

After-matching balancing tests are primarily concerned with the extent to which differences in the covariates in the two groups in the matched sample have been eliminated (assuming balance increases the likelihood of obtaining unbiased treatment effects). If differences still remain, then either the propensity score model should be estimated using a different approach (i.e., fine-tuning the specification of the propensity scores, because the current estimated score might not be an adequate balancing score), or a different matching approach should be used (because for a given data set, covariate differences are removed to a different extent by the different approaches of using the propensity score), or both. This might be difficult to systematically disentangle because of confounding resulting from the many possible combinations of the specification of the propensity score, the choice of the matching algorithm (greedy matching versus optimal matching), matching with or without replacement and matching structure (one-to-one, one-to-k, kernel matching, full matching etc.).

Note that as not all covariates are created equal (e.g., pre-program earnings might be more important than other demographic characteristics in an evaluation of a job-training program), researchers might also consider adjusting the matching approach in order to trade-off balance in one covariate for another. Lechner (1999) provides an illustration of how propensity score matching can include exact matching on some covariates.

4 A motivating example: results for the NSW–PSID data set

Consider the following question: suppose the actual experimental impact of the NSW was kept a secret and the problem of estimating a treatment effect of the NSW was

[5] The CIA is the assumption that treatment assignment is strongly ignorable after conditioning on a set of observed characteristics. It is also often referred to as the unconfoundedness assumption.

posed to several different labor economists. If all of them were restricted to the use of propensity score methods but not restricted to the type of balancing tests they could utilize, would they have obtained similar impact estimates and reached similar conclusions as Dehejia and Wahba (1999)?

For the purposes of this article, we restrict our analysis to the NSW–PSID subsample used in Dehejia and Wahba (1999). The NSW–PSID subsample is a combination of data from the NSW and the panel study of income dynamics (PSID). It consists of the male NSW treatment units and the largest of the three PSID subsamples (see Dehejia and Wahba (1999) for more details). The treatment D is participation in the NSW treatment group. Control variables are Age, Educ (years of education), Black (1 if black, 0 otherwise), Hisp (1 if hispanic, 0 otherwise), Married (1 if married, 0 otherwise), Nodeg (1 if no high school degree, 0 otherwise), RE75 (real earnings in 1975), RE74 (real earnings in 1974), and $U74$ (not employed in 1974).[6] The treatment group contains 185 observations and the comparison group 2,490 observations. To start, we first perform the DW test to check for the specification of the propensity score.

4.1 The DW test

A careful reader would have noted that when using the *same NSW–PSID data set* and implementing propensity score methods, Dehejia and Wahba (1999, 2002) and Dehejia (2005a) have at each instance used a *different specification* of the propensity score. There are therefore at least three specifications that pass this balancing test (and many more not specifically mentioned, as highlighted in Dehejia 2005a). All three specifications pass the DW test, as implemented in the Stata program by Becker and Ichino (2002), but give rise to different common support regions. Perhaps the subtle point to be noted from these different empirical implementations is that there are many possible specifications that can pass the DW test.[7]

In Dehejia and Wahba (1999), the specification used based on the logistic regression model is:

$$\text{prob}(D = 1|X) = F(\text{age}, \text{age}^2, \text{educ}, \text{educ}^2, \text{married}, \text{nodegree}, \text{black}, \text{hisp}, \text{RE74}, \text{RE74}^2, \text{RE75}, \text{RE75}^2, \text{U74*black})$$

[6] Note that RE74 is not literally real earnings in 1974, but actually consists of real earnings in months 13–24 before the month of random assignment. For persons randomly assigned early in the experiment, these months largely overlap with calendar year 1974. For persons randomly assigned later in the experiment, these months largely overlap with 1975. The variable U74 is an indicator variable equal to one when RE74 equals zero.

[7] An interesting suggestion by Diamond and Sekhon (2008) is to use genetic matching, which is an approach for performing multivariate matching that uses an evolutionary search algorithm to determine the weight given to each covariate. Their approach avoids the problem of multiple acceptable specifications by choosing weights to minimize a user-specified balance criterion. Software for implementing this "Genmatch" approach is described in Sekhon (2011). In related study, Graham et al. (2008) propose a flexible parametric variant of inverse probability weighting which they call "inverse probability tilting". Their proposed estimator involves computing weights by finding the solution to a globally concave programming problem with an unrestricted domain. They find that their alternative propensity score estimate increases the efficiency and robustness of inverse probability weighting.

where F is the cumulative logistic distribution. Although Dehejia and Wahba (1999) use the DW test as a diagnostic before employing matching methods, like nearest neighbor matching with replacement, they did not conduct any after-matching balancing tests. As argued in the previous section, such after-matching tests are more relevant as checks for balance than the DW test is when not stratifying on the propensity score because the matched sample is used to estimate the treatment effects.

4.2 The regression test

An alternative before-matching balancing test that has been suggested in the literature is a test that builds on a simple regression model. This test was suggested by Smith and Todd (2005b). The regression test involves regressing each of the conditioning variables in turn on a polynomial in the propensity score and the same polynomial interacted with the treatment indicator.

More formally, for each variable included in estimating the propensity score, the following regression can be estimated:

$$\begin{aligned} X = {} & \beta_0 + \beta_1 p(X) + \beta_2 p(X)^2 + \ldots + \beta_m p(X)^m \\ & + \alpha_0 D + \alpha_1 D p(X) + \alpha_2 D p(X)^2 + \ldots + \alpha_m D p(X)^m + \varepsilon \end{aligned}$$

where m is the order of the polynomial in $p(X)$ that is chosen by the researcher. An F-test is then conducted for the joint null that the coefficients on all of the terms involving the treatment dummy D equal zero. If any of the F-statistics exceeds a conventional critical value (e.g., 5%), it implies that balance has not been attained and a different propensity score specification should be used (just as in the algorithm for the DW test). Put another way, the test checks whether D provides any information on each X, conditional on different levels and non-linearities in the estimated propensity score.

The requirement that the order of the polynomial is chosen by the researcher is one unattractive feature of this test. In practice, the use of a cubic or quartic polynomial has been suggested as such polynomials should be sufficient to capture any potential non-linearities [e.g., see Sanders et al. (2008)]. In this article, for all implementations of the regression test in the following sections, we use a cubic polynomial (i.e., $m = 3$).

From an initial sample of 2,675 observations in the NSW–PSID data set, the common support region based on the $p(X)$ specification given from Dehejia and Wahba (1999) is $n = 1,331$.[8] Results for the before-matching regression test are shown in the first column of Table 1. In contrast to the DW test that found balance in all covariates in the common support region, the regression test finds that two variables—RE74^2 and Black*U74—have P values of the F-statistic which are less than 0.05, indicating imbalance.

In order to perform the after-matching balancing tests for the remainder of this section, we assume the use of the Dehejia and Wahba (1999) specification of $p(X)$.

[8] This is obtained using the *pscore.ado* Stata program written by Becker and Ichino (2002), using five initial blocks and $\alpha = 0.005$ for the t-test.

Table 1 Regression test before and after matching using the DW 1,999 specification

Variable	P value of F test before matching	P value of F test after nearest neighbor matching	P value of F test after kernel matching (Gaussian Kernel)
Age	0.994	0.722	0.696
Age^2	0.987	0.545	0.574
Educ	0.858	0.055	0.220
Educ^2	0.954	0.037	0.300
Married	0.219	0.014	0.100
Nodeg	0.883	0.002	0.863
Black	0.145	0.004	0.021
Hisp	0.832	0.014	0.871
RE74	0.307	0.823	0.982
RE75	0.994	0.925	0.996
RE74^2	0.006	0.809	0.701
RE75^2	0.952	0.815	0.999
Black*U74	0.034	0.000	0.075
n	1,331	240 (=370 when weighted)	1,331 (=370 when weighted)

The regression test is based on using a cubic in the polynomial of the propensity score and imposing common support. The bandwidth used for kernel matching is 0.06

We then apply two matching methods—nearest neighbor matching with replacement and kernel matching (using a Gaussian kernel)—based on this "balanced" specification of $p(X)$ and conduct after-matching balancing tests to determine if the treatment and comparison groups are still balanced after the use of these matching algorithms.[9] The after-matching tests we employ are: (i) the test for standardized differences, (ii) testing for the equality of each covariate mean between groups using t tests, (iii) testing for the joint equality of covariate means between groups using the Hotelling test or F test, and (iv) an after-matching version of the regression test described above that uses the weighted treatment and comparison groups (as opposed to the unweighted treatment and comparison groups in the common support region).

4.3 Standardized differences tests

Recall that the common support region based on the $p(X)$ specification given in Dehejia and Wahba (1999) is $n = 1,331$. After performing nearest neighbor one-to-one matching based on this "balanced" $p(X)$ specification within this common support region, the sample is reduced from $n = 1,331$ to an unweighted $n = 240$ (185 treated and 55 comparison group members), with the unmatched comparison group observations discarded. The weights on the comparison group adjust the n on the

[9] The bandwidth used for kernel matching in our motivating example is 0.06, similar to the bandwidth used in the illustrated example based on the NSW–PSID data set in Becker and Ichino (2002).

matched data set to $n = 370$ (185 treated and 185 weighted comparison group members) so that every treatment observation is paired with a comparison group observation. Similarly, using the estimated propensity scores from Dehejia and Wahba (1999) and performing kernel matching (using the Gaussian kernel), the matched data set has the sample reduced from $n = 1,331$ to a weighted $n = 370$. The difference between nearest neighbor matching and kernel matching is that in the former, unmatched comparison group observations are discarded and given zero weights, with some comparison group observations serving as the counterfactual for more than one treatment observation (so they have weights greater than one); whereas in the latter case, no comparison group members are given a zero weight, with comparison group observations who are more similar to a treatment counterpart given more weight, and comparison group observations who are less similar to a treatment counterpart given less weight.

The test of standardized differences will be used here to illustrate the reduction in bias that can be attributed to matching on $p(X)$. This test was first described in Rosenbaum and Rubin (1985) and checks the balance between the treatment group and the comparison group using a formula for the standardized difference in percentages:

$$B_{\text{before}}(X) = (100)\frac{\bar{X}_{\text{T}} - \bar{X}_{\text{C}}}{\sqrt{\frac{[V_{\text{T}}(X)+V_{\text{C}}(X)]}{2}}} \qquad B_{\text{after}}(X) = (100)\frac{\bar{X}_{\text{TM}} - \bar{X}_{\text{CM}}}{\sqrt{\frac{[V_{\text{T}}(X)+V_{\text{C}}(X)]}{2}}}$$

where for each covariate, $\bar{X}_{\text{T}}$ and $\bar{X}_{\text{C}}$ are the sample means for the full treatment and comparison groups, $\bar{X}_{\text{TM}}$ and $\bar{X}_{\text{CM}}$ are the sample means for the matched treatment and comparison groups, and $V_{\text{T}}(X)$ and $V_{\text{C}}(X)$ are the corresponding sample variances. Intuitively, the standardized difference considers the size of the difference in means of a conditioning variable, scaled by the square root of the variances in the original samples, which allows comparisons in the differences in X before and after matching. It requires defining what a "large" standardized difference is. Rosenbaum and Rubin (1985) suggest that a standardized difference of greater than 20% should be considered as "large."

Before matching, it is evident that there are large differences in the covariates between the treatment and comparison groups in the original sample, and many of the standardized differences have absolute values larger than 100% (Table 2). This is not surprising since we do not expect individuals in the comparison group reservoir to resemble the treatment group in general.

These differences are considerably reduced after nearest neighbor matching, with many of the standardized differences taking on values close to zero. But the variable Hisp that was balanced before matching is now imbalanced. Some other persistent covariate differences remain. The variables Nodeg and Black still have standardized differences larger than 20%, which is an indication that there are some differences in these covariates between the two groups. However, after kernel matching, the results are quite different with most of the differences in covariates removed. None of the standardized differences have absolute values larger than 20%.

Table 2 Test for standardized differences using the DW 1,999 specification

Variable	Standardized difference before matching	Standardized difference after nearest neighbor matching	Standardized difference after kernel matching (Gaussian Kernel)
Age	−100.9	−3.4	4.7
Age^2	−97.1	1.7	5.5
Educ	−68.1	1.0	−6.5
Educ^2	−78.5	1.0	−5.2
Married	−184.2	7.4	6.9
Nodeg	87.9	31.8	9.5
Black	148.0	−20.2	−6.7
Hisp	12.9	25.8	5.4
RE74	−171.8	1.6	−3.2
RE75	−177.4	0.1	−4.1
RE74^2	−85.7	0.9	−0.5
RE75^2	−82.9	0.0	−1.0
Black*U74	163.8	3.0	7.8
n	1, 331	240 (=370 when weighted)	1,331 (=370 when weighted)

The bandwidth used for kernel matching is 0.06

4.4 Test for equality of means before and after matching (t tests)

As in the previous section, after performing nearest neighbor matching and kernel matching based on the values of $p(X)$ that pass the DW test, we conduct the checks for balance based on individual t tests for each covariate used to estimate the propensity score.

Before matching, it is evident that there are large differences in the covariates between the treatment and comparison groups in the original full sample, as all the P values of the test for differences in individual covariate means based on the t test are highly significant (Table 3). After nearest neighbor matching, many of the significant differences disappear. But there are significant differences for the same three variables that the test of standardized differences found differences in the variables—Nodeg, Black, and Hisp. Again, after kernel matching, all of the significant covariate differences disappear.

4.5 Test of joint equality of means in the matched sample (Hotelling test)

Rather than testing for balance in each of the covariates individually, as done in the previous section, we now use a joint test for the equality of means in all the covariates in the $D = 1$ and 0 groups. An F test or Hotelling test can be used for this purpose.

Based on the Hotelling test, which tests for balance in the matched sample after nearest neighbor matching (Table 4, second column), the null of joint equality of

Table 3 *t* Test after matching using the DW 1,999 specification

Variable	*P* value of *t* test before matching	*P* value of *t* test after nearest neighbor matching	*P* value of *t* test after kernel matching (Gaussian Kernel)
Age	0.000	0.677	0.548
Age^2	0.000	0.815	0.429
Educ	0.000	0.897	0.400
$Educ^2$	0.000	0.889	0.457
Married	0.000	0.496	0.525
Nodeg	0.000	0.003	0.367
Black	0.000	0.015	0.465
Hisp	0.053	0.003	0.629
RE74	0.000	0.741	0.535
RE75	0.000	0.968	0.280
$RE74^2$	0.000	0.626	0.879
$RE75^2$	0.000	0.992	0.687
Black*U74	0.000	0.833	0.585
n	1,331	240 (=370 when weighted)	1,331 (=370 when weighted)

The bandwidth used for kernel matching is 0.06

means in the matched sample is rejected, indicating no balance in covariates between the $D = 1$ and 0 groups. However, when the Hotelling test is conducted after kernel matching (Table 4, third column), the null of joint equality of means in the matched sample is not rejected.[10]

4.6 After-matching regression test

An alternative version of the regression test described earlier is to use the weighted treatment and comparison groups. This approach gives more weight to comparison group observations that are more similar to treatment group observations. Results are shown in the second and third columns of Table 1. Once again, after matching, it is found that Dehejia and Wahba (1999) specification does not appear to balance all covariates.

4.7 Summary of balancing tests on the NSW–PSID data

The fact that the different tests give rise to different conclusions regarding balance is a cause for concern, as this could drastically affect the specification of the propen-

[10] As both the *t* test and the Hotelling test assume independence of observations, they might not be appropriate for making inferences when nearest neighbor matching with replacement or kernel matching is performed. In addition, as in Smith and Todd (2005b), the implementation of the Hotelling test here using Stata treats the matching weights as fixed rather than random.

Table 4 Hotelling Test after matching using the DW 1,999 specification

Variable	Mean for $D = 1$	Mean for $D = 0$ (weighted by nearest neighbor matching)	Mean for $D = 0$ (weighted by kernel matching using a Gaussian Kernel)
Age	25.82	26.12	25.40
Age^2	717.39	728.2	683.05
Educ	10.35	10.32	10.52
Educ^2	111.05	110.49	114.03
Married	0.19	0.16	0.16
Nodeg	0.71	0.56	0.66
Black	0.84	0.92	0.87
Hisp	0.06	0.005	0.048
RE74	2095.6	1936.6	2415.37
RE75	1532.1	1518.8	1938.02
RE74^2	28,100,000	22,600,000	31,000,000
RE75^2	12,700,000	12,600,000	19,400,000
Black*U74	0.60	0.59	0.57
Hotelling P value that means are different for the two groups	–	0.000	0.96
n	185	55 (=185 when weighted)	1,146 (=185 when weighted)

The bandwidth used for kernel matching is 0.06

sity score that is used, and hence the final estimate of the average treatment effect that is obtained. Based on using the stratification method, Dehejia and Wahba (1999) estimated the impact of the NSW to be $1,608. When using nearest neighbor matching with replacement, their estimate was $1,691. Although not done in their article, when using kernel matching with a Gaussian kernel, their impact estimate would have been $1,519. All three estimates are close to the experimental benchmark of $1,794 (see their Table 3).[11] The fact that the non-experimental point estimates they obtained were similar to the experimental point estimates helped spur great interest in propensity score methods. However, an interesting question is if they had used any of the after-matching tests, as done in Tables 1, 2, 3, and 4. Would they have rejected their specification and resulting estimate in that case? The contradictory balancing test results for the same data set are the motivation behind the Monte Carlo simulations performed in the next section.[12]

[11] The standard error of the experimental impact was $633.

[12] The results from this example using the specification in Dehejia and Wahba (1999) are not the same for the alternative specification used in Dehejia and Wahba (2002). In addition to passing the DW test, the latter provide after-matching results in their Tables 2 and 3 to show that this alternative specification also passes the after-matching t tests. In general, however, the point is that a propensity score specification that passes the DW test does not have to pass the after-matching balancing tests.

5 Size properties of some commonly used balancing tests

The design of the Monte Carlo experiments in this section investigates the performance of five commonly employed balancing tests when used together with three common ways of using the propensity score. The five balancing tests examined in this article are: (i) the DW test, (ii) the regression test, (iii) the test for standardized differences, (iv) testing for the equality of each covariate mean between matched treatment and comparison groups using t tests, and (v) testing for the joint equality of covariate means between matched treatment and comparison groups using the Hotelling test or F test. The three matching algorithms employed are: (a) propensity score stratification, (b) nearest neighbor matching, and (c) kernel matching. In particular, we simulate the use of the DW test and the before-matching regression test when stratification is done, and the use of the test for standardized differences, the t test, the Hotelling test, and the after-matching regression test after performing nearest neighbor matching and kernel matching.[13] Other matching algorithms (for example, local linear matching, one-to-k matching, full matching etc.) have been suggested in the literature, but we leave the detailed examination of the many other possible combinations to future study.

Suppose the outcome and selection equations can be written as:

$$Y = \alpha_0 + \delta D + \sum_{k=1}^{K} \alpha_k X_k + \varepsilon$$

$$D^* = \beta_0 + \sum_{k=1}^{K} \beta_k X_k + \mu$$

$$D = I(D^* > 0)$$

where δ is the treatment effect, ε and μ are error terms and i.i.d. with zero conditional means (conditioning onX_k), and $I(.)$is the indicator function. The outcome equation is not required for simulations done in this section that check for size properties of balancing tests. However, it will be used later in power simulations to determine the utility of balancing tests in distinguishing between good and bad estimates of the average treatment effect. Note that as this setup involves a homoskedastic outcome equation error and homogeneous treatment effects, the data generating process (DGP) has a finite semiparametric efficiency bound as defined in Hahn (1998). Busso et al. (2009) discuss the importance of this point with regards to simulation work on matching and reweighting estimators in more detail.

[13] Rather than using an arbitrary bandwidth like 0.06 for each replication, from this section onwards, the bandwidth used for kernel matching in each replication in the Monte Carlo simulations is based on Silverman (1986) rule-of-thumb approach. In general, bandwidth choice is a thorny issue for matching methods that rely on kernel regression as there is a bias-variance tradeoff. Too large a bandwidth implies including untreated units quite different from each treated unit in the estimation while too small a bandwidth implies using a few untreated units for each treated unit, with noisy estimates being the result. A recent article by Galdo et al. (2008) examines the issue of bandwidth selection for kernel matching in more detail.

5.1 The DW test

Applied researchers who have experience in using the original DW algorithm might have encountered some difficulty in finding the "optimal" number of intervals. This is partly because the original DW algorithm has no formal rule for setting the test level used to judge balance. The problem with setting the bar too high (i.e., a higher α level) is that this leads to a higher likelihood of rejections, making it harder to find balance. For example, if the bar is set at $\alpha = 0.10$, suppose the observed P value for the between-group difference in a covariate is 0.09. This will be considered to be a "fail" according to the algorithm. On the other hand, if the bar is set at $\alpha = 0.05$, the grade will be changed to a "pass." How should one choose an appropriate test level to gauge balance?

While it is not immediately clear that a Bonferroni adjusted test level should be used instead of an unadjusted test level to find the "optimal" number of intervals and to gauge balance, our experience was that using unadjusted levels is problematic.[14] According to the DW algorithm, a small number of initially unbalanced strata should be subdivided into finer and finer strata and balance retested within those finer strata. One drawback of the DW test is that with too many strata and too few observations in each one, the power of the test can be very low. Extensive simulation work with artificially generated data suggests that there appear to be problems with the original DW test. The results (not shown but available in the working paper version of this article or upon request) suggest that when conventional test sizes are used (e.g., $\alpha = 0.05$ or 0.10), the DW test performs poorly in terms of size and rejects the null much more often than it should. A possible quick-fix to the DW test is to set the bar lower (by using Bonferroni adjusted test levels) when determining the "optimal" number of intervals. This leads to quicker and coarser strata (because balance is easier to find) that would be useful for the second part of the DW test, where covariate differences within each strata are tested for. Simulation results suggest that with the Bonferroni correction made for the test level, the DW test simulations using artificially generated data come much closer to replicating their true sizes.

In order to examine the size properties of the DW test using real world data, we use the NSW–PSID data set and base our test size simulations on the following "true" specification of the propensity score based on the specification in Dehejia and Wahba (1999). That is, the following equation defines the true data generating process under the null hypothesis of a balanced specification,

$$\begin{aligned} D = 1\{&-7.552 + 0.3305\text{age} - 0.0063\text{age}^2 + 0.8248\text{educ} - 0.04832\text{educ}^2 \\ &- 1.8841\text{married} + 0.1299\text{nodegree} + 1.1329\text{black} + 1.9628\text{hisp} - 0.000105\text{RE74} \\ &- 0.000217\text{RE74}^2 + 2.36 \times 10^{-9}\text{RE75} + 1.58 \times 10^{-10}\text{RE75}^2 \\ &+ 2.14\text{U74} * \text{black} + \mu > 0\} \end{aligned}$$

[14] When conducting multiple statistical tests, the chance of finding at least one significant result due to chance fluctuation increases. The Bonferroni correction is one of the more simple and common ways of dealing with multiple comparisons which we later employ in our Monte Carlo simulations. See Schochet (2008) for an accessible and more detailed discussion of the Bonferroni adjustment and other approaches of dealing with multiple testing.

where μ is independent of the covariates and has a logistic distribution with mean 0 and scale parameter 1. Generating a balanced data set under the null to perform the simulations was done as follows. In the binary choice selection equation, because we assume that the error term in the selection equation is independent of the Xs, when we use the error term, along with arbitrary values of β and X to generate D, it is true that: $D \perp X|X\beta$. As only monotonic transformations are performed, it therefore follows that: $D \perp X|\text{logit}(X\beta)$ or $D \perp X|p(X)$. Therefore by construction, the data sets constructed using the actual Xs and the βs from the above equation satisfy the balancing property of propensity scores: $D \perp X|p(X)$.

Although we know the true value of the propensity score, we use the estimated propensity score in our simulations because previous studies (for example, Rosenbaum 1987; Heckman et al. 1997; Hirano et al. 2003) have suggested that the estimated score helps to remove any potential sample imbalances and can lead to better balance.

Unfortunately, it does not appear that finding a simple alternative way of setting test levels is enough to optimize the performance of the DW test. When the DW test with a Bonferroni adjustment is simulated using the real world data, unlike our findings for artificially generated data, based on 2,000 simulations, the test size is not close to the desired size of 5% but is found to be 23.8%.[15] A poor test size implies that many specifications of the propensity score would be rejected, even if they are correct, leading to an unnecessary and prolonged iterative specification search process for an elusive balanced propensity score specification.

5.2 The regression test

Using the same DGP as in the previous section, simulations were conducted to gauge the size properties of the before-matching regression test. The results are shown in the first column of Table 5. Although the average P value of the F test is greater than 0.05 for all variables (reflecting the reduction in covariate imbalance from the initial sample), using the rule where balance is rejected as long as any of the P value for any single variable is less than 0.05 results in a test size of 69.4%. Accounting for multiple comparisons does not appear to help much as even based on Bonferroni adjusted P values (Table 5, last row), the test still rejects the null of balance 37.1% of the time.

5.3 Standardized differences test

When performing the test for standardized differences, for most of the variables, the average standardized difference decreases after nearest neighbor or kernel matching

[15] As a robustness check, the same exercise was repeated using the balanced specifications for the propensity score model given in Dehejia and Wahba (2002) and Dehejia (2005a) to define the data generating process underlying the simulations, where the same NSW–PSID data were used. Similar results (19.4 and 22.9%, respectively) emerged.

Table 5 Monte Carlo results for the regression test before and after matching using the DW 1,999 Specification

Variable	Average P value of F test before matching	Average P value of F test after neighbor matching	Average P value of F test after kernel matching (Gaussian Kernel)
Age	0.731	0.150	0.194
Age^2	0.778	0.173	0.229
Educ	0.689	0.136	0.174
$Educ^2$	0.714	0.135	0.175
Married	0.428	0.311	0.396
Nodeg	0.632	0.147	0.185
Black	0.742	0.279	0.362
Hisp	0.355	0.246	0.293
RE74	0.423	0.376	0.543
RE75	0.211	0.286	0.423
$RE74^2$	0.719	0.366	0.668
$RE75^2$	0.442	0.374	0.653
Black*U74	0.362	0.291	0.327
Decision rule	% of times balance rejected	% of times balance rejected after nearest neighbor matching	% of times balance rejected after kernel matching
Reject if any P value < 0.05	69.4	98.1	95.2
Reject if any P value $< 0.05/13$	37.1	87.5	78.8

Based on 2000 replications. The second critical P value used is a Bonferroni adjusted P value. The regression test is based on using a cubic in the propensity score and imposing common support. The bandwidth used for kernel matching in each replication is based on Silverman (1986) rule-of-thumb approach

has been done. However, using a rule of rejecting balance as long as any variable has a standardized difference greater than 20% (or even 40%) leads to extremely high rejection rates (close to 100%) on the balanced data set (Table 6).

5.4 Test for equality of means before and after matching (t tests)

The after-matching version of the t test seems to perform well in terms of size distortion based on obtaining relatively high average P values after nearest neighbor matching or kernel matching. Once again, however, just as in the case of the before-matching regression test, using the rule of rejecting balance as long as any one variable is imbalanced (whether or not a Bonferroni adjustment is done) leads to high rejection rates (Table 7).

Table 6 Monte Carlo results for the test for standardized differences using the DW 1,999 specification

Variable	Average standardized difference before matching	Average standardized difference after nearest neighbor matching	Average standardized difference after kernel matching (Gaussian Kernel)
Age	−54.35	−34.66	−34.05
Educ	−26.70	−8.92	−10.06
Married	−123.78	−61.88	-61.73
Nodeg	52.07	27.40	28.62
Black	90.16	45.88	45.57
Hisp	0.81	−8.58	−8.01
RE74	−109.31	−64.36	−64.76
RE75	−111.59	−55.94	−56.20
Decision rule		% of times balance rejected after nearest neighbor matching	% of times balance rejected after kernel matching
Reject if any SD > 20		99.9	99.9
Reject if any SD > 40		98.7	98.8

Based on 2000 replications. The bandwidth used for kernel matching in each replication is based on Silverman (1986) rule-of-thumb approach

5.5 Test of joint equality of means in the matched sample (Hotelling test)

The results of the Hotelling test are shown in Table 8. Like the other balancing tests considered so far, it is found that the null of balance is rejected too often.

5.6 After-matching regression test

The after-matching version of the regression test appears to perform even worse than its before-matching counterpart. Simulation results using this test after nearest neighbor matching are displayed in the second column of Table 5, while the results after kernel matching are displayed in the third column of Table 5. In both cases, balance is rejected more than 90% of the time. Although adjusting for multiple comparisons helps somewhat (Table 5, last row), the rejection rates are still over 70%.

5.7 Summary of Monte Carlo results

Given that the data used in the simulations in this section based on real world data are balanced by construction, such frequent rejections of the null of balance indicate that there are size problems with balancing tests that are currently often employed in the literature. These results partly help to explain why different balancing tests can lead to different results (as we saw in Sect. 4). More importantly, the poor test sizes in the simulations suggest that current balancing tests cannot be fully relied upon and that it will be useful to look for balancing tests with better properties. The next two

Table 7 Monte Carlo results for the t test after matching using the DW 1,999 specification

Variable	Average P value of t test before matching	Average P value of t test after nearest neighbor matching	Average P value of t test after kernel matching (Gaussian Kernel)
Age	0.000	0.282	0.299
Age^2	0.000	0.332	0.355
Educ	0.019	0.245	0.258
Educ^2	0.005	0.251	0.263
Married	0.000	0.513	0.551
Nodeg	0.000	0.234	0.252
Black	0.000	0.491	0.517
Hisp	0.663	0.360	0.389
RE74	0.000	0.621	0.699
RE75	0.000	0.529	0.590
RE74^2	0.012	0.572	0.734
RE75^2	0.018	0.589	0.724
Black*U74	0.000	0.568	0.588
Decision rule		% of times balance rejected after nearest neighbor matching	% of times balance rejected after kernel matching
Reject if any P value < 0.05		81.9	77.3
Reject if any P value $< 0.05/13$		52.6	47.6

Based on 2000 replications. The second critical P value used is a Bonferroni adjusted P value. The bandwidth used for kernel matching in each replication is based on Silverman (1986) rule-of-thumb approach

Table 8 Monte Carlo results for the Hotelling test after matching using the DW 1,999 specification

Decision rule	% of times balance rejected after nearest neighbor matching	% of times balance rejected after kernel matching
Reject if P value < 0.05	76.9	64.8
Reject if P value < 0.01	66.0	53.3

Based on 2000 replications. The bandwidth used for kernel matching in each replication is based on Silverman (1986) rule-of-thumb approach

sections take a step in the direction of finding a more reliable balancing test that can be applied in practice. New balancing tests are suggested. After undergoing size and power simulations in Sects. 6 and 7, recommendations for their use in practice are made in Sect. 8.

6 Size properties of permutation versions of balancing tests

This section suggests permutation versions of the balancing tests as a replacement for some of the standard balancing tests described before. In particular, for the DW test, a permutation distribution of the t-statistic is used in place of the t distribution. Similarly, for two of the other parametric after-matching balancing tests based on the t test and the Hotelling test, permutation versions of the tests are proposed and their performance under Monte Carlo simulations examined. Permutation tests are a computer intensive statistical technique that was introduced by Fisher in the 1930s. As Good (2005) points out, the permutation test offers the advantage over the parametric t test in that it is exact even for very small samples and whether or not the observations come from a normal distribution. The parametric t test relies on the existence of a mythical infinite population from which all the observations are drawn. Closely related to the bootstrap (see Kennedy 1995), the main application of permutation tests is the two-sample problem (see Efron and Tibshirani 1993, chapter 15). In the next few sections, we first briefly describe the intuition underlying permutation tests, and then proceed to describing how the DW test, the after-matching t test, and the Hotelling test can be modified using permutation versions of the tests.

6.1 Permutation tests

Suppose that we observe two random samples $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ and $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ drawn from possibly different probability distributions F and G, and that having observed $\mathbf{y}$ and $\mathbf{z}$, we wish to test the null hypothesis of there being no difference between F and G. The null hypothesis can therefore be written as $\mathrm{H}_0 : F = G$. As noted earlier, this essentially is similar to the problem of checking for covariate balance between two groups (the additional complication being that it is not just one but many covariates for which balance needs to be tested). In practice, the t test for two independent samples is the most commonly employed test in this situation because it is reasonably robust and easy to compute. Researchers using such a test check the observed t-statistic against the critical value in the t distribution to determine whether the observed group difference is significant. In contrast, in permutation resampling, the test distribution is not assumed to follow the t distribution and is instead generated by randomly shuffling the group labels a large number of times.

A similar approach has been used by Abadie (2002) in the context of the Kolmogorov–Smirnov test statistic, where he proposes a bootstrap method to overcome the low power of the Kolmogorov–Smirnov test in the presence of point masses. Suppose a sample of size n consists of n_t observations in the treatment group and n_c observations in the control group. Abadie suggests a bootstrap approach that involves resampling observations with replacement, labeling the first n_t observations as treatment group members and the remaining n_c observations as control group members, and using the two generated samples to compute the test statistic. This procedure is repeated a large number of times. This essentially is what a permutation test involves, the key difference being that in permutation resampling, resampling is done without replacement. Abadie (2002) notes in his "summary and discussion of possible extensions" section

Table 9 Monte Carlo results on test sizes for permutation versions of the balancing tests using the DW 1,999 specification

Decision rule	% of times balance rejected after stratification on the propensity score	% of times balance rejected after nearest neighbor matching	% of times balance rejected after kernel matching
Permuted DW Test			
Reject if any P value < 0.05	84.5	–	–
Reject if any Bonferroni adjusted P value < 0.05	3.5	–	–
Permuted t test			
Reject if any P value < 0.05	–	20.0	17.0
Reject if any Bonferroni adjusted P value < 0.05	–	1.5	2.0
Permuted Hotelling Test			
Reject if P value < 0.05	–	1.0	0.0

Based on 200 replications. The bandwidth used for kernel matching in each replication is based on Silverman (1986) rule-of-thumb approach

that permutation versions of the bootstrap tests he proposes are equally valid and have the advantage that by construction, they provide exact levels in finite samples. Below, we discuss in more detail how we implement the permutation versions of the DW test, as well as the permutation versions of the after-matching t test and Hotelling test.

6.2 The permutation version of the DW test

We modify the DW algorithm in three respects: (i) correcting for multiple testing using the Bonferroni adjustment,[16] (ii) using a permutation version of the t test in selecting the "optimal" number of intervals to use, and (iii) using a permutation version of the t test in testing for the equality of covariates within each interval. As we saw in Sect. 5.1, only correcting for multiple comparisons in the DW test is not enough to ensure that the test has the correct size. Similarly, only using permutation tests on their own for choosing the appropriate number of blocks and for verifying balance within each block is also insufficient to allow the DW test to have the correct size. As can be seen in the first row of Table 9, using a permutation version of the DW test without correcting for multiple comparisons leads to balance being rejected 84.5% of the time.

The modified DW algorithm is given in the box below. Based on 200 simulations of the DW 1,999 specification, where 1,000 random permutations of the t test were performed, this modified DW test achieved a reasonable test size of 3.5% (see Table 9).[17]

[16] As Schochet (2008) highlights, many modified and sometimes more powerful versions of the Bonferroni adjustment have been developed. Modifying the DW test using these alternative corrections will be an interesting avenue for future research.

[17] Note that the test size simulations here do not involve modifying the logit specification (step 5) as the correct logit specification is assumed to be known and used in the simulations.

Algorithm for the Permutation Version of the DW Test

1. Start with a parsimonious logit specification to estimate the score.
2. Split the sample in k equally spaced intervals of the propensity score. For example, using $k = 5$ and dividing observations into strata of equal score range (0–0.2,…, 0.8–1). This is usually done over the region of common support.
3. Within each interval, use a permuted t test (with at least 1,000 random permutations) to test at the Bonferroni adjusted level α/k that the mean $p(X)$ values for treated and comparison units do not differ. If the test fails, split the interval in half and test again. The "optimal" number of intervals k^* is found when the mean $p(X)$ values for treated and comparison units do not differ in all intervals.
4. Let v be the number of covariates. Then within each interval, use a permuted t test (with at least 1,000 random permutations) and the optimal number of intervals k^* from step 3 to test at the Bonferroni adjusted level α/k^*v that for all covariates, the mean differences between treated and comparison units are not significantly different from zero.
5. If covariates are balanced between treated and comparison observations for all intervals, stop. If covariates in any interval are not balanced (i.e., we are using the maximal t-statistic), modify the logit by using a less parsimonious specification (i.e., adding interaction terms and/or higher-order terms of the covariate) and reevaluate.

6.3 The permutation versions of the after-matching t-test and Hotelling test

The after-matching t test is a standard two sample t test that uses the matched treated and comparison group units. As the standard t test did not fare well in the Monte Carlo simulations in Sect. 4, we modify the test by using the permutation version of the t test. In addition, to account for multiple testing, we use the Bonferroni adjustment to divide the chosen test level by the number of variables.

The after-matching Hotelling test is the multivariate analog of the t test, where instead of testing for the equality in covariate means between matched treated and comparison units one at a time, the joint equality of all covariate means is tested. As for the t test, a permutation version of the Hotelling test can be used, since the standard Hotelling distribution did not give good size levels in the simulations in Sect. 5. However, unlike the t test, because it tests for a joint null hypothesis, no correction for multiple testing needs to be done.

Using 1,000 random permutations of the t test and Hotelling test, we see in Table 9 that much more reasonable test sizes are achieved. Using a Bonferroni adjusted test size, the t test rejects the null of balance only 1.5% of the time under nearest neighbor matching, and only 2.0% of the time under kernel matching. The fact that these rejection rates are all less than the nominal size of 5% could be due to the conservative nature of the Bonferroni adjustment. The corresponding rejection rates for the Hotelling test were 1.0 and 0.0%.

7 Power of the permutation tests and bias on impacts

The previous section showed via simulations that the permutation versions of the DW test, t test, and Hotelling test have much better test sizes using the NSW–PSID data than their parametric versions. We next turn to performing simulations to illustrate the link between balance/imbalance in covariates (without knowledge of the outcome) and the subsequent bias in estimates of average treatment effects. This allows us to determine the utility of balancing tests. Is imbalance in covariates closely related to large biases in the evaluation estimator? Conversely, is balance in covariates related to small biases in the evaluation estimator? As the true DGP is unknown in observational studies, power results (when we model $p(X)$ incorrectly) are potentially more important in practice than the size results, as tests with good sizes but low power can be of limited use.

The focus in this section is on the results of Monte Carlo simulations where the CIA holds. This is because where it does not hold (e.g., an omitted variable problem), the estimated average treatment effect on the treated (ATT) will be biased and the results of a balancing test are of little importance. The CIA underlying matching estimators is often thought of in terms of including the correct set of covariates to estimate the propensity score. But the version of the CIA based on the propensity score (as opposed to the covariates X) can be viewed as also requiring the correct choice of link function and index function in modeling the propensity score.

Using a propensity score model comprised of two covariates from artificially generated data, Zhao (2008) finds that misspecification of the propensity score model is relatively unimportant when the CIA holds in the sense that the poorly estimated coefficients have little influence on the estimated ATT. We expand this line of inquiry further in this section by examining if the proposed permutation tests have good sizes and high power using real world data. In particular, we perform two sets of simulations, one exploring misspecifications in the link function, and the other exploring misspecifications in the index function.

7.1 Misspecifications in the link function

In our first set of DGPs, we generate the true propensity score using a heteroskedastic probit model to allow us to control the misspecification of the link function by varying degrees. Assume the true DGP for treatment assignment is:

$$\begin{aligned} D = 1\{&-2.4869 + 0.1616\text{age} - 0.0031\text{age}^2 - 0.005\text{educ} + 0.3744\text{nodegree} \\ &-0.9630\text{married} + 1.2285\text{black} + 1.219\text{hisp} - 0.00005\text{RE74} \\ &-0.0001\text{RE75} + \upsilon > 0\} \end{aligned}$$

where υ is independent of the covariates. While a standard probit model assumes that υ has a standard normal distribution, the heteroskedastic probit model generalizes the standard probit model by assuming that the variance of the error term is no longer fixed at one but allowed to vary as a function of the independent variables. Following Harvey (1976), assume that the variance of υ has a multiplicative functional form

$$\text{Var}(\upsilon) = \sigma^2 = \exp(z\gamma)^2,$$

where z is a vector of covariates that define groups with different error variances in the underlying latent variable, and γ is a vector of parameters to be estimated. This is a convenient functional form because if all the elements of γ are equal to zero, then $e^0 = 1$ and the heteroskedastic probit model is reduced to a standard probit model.

Similarly, assume the true DGP for the outcome variable is given by:

$$\begin{aligned} Y = &-2872.058+1000\text{D}+106.30\text{age} - 2.62\text{age}^2+586.52\text{educ}+629.18\text{nodegree} \\ &+975.46\text{married}-518.28\text{black}+2248.67\text{hisp}+0.2757\text{RE74}+0.5659\text{RE75}+\varepsilon, \end{aligned}$$

where ε is $N(0, 300)$. The coefficients for these equations are based on the actual coefficients from the NSW–PSID data set, using the originally assigned treatment group values and earnings in 1978 (RE78) as the outcome. The true treatment effect is set at \$1,000.

For each matching algorithm, four scenarios are used to determine how results of balance and imbalance in covariates from permutation tests relate to bias in the estimated ATT. These involve generating the data assuming that the true DGP is a heteroskedastic probit model (while varying a parameter that captures the degree of heteroskedasticity) but estimating the propensity score using a logit model (with the exact same set of variables used in the true DGP). In particular, suppose that:

$$\begin{aligned} z\gamma = c\Big[&0.1616\text{age}-0.0031\text{age}^2-0.0056\text{educ} + 0.3744\text{nodegree}-0.9630\text{married} \\ &+1.2285\text{black} + 1.219\text{hisp} - 0.00005\text{RE74} - 0.0001\text{RE75}\Big] \end{aligned}$$

and that we use values of $c = 0$,1,2, and 4. Figure 1 shows how heteroskedasticity can lead to substantial misspecification in the estimated propensity score. Based on a single simulation, the figures in the left panel of Fig. 1 compares the densities of the true propensity score versus the estimated propensity score, while in the right panel show an alternative way the extent to which the propensity score is estimated incorrectly (they should be on a 45° line if the estimated propensity score equals the true propensity score). Note that when $c = 0$, the model is a standard probit model. As the graphs for the standard probit and standard logit models are rather similar, the extent to which the graph with $c = 0$ is different from the graphs when $c \neq 0$ can be used as an indication of the severity of model misspecification. Recently, Koenker and Yoon (2009) emphasize the importance of the link function in binary choice models. They highlight that an incorrect choice of a link function may result in misleading propensity scores and consequently misleading estimates of treatment effects.

For all simulations performed in this section, the estimated ATT is obtained by taking the mean difference $(Y|D = 1) - (Y|D = 0)$ averaged over 200 replications. When $c = 0$, we see the effect of using a logit model when true model is a probit. As expected, as the two link functions are very similar in distribution, there is low power for detecting the incorrect use of the logit model. Furthermore, bias in the ATT is low regardless of the matching algorithm used (Table 10).

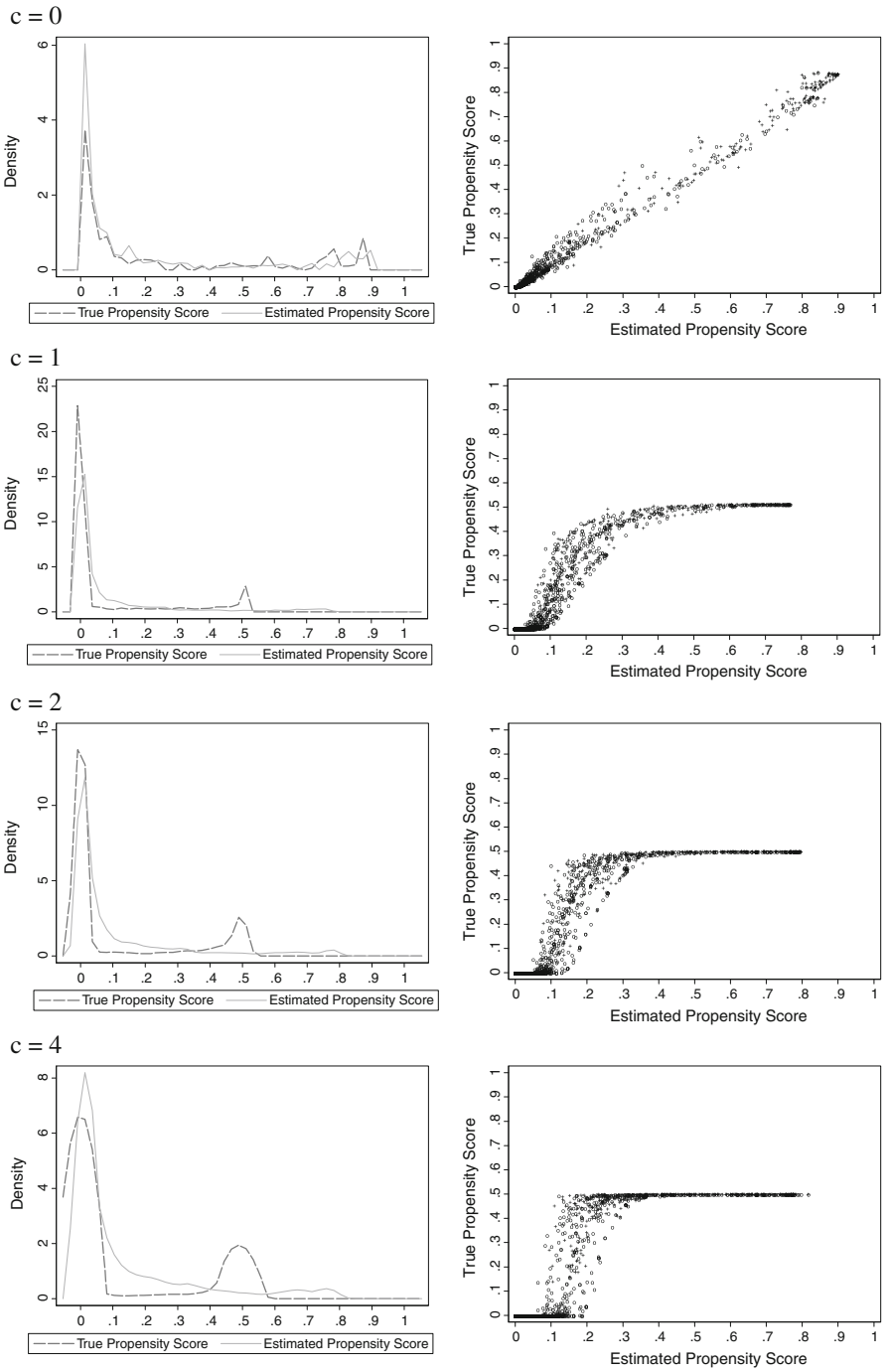

Fig. 1 Four scenarios for the power simulations (see Table 10 for the corresponding results). In the *rightpanel*, "+" denotes treatment observations and "o" denotes control observations

Table 10 Analysis of power and bias for permutation versions of the balancing tests: Misspecification of the link function

No balancing test done			Using specifications that are balanced			
	Bias	RMSE		% of times balance rejected	Bias when balanced	RMSE when balanced
Stratification			Stratification (permuted DW test)			
$c = 0$	−76.41	180.11	$c = 0$	2.0	−75.84	180.80
$c = 1$	184.69	214.36	$c = 1$	16.0	192.13	217.75
$c = 2$	271.23	296.99	$c = 2$	60.0	290.93	318.11
$c = 4$	481.08	493.87	$c = 4$	71.0	509.78	521.54
Nearest neighbor matching			Nearest neighbor matching (permuted t test)			
$c = 0$	90.42	291.88	$c = 0$	0	90.42	291.88
$c = 1$	325.30	408.66	$c = 1$	1.0	320.85	402.37
$c = 2$	381.38	457.39	$c = 2$	3.0	375.30	448.22
$c = 4$	424.70	501.15	$c = 4$	0.5	422.65	498.92
			Nearest neighbor matching (permuted Hotelling test)			
			$c = 0$	0	90.42	291.88
			$c = 1$	0	325.30	408.66
			$c = 2$	1.5	378.80	454.65
			$c = 4$	0	424.70	501.15

Table 10 continued

No balancing test done			Using specifications that are balanced			
	Bias	RMSE		% of times balance rejected	Bias when balanced	RMSE when balanced
Kernel matching			Kernel matching (Permuted t test)			
$c = 0$	52.02	225.47	$c = 0$	0	52.02	225.47
$c = 1$	250.84	272.96	$c = 1$	0	250.84	272.96
$c = 2$	309.32	326.72	$c = 2$	0.5	308.55	325.90
$c = 4$	477.49	488.18	$c = 4$	0.5	476.79	487.44
			Kernel matching (permuted Hotelling test)			
			$c = 0$	0	52.02	225.47
			$c = 1$	0	250.84	272.96
			$c = 2$	0	309.32	326.72
			$c = 4$	0	477.49	488.18

Based on 200 replications with 1,000 permutations of the test statistic. The sample is restricted to those in the region of common support. The true treatment effect is $1,000. The true model is a probit with varying degrees of heteroskedasticity (reflected by the value of c). The estimated model in all instances is a logit. The bandwidth used for kernel matching in each replication is based on Silverman (1986) rule-of-thumb approach. Tests are conducted at the 5% level—the DW test and t test use Bonferroni adjusted P values

When $c \neq 0$, we measure the power of the permutation tests to detect more severe misspecification in the estimated propensity score. The permutation version of the DW test appears to have good power to detect increasing deviations from the true propensity score. As heteroskedasticity in the true model increases (i.e., values of c deviate further away from zero) so that the difference between the true and estimated model becomes more stark, the null of balance is rejected substantially more often. This is an important result because the increase in the rejection of balance corresponds to an increase in the bias and root mean square error (RMSE). In this way, the test can help weed out more biased estimates of the ATT by indicating an imbalance in covariates.

It is important to keep in mind, however, that finding balance in covariates does not necessarily imply that the ATT one obtains is less biased than the ATT one obtains without conducting any balancing tests. For example, in Table 10, for the case of stratification when $c = 4$, it can be seen that both bias and RMSE are higher using specifications that pass the balancing test as opposed to the scenario when no balancing test was conducted. The value of the balancing test in this case is that it increases the probability of detecting a possible functional form misspecification in the propensity score.

On the other hand, it appears that permutation versions of the t test and the Hotelling test have much less power in detecting increasing heteroskedasticity in the true model. For example, it is clear in Table 10 that the bias in the ATT can be substantial when $c \neq 0$ but these balancing tests provide no indication that this could be occurring. It is quite puzzling why the permutation versions of the t test and the Hotelling test perform so badly. One possible explanation is that in nearest neighbor matching with replacement and kernel matching, observations are re-used in some instances. Although distributional assumptions are relaxed with the use of permutation versions of these tests, independence of observations is still a required assumption. It might be useful to investigate in future research how these tests perform under settings where the independence of observations assumption might be more plausible.

7.2 Misspecifications in the index function

In our second set of DGPs, we focus on scenarios where the choice of link function is not an issue, but where we are interested in determining the appropriate set of non-linear terms to include in the propensity score model. The following six DGPs are considered.

In DGPs 1–3, the true specification of $p(X)$ is the Dehejia and Wahba (1999) logit specification. We estimate $p(X)$ using the correct logit link function, but fail to include some non-linear terms involving variables that are included at a lower order. In DGP 1, we incorrectly omit the U74*black term. In DGP 2, we incorrectly omit both U74*black and RE74^2. In DGP 3, we incorrectly omit RE74^2, RE75^2, educ^2, and U74*black from the propensity score model.

In DGPs 4–6, we estimate $p(X)$ in each case using the Dehejia and Wahba (1999) logit specification, but now, we have the reverse case where the true specification does not include certain non-linear terms. In DGP 4, the true DGP does not include the U74*black term. In DGP 5, the true DGP does not include both U74*black and

RE74^2. Finally, in DGP 6, RE74^2, RE75^2, educ^2, and U74*black are not included in the true DGP. Put another way, DGPs 4–6 consider scenarios where the estimated propensity score model is increasingly over-specified in some way.

When relevant non-linear terms are omitted from the estimated propensity score model (DGPs 1–3), bias can be quite high and Table 11 shows that the permutation version of the DW test has some power to detect such misspecification. For example, in DGP 1 when the U74*black term is omitted incorrectly, the null of balance is rejected 17% of the time. On the other hand, when irrelevant non-linear terms are included in the estimated propensity score model (DGPs 4–6), the permutation version of the DW test has not much power to detect the misspecification. From examining the estimates of bias in Table 11, as with the case of detecting misspecifications in the link function, the permutation versions of the *t* test and the Hotelling test do not appear to have much power in detecting misspecifications in the index function.

It is interesting to note that when the estimated propensity score model is over-specified (DGPs 4–6), bias and RMSE are in all cases relatively lower than when the model is under-specified (DGPs 1–3), similar to the findings in Millimet and Tchernis (2009).

8 Conclusions and recommendations for practice

This article was motivated by Smith and Todd's (2005b) observation that multiple versions of the balancing test exist with little known about their properties. Based on simulations that use the NSW–PSID data that is well known in the evaluation literature, the finding that commonly employed balancing tests have poor size properties suggests a need for an overhaul of balancing tests. This finding is perhaps not surprising in light of the fact that parametric tests are not recommended for use in the parallel context of determining if estimated average treatment on the treated impacts from matching (which use the exact same weights from matching) are statistically significant.

In this article, permutation tests were proposed as an approach for conducting balancing tests. As any iterative adjustment process involving balancing tests occurs independently of the outcome, it should have no systematic effect on the outcome. However, to date, little or no study has been done on establishing a link between balancing tests and bias of the evaluation estimator.

In practice, applied researchers never know what the true specification of the propensity score is, even if they might have a rich data set which arguably captures most of the variables that could make the CIA plausible. Even in the case when the correct set of covariates is included for matching estimators to be identified, with misspecification of the link function or index function, simulations in this article using real world data show that bias in the ATT can be large.

This article shows that the permutation version of the DW test has good size properties. In addition, the test appears to have good power for detecting a misspecification in the link function and some power for detecting an omission of relevant non-linear terms involving variables that are included at a lower order. It is considerably less powerful with regards to detecting the inclusion of irrelevant non-linear terms, but this might not be a serious weakness as there is some evidence that there is less of a

Table 11 Analysis of power and bias for permutation versions of the balancing tests: Misspecification of the index function

No balancing test done			Using specifications that are balanced			
	Bias	RMSE		% of times balance rejected	Bias when balanced	RMSE when balanced
Stratification			Stratification (Permuted DW test)			
DGP 1	333.69	423.52	DGP 1	17.0	325.52	420.91
DGP 2	347.45	425.27	DGP 2	17.0	340.81	417.92
DGP 3	248.17	376.91	DGP 3	15.0	234.71	371.89
DGP 4	-101.41	212.13	DGP 4	2.0	-104.62	212.57
DGP 5	-95.64	205.94	DGP 5	4.0	-96.04	204.73
DGP 6	-89.60	209.09	DGP 6	2.5	-86.64	207.50
Nearest neighbor matching			Nearest neighbor matching (Permuted *t* test)			
DGP 1	455.17	572.04	DGP 1	2.0	454.40	573.04
DGP 2	426.32	560.44	DGP 2	1.5	423.26	558.78
DGP 3	356.32	551.59	DGP 3	1.0	351.89	546.55
DGP 4	99.19	286.57	DGP 4	0	99.19	286.57
DGP 5	62.76	262.16	DGP 5	0	62.76	262.16
DGP 6	27.31	252.28	DGP 6	0.5	25.86	251.92
			Nearest neighbor matching (Permuted Hotelling test)			
			DGP 1	0	455.17	572.04
			DGP 2	0	426.32	560.44
			DGP 3	0	356.32	551.59
			DGP 4	0	99.19	286.57
			DGP 5	0	62.76	262.16
			DGP 6	0	27.31	252.28

Table 11 continued

No balancing test done			Using specifications that are balanced			
	Bias	RMSE		% of times balance rejected	Bias when balanced	RMSE when balanced
Kernel matching			Kernel matching (permuted *t* test)			
DGP 1	456.84	557.56	DGP 1	0.5	455.38	556.44
DGP 2	428.94	529.95	DGP 2	0	428.94	529.95
DGP 3	358.03	530.09	DGP 3	0	358.03	530.09
DGP 4	47.41	238.25	DGP 4	0	47.41	238.25
DGP 5	3.21	206.34	DGP 5	0.5	22.42	206.47
DGP 6	0.72	228.60	DGP 6	0	0.72	228.60
			Kernel matching (permuted Hotelling test)			
			DGP 1	0	456.84	557.56
			DGP 2	0	428.94	529.95
			DGP 3	0	358.03	530.09
			DGP 4	0	47.41	238.25
			DGP 5	0	23.21	206.34
			DGP 6	0	0.72	228.60

$DGP1$ U74*black relevant but omitted, $DGP2$ U74*black, $RE74^2$ relevant but omitted, $DGP3$ U74*black, $RE74^2$ $RE75^2$, $educ^2$relevant but omitted, $DGP4$ U74*black included but irrelevant, $DGP5$ U74*black, $RE74^2$ included but irrelevant, $DGP6$ U74*black, $RE74^2$ $RE75^2$, $educ^2$included but irrelevant

Based on 200 replications with 1,000 permutations of the test statistic. The sample is restricted to those in the region of common support. The true treatment effect is $1,000. The estimated model in all instances is a logit. The bandwidth used for kernel matching in each replication is based on Silverman (1986) rule-of-thumb approach. Tests are conducted at the 5% level—the DW test and *t* test use Bonferroni adjusted *P* values

penalty associated with over-specifying the propensity score model. The evidence in this article suggests against a reliance on parametric versions of the DW test, the regression test, the test for standardized differences, t tests, and Hotelling tests, especially if these tests are used in conjunction with a rule-of-thumb rejection rule. The findings therefore suggest that for future applied study involving propensity score matching, even if other matching algorithms are used to estimate the ATT, it might be useful to estimate the ATT based on stratification as a robustness check. This is because assuming one has a rich enough set of covariates to fulfill the key assumption underlying matching estimators, simulation results suggest that the permutation version of the DW test can be useful by indicating an imbalance in covariates if there is a misspecification in the link function, or if there are relevant higher-order (and/or interaction) terms that should be in the model (i.e., a guard against under-specification).

Given that there exist other balancing tests in the literature for matching estimators not considered in this article, and that the Monte Carlo simulations for the proposed test are limited to a single real world data set due to finite time constraints, future research on assessing the properties of other balancing tests and on further understanding the properties of the proposed test will clearly be useful.

Acknowledgments Many thanks to the editor Jeff Smith and two anonymous referees for their constructive and detailed comments that greatly improved the article, to Jim Powell and Chris Skeels for very useful discussions, to Jeff Borland, Denzil Fiebig, Ben Hansen, Michael Lechner, Chris Ryan, Yi-Ping Tseng for helpful comments, and to conference participants at the 2005 PhD Conference in Economics and Business (Perth), the 2006 Australasian Meeting of the Econometric Society (Alice Springs) and the 2006 ZEW Conference on Policy Evaluation (Mannheim) for comments. This article is based on a chapter from my PhD dissertation and was partially supported by a University of Melbourne Faculty Grant. All errors are my own.

References

Abadie A (2002) Bootstrap tests for distributional treatment effects in instrumental variable models. J Am Stat Assoc 97:284–292

Becker S, Ichino A (2002) Estimation of average treatment effects based on propensity scores. Stata J 2:358–377

Begg C (1990) Significance tests of covariate imbalance in clinical trials. Controlled Clin Trials 11:223–225

Busso M, DiNardo J, McCrary J (2009) New evidence on the finite sample properties of propensity score matching and reweighting estimators. IZA Discussion Paper No. 3998

Dehejia R (2005a) Practical propensity score matching: a reply to Smith and Todd. J Econom 125:355–364

Dehejia R (2005b) Does matching overcome Lalonde's critique of non-experimental estimators? A Postscript. Manuscript

Dehejia R, Wahba S (1999) Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. J Am Stat Assoc 94:1053–1062

Dehejia R, Wahba S (2002) Propensity score matching methods for nonexperimental causal studies. Rev Econ Stat 84:151–161

Diamond A, Sekhon J (2008) Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Unpublished working paper

Diaz J, Handa S (2006) An assessment of propensity score matching as a non experimental impact estimator: evidence from Mexico's PROGRESA program. J Hum Resour 41:319–345

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman and Hall, New York

Galdo J, Smith J, Black D (2008) Bandwidth selection and the estimation of treatment effects with unbalanced data. Ann d'Econ et Stat 91(92):189–216

Graham B, Pinto C, Egel D (2008) Inverse probability tilting for moment condition models with missing data. NBER Working Paper No. 13981
Good P (2005) Permutation, parametric, and bootstrap tests of hypotheses. Springer, New York
Hahn J (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66:315–331
Hansen B (2008) The essential role of balance tests in propensity-matched observational studies: comments on "a critical appraisal of propensity-score matching in the medical literature Between 1996 and 2003" by Peter Austin, Statistics in Medicine. Stat Med 27:2050–2054
Hansen B, Bowers J (2008) Covariate balance in simple, stratified and clustered comparative studies. Stat Sci 23:219–235
Harvey A (1976) Estimating regression models with multiplicative heteroscedasticity. Econometrica 44:461–465
Heckman J, Hotz J (1989) Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. J Am Stat Assoc 84:862–874
Heckman J, Ichimura H, Todd P (1997) Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Rev Econ Stud 64:605–654
Hirano K, Imbens G, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71:1161–1189
Hill J, Reiter J (2006) Interval estimation for treatment effects using propensity score matching. Stat Med 25:2230–2256
Hosmer D, Lemeshow S (1980) Goodness of fit tests for the multiple logistic regression model. Commun Stat Theory Methods A9:1043–1069
Imai K, King G, Stuart E (2008) Misunderstandings among experimentalists and observationalists about causal inference. J Royal Stat Soc A 171:481–502
Kennedy P (1995) Randomization tests in economics. J Bus Econ Stat 13:85–94
Kleyman Y (2009) Testing for covariate balance in observational studies. Ph.D Dissertation, Department of Statistics, University of Michigan
Koenker R, Yoon J (2009) Parametric links for binary choice models: a Fisherian-Bayesian Colloquy. J Econom 152:120–130
Landwehr J, Pregibon D, Shoemaker A (1984) Graphical methods for assessing logistic regression models. J Am Stat Assoc 79:61–71
Lechner M (1999) Earnings and employment effects of continuous on-the-job training in East Germany after unification. J Bus Econ Stat 17:74–90
Millimet D, Tchernis R (2009) On the specification of propensity scores: with applications to the analysis of trade policies. J Bus Econ Stat 27:397–415
Rosenbaum P (1987) Model-based direct adjustment. J Am Stat Assoc 82:387–394
Rosenbaum P (2002) Observational studies. Springer, New York
Rosenbaum P, Rubin D (1983) The central pole of the propensity score in observational studies for causal effects. Biometrika 70:41–55
Rosenbaum P, Rubin D (1984) Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 79:516–524
Rosenbaum P, Rubin D (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 3:33–38
Rubin D (1997) Estimating causal effects from large data sets using propensity scores. Ann Intern Med 127:757–763
Rubin D (2001) Using propensity scores to help design observational studies: application to the Tobacco Litigation. Health Serv Outcomes Res Methodol 2:169–188
Sanders S, Smith J, Zhang Y (2008) Teenage childbearing and maternal school outcomes: evidence from matching. Unpublished working paper
Schochet P (2008) Technical methods report: Guidelines for multiple testing in impact evaluations. Report prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences
Sekhon J (2011) Multivariate and propensity score matching software with automated balance optimization: the matching package for R. J Stat Softw 42:1–52
Senn S (1994) Testing for baseline balance in clinical trials. Stat Med 13:1715–1726
Shaikh A, Simonsen M, Vytlacil E, Yildiz N (2009) A specification test for the propensity score using its distribution conditional on participation. J Econom 151:33–46

Silverman B (1986) Density estimation for statistics and data analysis. Chapman & Hall, London
Smith J, Todd P (2005a) Does matching overcome Lalonde's critique of nonexperimental estimators. J Econom 125:305–353
Smith J, Todd P (2005b) Rejoinder. J Econom 125:365–375
Zhao Z (2008) Sensitivity of propensity score methods to the specifications. Econ Lett 98:309–319