Propensity Score Matching and Variations on the Balancing Test

Wang-Sheng Lee*

Melbourne Institute of Applied Economic and Social Research
The University of Melbourne

First version: November 3, 2005
This version: August 11, 2006

Abstract

This paper focuses on the role of balancing tests when employing propensity score matching methods. The idea behind these tests are to check to see if observations with the same propensity score have the same distribution of observable covariates independent of treatment status. Currently, multiple versions of the balancing test exist in the literature. One troubling aspect is that different balancing tests sometimes yield different answers. This paper highlights the importance of distinguishing between balancing tests that are conducted before matching and after matching, and provides a Monte Carlo examination of four commonly employed balancing tests. We highlight the poor size properties of these commonly employed balancing tests and demonstrate how non-parametric versions of before and after matching tests provide much better test sizes. Finally, we illustrate how balancing tests are of little utility if the conditional independence assumption underlying matching estimators is not fulfilled.

Key words: Matching; Propensity score; Monte Carlo simulation.

*Melbourne Institute of Applied Economic and Social Research, The University of Melbourne, Level 7, 161 Barry Street, Carlton, Victoria 3010, Australia. Thanks to Jeff Borland, Denzil Fiebig, Chris Ryan and Yi-Ping Tseng for comments and Jim Powell and Chris Skeels for very useful discussions. This research was partially supported by a University of Melbourne Faculty Seeding Grant. All errors are my own.

Propensity Score Matching and Variations on the Balancing Test

1. Introduction

Recent papers by Dehejia and Wahba (1999, 2002) have generated great interest in the economics profession regarding the ability of propensity score matching methods to potentially produce unbiased estimates of a social program's impact, for example, when estimating the effect of a job training program or disability program. Matching is a method of sampling from a large reservoir of potential controls in which the goal is to select a subset of the control sample that has covariate values similar to those in the treated group. 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 thus a scalar function of *X*. As actual propensity scores are not known, the first step in a propensity score analysis is to estimate the individual scores, and there are various ways to do this in practice. The most common approach is to use logistic or probit regression, although other methods like neural nets might be employed. When all relevant differences between treatment and comparison group members that affect outcomes are captured in the observable covariates (i.e., outcomes are independent of assignment to treatment, conditional on pre-treatment covariates) matching on the propensity score can yield an unbiased estimate of the treatment impact.

Faith in using econometric methods to evaluate social programs had weakened in the 1980s because Lalonde (1986) had found that a range of standard non-experimental evaluation estimators produced impact estimates that were very different from the experimental benchmark estimates. As a result, there was a marked shift in the U.S. towards a preference for using experimental designs for evaluating social programs (for example, the plethora of experimental studies evaluating welfare reform in the U.S. in the 1990s). However, Lalonde had not considered propensity score methods. The striking result of Dehejia and Wahba (1999, 2002) was that utilising the same National Supported Work Demonstration (NSW) data set as Lalonde, they provided empirical evidence that challenged Lalonde's conclusion. Dehejia and Wahba (1999, 2002) found that using propensity score methods, they could come close to replicating the experimental benchmark results.

However, questions have been raised regarding the robustness of this finding. Smith and Todd (2005a) argue that the Dehejia and Wahba (1999, 2002) results are sensitive to the choice of subsample of the Lalonde data employed in the estimation. They also point out that the Dehejia and Wahba (1999, 2002) finding is somewhat surprising in light of the lessons learnt from the analyses of Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998) based on their analyses of the U.S. National Job Training Partnership Act study. Those studies concluded that in order for matching estimators to have low bias, it is necessary to have a rich set of variables related to program participation and labour market outcomes, that the non-experimental comparison group be drawn from the same local labour markets as the participants, and that the dependent variable be measured in the same way for participants and non-participants. None of these conditions hold in the NSW data set analysed by Lalonde (1986) and Dehejia and Wahba (1999, 2002).

A more recent exchange between Smith and Todd (2005b) and Dehejia (2005a, 2005b) 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.[1]

> "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)

One version of the balancing test used in Dehejia and Wahba (1999, 2002) gained prominence because their analysis seemed to suggest that as long as this diagnostic tool does not reject the propensity score specification employed, then we might reasonably expect to obtain an unbiased impact estimate. This is because in their work, by showing how their specification of the propensity score passed this balancing test (and not any other diagnostic), they were able to replicate the experimental benchmark impact estimates of the NSW.

Consider the following situation: suppose the actual experimental impact of the NSW had been kept a secret and the problem of estimating a treatment effect of the NSW was posed to several different labour 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 utilise, would they have obtained similar impact estimates and reached similar conclusions as Dehejia and Wahba (1999, 2002)?

The main objective of this paper is to make a more detailed study of some of the more commonly used balancing tests in the literature so that we might better understand whether they are important, when they might work well, and when they might not work so well. There have been several other recent efforts that address the issue of determining balance when using matching estimators. Hansen (2006) discusses an omnibus measure of balance that is a weighted sum of squares of differences of means of cluster means. Imai, King and Stuart (2006) argue that hypothesis tests should not be used as a balance stopping rule and suggest a more general approach based on quantile-quantile (QQ) plots. Sekhon (2006) and Diamond and Sekhon (2005) stress the importance of testing balance in several different ways as no measure is equally sensitive to departures from balance. They discuss how a matching algorithm called Genetic Matching can maximise a variety of balance measures ranging from *t*-tests to QQ plots.

In section 2, we set the scene by describing the theory and intuition behind propensity score methods. In section 3, we discuss the difference between a before matching and after matching balancing

[1] We define precisely what a balancing test is in the next section.

test, a difference that has not been sufficiently highlighted in the literature. Several balancing tests are examined in detail in this paper: (i) the balancing test used in Dehejia and Wahba (1999, 2002) that are based on testing for mean differences within strata of the propensity score, (ii) the test for standardised differences (Rosenbaum and Rubin 1985), (iii) testing for the equality of each covariate mean between treatment and comparison groups using *t*-tests (Rosenbaum and Rubin 1985), and (iv) testing for the joint equality of covariate means between treatment and comparison groups using the Hotelling test or *F*-test (Smith and Todd 2005b). In section 4, we provide a motivating example using the NSW data, showing how the use of these different balancing tests can give different results. In section 5, we perform Monte Carlo simulations using generated data. In section 6, we perform similar simulations using real world data. As the simulations in sections 5 and 6 show that conventional balancing tests have poor size properties, we introduce non-parametric balancing tests (permutation tests) in section 7 and demonstrate via simulation that they have much better test sizes. Power simulations are also performed. Finally, section 8 concludes.

2. Theory and Intuition Behind Propensity Scores

The basic idea of propensity score matching is an attempt in a non-experimental context to replicate the setup of a randomised experiment. In order to make clear the conceptual differences between an experiment, covariate matching, and propensity score matching, we briefly discuss each in turn.

2.1 Randomised Studies

In an experiment, for all observable and unobservable covariates X_{All} , we have

$$D \perp X_{All}$$

where $\perp$ denotes independence, and D is the treatment group indicator, with $D = 1$ indicating the treatment group, and $D = 0$ the comparison group. In addition,

$$D \perp Y(0), Y(1)$$

where $Y(0)$ is the outcome in the untreated state, and $Y(1)$ is the outcome in the treated state. This implies that

$$f(Y(0) \mid D=1) = f(Y(0) \mid D=0)$$

where f denotes the density function. This allows us to estimate the average treatment effect: $E[Y(1) \mid D=1] - E[Y(0) \mid D=1]$ using $E[Y(1) \mid D=1] - E[Y(0) \mid D=0]$. Such a design is known in the experimental design literature as a single-factor experiment where all observations are classified into either the treatment or control group. Social experiments most often employ the single-factor experimental design. Because of randomisation, the true propensity score is a constant (and identically 0.5 if the treatment-control ratio is 50-50). A check to determine if randomisation was properly done can be performed by comparing the overall treatment and control covariate distributions.

A closely related design is the randomised complete block design. Blocking is the grouping of subjects into homogeneous groups (homogeneous with respect to some characteristic that contributes to the variability of the outcome variable). Broadly, the idea is that in the analysis, we will be able to subtract out the contribution of the block effect to the outcome variable, thus reducing variability. Stated more precisely, such a design allows one to attribute some of the variability to the characteristics that was blocked on. In a block experiment, a situation is created where the treatment groups are randomised within subclasses defined by the blocking variables B. That is,

$$D \perp X_{All} \mid B$$

In addition,

$$D \perp Y(0), Y(1) \mid B$$

This implies that

$$f(Y(0) \mid D=1, B) = f(Y(0) \mid D=0, B)$$

This allows us to estimate the average treatment effect of the full sample

$$E[Y(1) \mid D=1] - E[Y(0) \mid D=1]$$

using the sum of the average treatment effect in each block,

$$E[Y(1) \mid D=1, B] - E[Y(0) \mid D=0, B].$$

2.2 Covariate Matching

We can attempt to recreate the ideal of balanced treatment groups in an observational study. This requires making the assumption of strongly ignorable treatment allocation (SITA) based on an observable set of covariates *X*, a subset of X_{All} .

$$D \perp Y(0), Y(1) \mid X$$

In this case, we can estimate the average treatment effect: $E[Y(1) \mid D=1] - E[Y(0) \mid D=1]$ using $E[Y(1) \mid D=1, X] - E[Y(0) \mid D=0, X]$.[2] Let the potential outcomes be written as follows.

$$Y_i(1) = f_1(X_i) + \nu_i$$
$$Y_i(0) = f_0(X_i) + \omega_i$$

[2] In addition, it is standard to make the additional Stable Unit Treatment Value (SUTVA) assumption. This requires that the treatment status of any unit be independent of potential outcomes for all other units, and that treatment is defined identically for all units.

where v_i and ω_i are independent error terms. The basic ideas of covariate matching are

$$X_i = X_j \quad \Rightarrow \quad f_t(X_i) = f_t(X_j), \quad t = 0,\ 1$$

and

$$d(X_i, X_j) < \varepsilon \quad \Rightarrow \quad d'(f_t(X_i),\ f_t(X_j)) < \delta, \quad t = 0,\ 1$$

where d and d' are some distance metrics. The former justifies exact matching, while the latter shows how the continuity of f_t justifies neighbourhood matching when exact matching is infeasible (see Zhao 2004 for more discussion).

2.3 Propensity Score Matching

The introduction of propensity scores by Rosenbaum and Rubin (1983) was primarily an extension of Cochran (1968), who considered the problem of comparing the means of an outcome *Y* in two groups, and the problem of bias because *Y* is related to a variable *X* whose distribution differs in the two groups. Cochran showed that adjustment by subclassification was a useful device for removing bias. Initially, the groups are divided into a few subclasses based on the distribution of *X*. Next, the mean value of *Y* is calculated separately within each subclass. Finally, a weighted mean of these subclass means is calculated for each group, using the same weights for each group, where the weights are proportional to the number of subjects in the subgroup. With 5 subclasses, Cochran showed that approximately a 90 percent reduction in bias could be obtained.

Cochran had only considered adjustments on a single variable *X*. However, a major problem with subclassification is that as the number of covariates increases, the number of subclasses grows dramatically. For example, considering only binary covariates, with *k* variables, there will be 2^k subclasses, and it is highly unlikely that every subclass will contain both treated and comparison units. Subsequently, Rosenbaum and Rubin (1983) presented a useful extension allowing adjustment by subclassification for multivariate *X*. The central idea there is to replace multivariate *X* with a scalar function of *X* called the propensity score because it gives the probability of being in the treatment versus comparison group at each value of *X*.[3] Assuming SITA holds, matching on the propensity score, defined as

$$p(X) = prob(D = 1 \mid X)$$

leads to the following two conditions:

$$D \perp Y(0), Y(1) \mid p(X)$$

and

[3] Although propensity scores can be used in several different ways (for example, for pair matching or covariance adjustment, both mentioned in Rosenbaum and Rubin (1983)), the intellectual link between Cochran's work on stratification and propensity scores is made very clear in Rubin (1997).

$$D \perp X \mid p(X)$$

The first statement is SITA conditional on the propensity score, and is a direct result of assuming SITA.[4] It is also often referred to as the Conditional Independence Assumption (CIA) in the propensity score literature. The second relation is the balancing property of propensity scores. As compared to covariate matching, propensity score matching avoids the problem of the curse of dimensionality when *X* is high dimensional.

The framework underlying propensity score methods is closely related to the framework underlying a randomised experiment. When subclasses have similar or close values of *p(X)*, theorem 2 of Rosenbaum and Rubin (1983) shows that *X* has the same distribution of $D = 1$ and $D = 0$ units in each subclass. The idea behind propensity score stratification is that by assembling groups of $D = 1$ and $D = 0$ units that are similar with respect to *X* within subclasses, we are trying to reconstruct a series of completely randomised experiments with changing probabilities of treatment assignment.[5] It is reasonable to expect that observations with the same propensity scores should have the same distribution of observable covariates. Adjustment on the propensity score is in this case sufficient to produce unbiased estimates of the average treatment effect.[6] Examples of the method of stratification on propensity scores are Rosenbaum and Rubin (1983, 1984), Rubin (1997), and Dehejia and Wahba (1999).

Despite similarities in the ideas between propensity score matching and covariate matching, the theory behind propensity score matching is quite different from that behind covariate matching (Zhao 2004). The basic ideas of propensity score matching are:

$$f(X_i \mid D_i{=}1, p(X_i) = p) = f(X_i \mid D_i{=}0, p(X_i) = p) = f(X_i \mid p)$$

and

$$d(p_k, p_l) < \varepsilon \;\; \Rightarrow \;\; d'(f(X_i \mid p_k),\; f(X_j \mid p_l)) < \delta$$

The former says that when matching is exact at the propensity score *p*, then the distribution of *X* will be the same for the treated sample and the comparison sample at *p*. The latter equation states that if exact matching on *p* is impossible and instead matching is on some neighbourhood of *p*, then the distribution of *X* is still approximately the same for the treated sample within the neighbourhood of *p*.

[4] Theorem 3 in Rosenbaum and Rubin (1983) shows that if treatment assignment is strongly ignorable given *X*, then it is strongly ignorable given any balancing score. A balancing score *b(X)* is a function of the observed covariates *X* such that the conditional distribution of *X* given *b(X)* is the same for the treatment and comparison units. The most trivial balancing score is *b(X)* = *X*. More interesting balancing scores are many-to-one functions of *X*. By definition, *X* is the finest balancing score, whereas the propensity score, with dimension equal to one, is the coarsest balancing score.

[5] This method is also sometimes referred to as blocking on the propensity score or as subclassification on the propensity score.

[6] However, in practice, subclasses will generally not be homogeneous in *p(X)*, so the directly adjusted estimate may contain some residual bias due to *X*.

It is also worth noting that propensity score matching differs from a single-factor randomised experiment in two important ways. First, a randomised experiment balances the distributions of both observables and unobservables between treated and control samples, but propensity score matching only balances the observables. Second, a single-factor randomised experiment balances the distributions for the whole sample, but propensity score matching balances the distribution at each individual propensity score value. In other words, the estimate of propensity score matching can be thought of as a weighted average of the estimates from many miniature randomised experiments (at different p's). Put another way, the subclasses of propensity scores can be thought of as recreating a randomised block experiment, where there are a series of completely randomised experiments with different propensities.

The overall quality of the estimation depends on the quality of each of these miniature experiments. Just as a substantial sample size is needed to obtain a meaningful estimate from a single-factor randomised experiment, a sufficiently large sample size is required at each p in order to obtain a meaningful propensity score matching estimate.[7] Stratifying on the propensity score in effect returns to the template of the randomised block experiment, where the blocking variable B is now $p(X)$. Alternatively, pairs of treatment-control units can be created that are matched on the propensity scores, thereby recreating a paired comparison experiment (Rubin 2004). Matched pairs experiments are a special case of randomised block experiments, where the size of the block is two.[8]

The propensity score plays a critical role in capturing the assignment mechanism, similar to the seminal selection model of Heckman (1979). However, a key difference is that such conditional choice probability models deal with the probability of treatment assignment based on the principle of control functions (Heckman and Robb 1986) and use the probability in a different way. In particular, they focus on exclusion restrictions and do not focus on creating balance in the observed covariates.

The primary purpose of the propensity score is that it serves as a balancing score. Consequently, *the idea behind balancing tests is to check whether the propensity score is an adequate balancing score,* that is, to check to see if at each value of the propensity score, X has the same distribution for the treatment and comparison groups. More formally, we are interested in verifying if:

$$D \perp X \mid p(X) \qquad (1)$$

[7] Propensity score methods generally work better in larger samples. This is because the distributional balance of observed covariates created by subclassifying on the propensity score is an expected balance, just as the balance of all covariates in a randomized experiment is an expected balance. In a small randomized experiment, random imbalances of some covariates can be substantial despite randomization. Analogously, in a small non-experimental study, substantial imbalances of some covariates may be unavoidable despite subclassification using a sensibly estimated propensity score. Based on Monte Carlo simulations, Zhao (2004) found that propensity score methods were not superior to covariate matching for small sample sizes (n = 500), but performed better for larger sample sizes (n = 1,000, n = 2,000).

[8] Other matching weights developed in the econometric literature, like kernel matching and local linear matching, have less of a direct parallel with the experimental design literature.

where X is a set of covariates that are chosen to fulfil the CIA.[9] The basic intuition is that after conditioning on *p(X)*, additional conditioning on X should not provide new information on D. The propensity scores themselves serve only as devices to balance the observed distribution of covariates between the treated and comparison groups. The success of propensity score estimation is therefore assessed by the resultant balance rather than by the fit of the models used to create the estimated propensity scores. Given that propensity score methods are typically used to estimate some kind of a treatment effect, *balancing tests are really a means to an end*, and can be considered useful only if passing a balancing test leads to more unbiased treatment effect estimates. This should be the yardstick by which balancing tests are ultimately assessed. For example, if passing a particular balancing test is related to obtaining more biased treatment effect estimates, then it is clear that such a balancing test is best avoided.

3. When Should a Balancing Test be Conducted?

A property of conditional independence relations is that $D \perp X \mid p(X) = X \perp D \mid p(X)$ and the latter implies that:

$$f(X \mid D, p(X)) = f(X \mid p(X))$$

In other words, after conditioning on *p(X)*, if (1) is true, then:

$$f(X \mid D=1) = f(X \mid D=0) \qquad (2)$$

Conceptually, verifying balance involves checking if (2) holds, usually after invoking the common support assumption.[10] Just as there are different ways of verifying balance in covariates when conducting an experiment to ensure that randomisation 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. The intuition of how balance should be verified from randomised 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

[9] It is important to distinguish the CIA from the balancing property of propensity scores. One does not imply the other. For example, it is possible to obtain balance for samples of data where the CIA is valid or where it does not hold. The simplest case is when X is a univariate variable, where it is clear that the CIA does not hold and where it is very easy to attain balance. Similarly, even if the CIA is fulfilled, the balancing property might not hold because *p(X)* could be an inadequate balancing score, perhaps because the functional form of X is not represented correctly when estimating *p(X)*. See Smith and Todd (2005a) for further clarification.

[10] There are various ways of defining common support. One method, used in Dehejia and Wahba (1999, 2002), is based on discarding the D = 0 observations with *p(X)* lower than the minimum of the D = 1 observations and discarding the D = 0 observations with *p(X)* higher than the D = 1 observations. Another is based on the notion of trimming (Smith and Todd 2005a) where the region of common support is defined as those values of *p(X)* that have a positive density within the D = 1 and D = 0 observations. We define common support using the former in this paper.

depending on the type of matching that is performed. In this section, an important distinction is made between a before matching balancing test 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. If important within-stratum differences are found on some covariates, then either the propensity score model needs to be reformulated or it would be concluded that the covariate distributions do not overlap sufficiently to allow subclassification to adjust for these covariates. Because of the curse of dimensionality and the difficulty in finding exact matches with more than a few covariates, instead of comparing estimates of the full multidimensional densities, researchers usually examine various low dimensional summaries of each variable in X, for example, mean differences. If a low dimensional summary differs between the $D = 0$ and $D = 1$ groups, then (2) probably does not hold. Note, however, that even if many low dimensional summaries are the same for the treatment and comparison group, we still cannot be certain that (2) holds because these summaries do not test for overall distributional equality of X between the two groups.

Rosenbaum and Rubin (1984) and Rubin (1997) suggest a process of recycling between checking for balance on the covariates and reformulating 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 is also the basis of choosing the specification of the propensity score in Dehejia and Wahba (1999). The algorithm behind this so-called balancing test (henceforth the DW test – although it is really more a specification test for the propensity score) is given in more detail in the appendix of Dehejia and Wahba (2002).[11] A key advantage of the matching approach, as opposed to model-based methods, is that outcome data is not involved so repeated analyses attempting to balance covariates do not bias estimates of the treatment effect on outcome variables. The intuition behind this check for balance within strata is the close analogy between randomized block experiments and propensity score methods. As described in Dehejia and Wahba (2002), it is suggested that in practice, one could start with k groups based on equally spaced intervals of the estimated propensity score. Groups are further split into halves if it is found that the average propensity scores of treated and control units differ within each group.[12]

The DW test is, however, not a completely new invention and has some close relatives in the statistical literature. These tests all involve some partitioning in the '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. In the implementation of the Hosmer and Lemeshow test, groups

[11] The check for balance is usually done in the region of common support for X. Interestingly, if the model for participation is predicted very well, the $D = 1$ and $D = 0$ observations might have very little overlap. This intuitively tells us that the best predictor score is not necessarily a useful propensity score if it does not serve as an adequate summary score of X.

[12] The Stata ado program *pscore* written by Becker and Ichino (2002) has an algorithm for implementing the DW test.

are defined either by using *g* fixed cutpoints in the interval (0, 1) or by using *g* equal sized groups. Typically, *g* is chosen to be 10 in both instances. Another example is a graphical method based on local mean deviance plots for logistic regression models, suggested by Landwehr, Pregibon, and Shoemaker (1984). Their method is based on a partition of the deviance into a pure-error component and a lack of fit component using clusters of neighbouring points.

A weakness of these tests that involve creating groups is that the results of tests may depend heavily on the choice of cutpoints. The version of the Hosmer and Lemeshow test that defines groups based on fixed cutpoints of the predicted probabilities could lead to groups which have zero or small expected values when the estimated probabilities are clustered into a few intervals. This often leads to large values of the test statistic, leading to a corresponding over rejection of the null. On the other hand, the version of the test based on equal sized groups may lead to over rejection of the null because of small estimated expected frequencies in the lower and upper deciles of the predicted probability (Hosmer, Lemeshow and Klar 1988).

Another issue with the DW test is the issue of multiple comparisons, which affects the significance level of the test. To the best of our knowledge, there have been no formal attempts to address this issue in the context of the DW test. Appendix A describes the intuition behind the Bonferroni correction – one of the more simple and common ways of dealing with multiple comparisons – that we later employ in our Monte Carlo simulations. See Westfall and Young (1993) for other approaches of dealing with multiple testing.

We are unaware of any work that has focused on estimating the statistical properties of the DW test. Given the rather ad hoc nature the number of groups is chosen, and the issue of multiple comparisons, this issue is important if the DW test is to be routinely used in applied work. Sections 5 and 6 present Monte Carlo simulations of the DW test and reveal its shortcomings. In section 7, we suggest a modification to the DW test which we show attains much better test sizes.

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 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 *n*R comparison units (with R ≥ 1). With stratification, assuming no observations are removed due to a lack of common support for the sake of argument, (*n* + *n*R) units will be used in the estimation of the treatment effect. On the other hand, with other matching approaches, like nearest neighbour 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 realise that ensuring balance for the full sample 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 balancing test (for example, it was appropriately used in Dehejia and Wahba (1999) because stratification on $p(X)$ was employed, but not in Dehejia and Wahba (2002) because other matching methods other than stratification were used). Checking for balance in the full sample is not critical because it is a different matched (or weighted) sample that is being used to estimate the treatment effect. 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 in Rubin (1997), Dehejia and Wahba (1999, 2002), Michalopoulos, Bloom and Hill (2004), and Dehejia (2005a). Tests that were conducted after matching were subsequently also labelled by Smith and Todd (2005a) as balancing tests. In their Table 3, for example, they provide results of “balancing tests from single nearest neighbour 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. Such a view of balancing tests has been picked up by others in the applied literature, for example, Ham, Li, and Reagan (2003), and Ho, Imai, King, and Stuart (2005). These tests are closely related to the pre-program alignment test suggested by Heckman and Hotz (1989), where the focus is comparing differences in pre-program outcomes between the treatment and comparison groups. Here, the after matching balancing tests go one step further in comparing differences in time-invariant covariates (that are known to be not affected by the treatment) between the treatment and comparison groups.

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.[13] It is of course also possible that no amount of adjustment can lead to balance on the matched samples, where it might be necessary to conclude that propensity score matching methods cannot solve the selection problem.

[13] 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.).

In summary, the literature has suggested various ways of checking for (1). However, the loose interpretation of what conditioning on *p(X)* means has led researchers to not distinguish between before matching and after matching balancing tests. A before matching balance test is really more of a specification test when used in conjunction with the method of stratification on the propensity score. Here, the interpretation of conditioning on *p(X)* is to use intervals of *p(X)*. It should be distinguished from an after matching balance test, which is really a check to see if a matched comparison group can be considered to represent a plausible counterfactual. In this case, conditioning on *p(X)* refers to weighting on *p(X)*.

4. A Motivating Example: Results for the NSW-PSID data set

We return to the question posed in the introduction in this section: would a re-analysis of the NSW-PSID data used in Dehejia and Wahba (1999, 2002) and Smith and Todd (2005a) by many different analysts lead to very different results?[14]

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

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

where *F* is the cumulative logistic distribution. In contrast, in Dehejia and Wahba (2002), the specification used is:

$$prob(D=1\mid X)=F(\text{age, age}^2\text{, educ, educ}^2\text{, married, nodegree, black, hisp, RE74, RE74}^2\text{,}\\ \text{RE75, RE75}^2\text{, U74, U75, U74*hisp})$$

Finally, in Dehejia (2005a), the specification used is:

$$prob(D=1\mid X)=F(\text{age, educ, married, black, hisp, RE74, RE75, married*U75, nodegree*U74})$$

[14] The PSID data, or Panel Study of Income Dynamics data, are based on a nationally representative U.S. longitudinal survey and was used by Lalonde (1986) to construct one of his comparison groups for the treated individuals in the NSW. Diamond and Sekhon (2005) have also recently reanalysed the NSW data and revisited the results of Dehejia and Wahba (1999, 2002), but do not focus on the issue of balancing tests.

All three specifications pass the DW test, as implemented in the Stata program by Becker and Ichino, but give rise to different common support regions.[15] Perhaps the subtle point Dehejia and Wahba are trying to make is that there are many possible specifications that can pass the DW test.

Although Dehejia and Wahba (1999, 2002) use the DW test as a diagnostic prior to employing matching methods, like nearest neighbour 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. 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 $p(X)$. We then apply two matching methods – nearest neighbour 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. The after matching tests we employ are: (i) the test for standardised differences, (ii) testing for the equality of each covariate mean between groups using t-tests, and (iii) testing for the joint equality of covariate means between groups using the Hotelling test or F-test.

4.2 Standardised Test of Differences

The common support region based on the $p(X)$ specification given above from Dehejia and Wahba (1999) is n = 1331. After performing nearest neighbour one-to-one matching based on this 'balanced' $p(X)$ specification within this common support region, the sample is reduced from n = 1331 to an unweighted n = 242 (185 treated and 57 comparison group members), with the unmatched comparison group observations discarded. The weights on the comparison group adjust the n on the 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 = 1331 to a weighted n = 370. The difference between nearest neighbour matching and kernel matching are that in the former, unmatched comparison group observations 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). 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.[16]

[15] When using a t-test level of α = 0.005, from an initial sample size of n = 2,675 (with n = 185 for D = 1 and n = 2,490 for D = 0), the first specification gives rise to a common support of n = 1,331 observations, the second has n = 1,243, and the third has n = 1,458.

[16] If an Epanechnikov kernel is used instead of the Gaussian kernel, some comparison group observations not within a certain radius (specified by the researcher) of a treated observation's propensity score will be discarded.

The test of standardised 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 standardised difference:

$$B_{before}(X) = 100.\frac{\overline{X}_T - \overline{X}_C}{\sqrt{\frac{[V_T(X)+V_C(X)]}{2}}} \qquad B_{after}(X) = 100.\frac{\overline{X}_{TM} - \overline{X}_{CM}}{\sqrt{\frac{[V_T(X)+V_C(X)]}{2}}}$$

where for each covariate, $\overline{X}_T$ and $\overline{X}_C$ are the sample means for the full treatment and comparison groups, $\overline{X}_{TM}$ and $\overline{X}_{CM}$ are the sample means for the matched treatment and comparison groups, and $V_T(X)$ and $V_C(X)$ 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" standardised difference is. Rosenbaum and Rubin (1985) suggest that a standardised difference of > 20 should be considered as "large."

Table 1: Test for Standardised Differences

Variable	Standardised Difference Before Matching	Standardised Difference After Nearest Neighbour Matching	Standardised Difference After Kernel Matching (Gaussian kernel)
Age	-100.9	-3.5	5.0
Educ	-68.1	1.5	-7.3
Married	-184.2	7.4	6.2
Nodeg	87.9	**31.8**	9.2
Black	147.9	**-20.2**	-6.2
Hisp	12.9	**25.8**	5.0
RE74	-171.8	2.2	-4.3
RE75	-177.4	0.04	-6.5
n	2675	242 (= 370 when weighted)	1331 (= 370 when weighted)

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 standardised differences have absolute values larger than 100 (Table 1). 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 neighbour matching, with many of the standardised 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 standardised 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 standardised differences have absolute values larger than 20.

4.3 Test for Equality of Means Before and After Matching (t-tests)

As in the previous section, after performing nearest neighbour 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. After nearest neighbour matching, many of the significant differences disappear. But there are significant differences for the same three variables that the test of standardised differences found differences in – the variables *Nodeg*, *Black* and *Hisp*. Again, after kernel matching, all of the significant covariate differences disappear.

Table 2: *t*-Test After Matching

Variable	*p*-Value of *t*-Test Before Matching	*p*-Value of *t*-Test After Nearest Neighbour Matching	*p*-Value of *t*-Test After Kernel Matching (Gaussian kernel)
Age	0.000	0.666	0.549
Age^2	0.000	0.804	0.429
Educ	0.000	0.856	0.401
Educ^2	0.000	0.834	0.457
Married	0.000	0.496	0.526
Nodeg	0.000	**0.003**	0.368
Black	0.000	**0.015**	0.466
Hisp	0.053	**0.003**	0.629
RE74	0.000	0.630	0.536
RE75	0.000	0.991	0.281
RE74^2	0.000	0.340	0.879
RE75^2	0.000	0.958	0.687
*Black*U74*	0.000	0.833	0.586
n	2675	242 (= 370 when weighted)	1331 (= 370 when weighted)

4.4 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 $D = 0$ groups. A *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 neighbour matching (Table 3, second column), the null of joint equality of means in the matched sample is rejected, indicating no balance in covariates between the $D = 1$ and $D = 0$ groups. (However, when the difficult to balance variables *Nodeg*, *Black* and *Hisp* are removed and not used in the joint test, the Hotelling test does not reject the null that the means in the two groups are equal). When the Hotelling test is conducted after kernel matching (Table 3, third column), the null of joint equality of means in the matched sample is not rejected.

Table 3: Hotelling Test After Matching

Variable	Mean for $D = 1$	Mean for $D = 0$ (weighted by nearest neighbour matching)	Mean for $D = 0$ (weighted by kernel matching using Gaussian kernel)
Age	25.82	26.13	25.40
Age^2	717.39	728.8	683.05
Educ	10.35	10.31	10.52
Educ^2	111.05	110.21	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	1873.1	2415.37
RE75	1532.1	1528.5	1938.02
RE74^2	28100000	18800000	31000000
RE75^2	12700000	12900000	19400000
*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	57 (= 185 when weighted)	1146 (= 185 when weighted)

4.5 Summary of Balancing Tests on the NSW-PSID Data

The summary of the four balancing tests we conducted on the NSW-PSID data is given in Table 4. The first test was based on a sample of n = 1331 (the common support region of n = 2675, the full sample) while the second to fourth tests were conducted on the nearest neighbour with replacement matched sample (weighted n = 370) and the kernel matched sample (weighted n = 370). 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 propensity score that is used, and hence the final estimate that is obtained.

Table 4: Summary of Balancing Test Results

Test	Result
DW specification test	Pass
Standardised differences test (nearest neighbour)	Fail
Standardised differences test (kernel)	Pass
t-test (nearest neighbour)	Fail
t-test (kernel)	Pass
Hotelling test (nearest neighbour)	Fail
Hotelling test (kernel)	Pass

Based on using the stratification method, Dehejia and Wahba (1999) estimated the impact of the NSW to be $1608. When using nearest neighbour matching with replacement, their estimate was $1691. Although not done in their paper, when using kernel matching with a Gaussian kernel, their impact estimate would have been $1519. All three estimates are close to the experimental benchmark of $1794 (see their Table 3). This explains the recent great interest in propensity score methods that was spurred by their paper. An interesting question is if they had checked their estimate based on nearest neighbour matching using any of the after matching tests, as done in Tables 2-4. Would they have rejected their estimate in that case? Was obtaining a close estimate based on matching in this case a fluke? The contradictory balancing test results for the same data set is the motivation behind the Monte Carlo simulations performed in the next section.

5. Monte Carlo Simulations Based on Generated Data

Is the positive correlation between the DW test and the three after matching balancing tests when kernel matching is employed in the NSW-PSID data a robust relationship? Is there no relationship to be expected between the DW test and the three after matching balancing tests with nearest neighbour matching? We investigate these questions in more detail in this section using Monte Carlo simulations based on artificially generated data. Such simulations serve as a useful way of controlling for factors which we wish to hold constant and examining how variation in the factors we wish to study affect the balancing test results. The simulations assume the CIA holds (i.e., we know which *X*s to use to estimate the true propensity score) and focuses on varying the sample size (*n* = 500, 1000, 2000), the number of covariates (2, 6, 12), the correlation between covariates, and the test level if a test statistic is used (employing the Bonferroni correction where necessary).[17] The goal of these simulations is to provide guidance for researchers on how closely related the results of different balancing tests are when used on the same data set under different scenarios.

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 on X_k), and $I(.)$ is the indicator function. The design of the Monte Carlo experiments in this section investigates the performance of the four balancing tests highlighted in the previous section when used together with three common ways of using the propensity score – propensity score stratification, nearest neighbour matching, and kernel matching. In particular, we simulate the use of the DW test when stratification is done, and the use of the test for standardised differences, the *t*-test, and the Hotelling test when performing nearest neighbour matching and kernel matching. Other balancing tests (for example, a regression test (Smith and Todd 2005a) or the Kolmogorov-Smirnov test (Diamond and Sekhon 2005)) and 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 these many other possible combinations to future work.

[17] Varying the number of covariates and distribution of covariates in the simulations represents an important advance over previous Monte Carlo work in the matching literature and is an attempt to make the simulations encompass more realistic scenarios. Drake (1993) and Zhao (2004), for example, use two N(0, 1) covariates in their simulations, while Frölich (2004) uses one covariate drawn from the Johnson S_B distribution.

We examine a total of nine different covariate distributions to help us get a better idea of how balancing tests might perform under a variety of settings. The first three scenarios use variables generated from the uniform distribution and vary their range from U(-6, 6) to U(-1, 1). The next three use normally distributed covariates and vary the variance of the covariates from N(0, 6) to N(0, 1). Finally, the last three use standard normal covariates, but vary the correlation between the covariates (ρ = 0.3, 0.5, 0.7). Given the distinction made between before matching and after matching balancing tests, and an interest in determining how similar the results of these tests are when conducted on the same data set, we perform all four tests on the same nine simulated covariate distributions. The data sets are generated such that they are balanced under the null.[18] For each data set, we begin by employing the stratification method in conjunction with the DW test. This replicates the work done in Dehejia and Wahba (1999). In this case, the DW test is both a before matching and after matching test because it uses the same sample (i.e., the common support region of the full sample) to first do the DW test and then use those same blocks from the DW test to estimate the treatment effect. The simulations of the DW test allow us to estimate the size of the DW test.

Next, using the same propensity score specification used in the DW test simulations, we perform nearest neighbour matching (with replacement) and kernel matching (using the Gaussian kernel). This attempts to replicate the procedure in Dehejia and Wahba (2002) where the DW test is first used as a test for *p(X)* prior to using the other matching algorithms to estimate the treatment effect. In these simulations, the focus is on the matched data sets, and the after matching balancing tests employed are the test for standardised differences, the *t*-test, and the Hotelling test. The DW test is not simulated in the matched data set because in practice, it is only used in the original full data set. Given that the original data set is balanced under the null, nearest neighbour matching or kernel matching essentially extracts a portion of data from a balanced data set. This should give rise to a matched data set that is balanced, but does not guarantee that it be so. For example, the change in the data set as a result of matching could cause the densities of *X* at certain values of *p(X)* to become too sparse so that it is no longer true that the distribution of *X* is approximately the same between groups within a neighbourhood of values of *p(X).* The simulations of the test for standardised differences, the *t*-test, and the Hotelling test allow us to estimate their respective test sizes, as well as their relationship with the DW test result. A key question of interest is whether we should expect any correspondence between a test done on a full sample (a before matching test) and a test done on a matched sample (an after matching test).

[18] Generating a balanced data set under the null in order 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 *X*s, when we use the error term, arbitrary values of β and *X* to generate *D*, it is true that:

$$D \perp X \mid X\beta$$

As only monotonic transformations are performed, it therefore follows that

$$D \perp X \mid \text{logit}(X\beta) \quad \text{or} \quad D \perp X \mid p(X)$$

Therefore by construction, these data sets satisfy the balancing property of propensity scores: $D \perp X \mid p(X)$.

We choose the parameters so that the distribution of *p(X*) and the treatment-comparison group ratio in the simulations reflect real world scenarios. In all cases, the coefficients are $\beta_k = 1$ in the selection equation. Given these coefficients, the intercept is chosen so that approximately 20 percent of the observations are in the treatment group and 80 percent of the observations are in the comparison group. The selection equation is specified such that for treatment assignment, observations with large values of $\sum_{k=1}^{K} X_k$ are likely to be assigned to treatment, while those with small values are likely to be assigned to control. This creates a data set in which there are relatively few controls with large propensity score values and relatively few treated units with small propensity score values, but a sizeable overlap of common support, a pattern often observed in practice. To illustrate, Figures 1 to 9 depict the distribution of *p(X)* in the treatment and comparison groups for the case of six covariates and *n* = 2,000. [19]

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) have suggested that the estimated score helps to remove any potential sample imbalances and can lead to better balance.

Finally, we also examine the relationship between the DW test result and the unbiasedness of the treatment effect estimate to determine if any relationship between the balance test result and the bias of the average treatment effect exists.

5.1 Results for Balance based on Generated Data

The results of simulating the DW test is shown in Table 5. For the simulations with 2 covariates (top panel), we see that using a conventional test level of α = 0.05 (first three columns), the DW test performs terribly in terms of size and rejects the null much more often that it should. But with the Bonferroni correction made for the test level (top panel, last three columns), the DW test simulations come much closer to replicating their true sizes. (The chosen test size is divided by 10 because assuming there are five blocks of the propensity score and two covariates to compare within each block, there are a total of 5x2 = 10 comparisons to be made). The simulations for the case of 6 covariates (middle panel) and 12 covariates (bottom panel) tell the same story. For the sample sizes considered (500, 1000, 2000), it appears that the correction for multiple comparisons helps the DW test to achieve a more correct size.

Tables 6-8 focus on the situation when nearest neighbour matching with replacement is employed. Each row in the table corresponds exactly to the rows in Table 5 in terms of the setup, the only difference being that a matched sample is used in Tables 6-8 while the full sample over the common support is used in Table 5. For example, for the case of 12 U(-1, 1) covariates, it could be the case that in Table 5, when simulating the DW test, the sample used in a simulation could be 985 (the common support when *n* = 1000) whereas in Table 6, the corresponding sample used to simulate the *t*-test could be around *n* = 400.

[19] The intercept needs to be varied when we use a different number of covariates or sample size in order to maintain the 20-80 treatment-comparison group ratio in the two groups.

The idea is that if the data is originally balanced before matching (and it is because we generated it to be so), we are interested to see how tests for balance after matching perform.

For the case of 2 covariates (top panel of Table 6), there is a fair correlation between the results of the DW test on the original sample and the *t*-test on the matched sample. The exceptions are when the distribution of the covariates have larger variances: U(-6, 6), N(0, 6) and N(0, 3). However, in the case of 6 and 12 covariates (middle and bottom panel of Table 6), there is a very low correlation with the corresponding results of the DW test in Table 5, with cases that would attain balance by the DW test failing the after matching tests.

When simulating the Hotelling test (Table 7) and the test for standardised differences (Table 8) after performing nearest neighbour matching with replacement, the story is much the same, there being a fair correlation between the DW test in the case of 2 covariates, but no correlation in the case of 6 and 12 covariates. For the test for standardised differences, even a more lenient rule that considers a standardised difference > 40 as large (as opposed to 20) does not alter the conclusion. Increasing the sample size does not appear to affect this relationship between the before and after matching tests by very much.

Tables 9-11 replicate Tables 6-8, except that kernel matching is used in place of nearest neighbour matching. The results are very similar and there is in general a very low correlation with the DW test results in the scenarios we examine.

It is rather puzzling why the three after matching balancing tests have such high rejection rates since under the null, the before matching data set is balanced. We can think of two possible explanations. The first is that the rules we have devised for the *t*-test and test for standardised differences are too rigid in that we specified that as long as any one covariate is found to have imbalance (even if there is a large reduction in covariate differences in general), the balancing test fails. A second possible explanation is that even when you start with a balanced data set, in the sense that the covariates are independent of the treatment variable given the propensity score, matching itself can create an imbalance or make balance worse. But what is still puzzling is the high rates of rejection when the number of covariates is greater than 2, even when the covariates are N(0, 1) and do not have unusually large variances. For example, for the case of 12 N(0, 1) covariates and *n* = 2,000, the test of standardised differences rejected balance 96.2% of the time when 'large' was defined as a value of greater than 20 and 47.0% of the time when 'large' was defined as a value of greater than 40 (Table 8, bottom panel, sixth row).[20]

[20] Suspecting that a sample size of 2,000 was not enough, we experimented with a sample size of *n* = 10,000 and *n* = 50,000 for the case of 12 N(0, 1) covariates to see if that made a difference. This changed the results of the test for standardised differences but not the results of the *t*-tests and Hotelling test. For *n* = 10,000 and using the "large equals a difference greater than 20" definition, balance was rejected 43% of the time, while using the "large equals a difference greater than 40" definition, balance was rejected 0.8% of the time. The corresponding results for *n* = 50,000 were 1% and 0%. So perhaps there is a stronger correlation between the DW test and test for standardised differences as *n* increases, but not in realistic sample sizes used in practice.

5.2 Summary of Monte Carlo Simulations Based on Generated Data

Of the four balancing tests examined in the Monte Carlo simulations, it appears that the DW before balancing test combined with a Bonferroni adjustment is the only test that is reliable. This is because adjusting for multiple tests appears to give the DW test the correct size. In contrast, the three after matching balancing tests considered – the test of standardised differences, the *t*-test and the Hotelling test – all do not attain the correct test sizes and appear to be of doubtful value as a balancing diagnostic.

However, while simulation results based on artificially generated data can be useful, it is perhaps more interesting to see how the balancing test performs when doing simulations using real world data (i.e., distributions of covariates from actual data sets that do not conform to textbook statistical distributions). If the promising results of the DW test with a Bonferroni adjustment hold up when using real world data, then the implications for applied researchers would be to use it as a diagnostic and not any of the after balancing tests.

6. Monte Carlo Simulations based on Real World Data

Using the Dehejia and Wahba (1999) specification for the NSW-PSID data set that was previously analysed, we base our simulations on the following 'true' specification of the propensity score. That is, equation (3) defines the latent variable in the true data generating process under the null hypothesis of a balanced specification,

$$\begin{aligned}\Pr(D=1\mid X) = F(&-7.552 + 0.3305\text{age} - 0.0063\text{age}^2 + 0.8248\text{educ} - 0.04832\text{educ}^2 - 1.8841\text{ married} + \\ &0.1299\text{ nodegree} + 1.1329\text{black} + 1.9628\text{hisp} - 0.000105\text{RE74} - .000217\text{RE74}^2 + \\ &2.36{*}10^{-9}\text{RE75} + 1.58{*}10^{-10}\text{RE75}^2 + 2.14\text{U74}{*}\text{black}\,)\end{aligned} \tag{3}$$

where F is the cumulative logistic distribution. In the same way that we generated a balanced data set using the artificial data, a balanced data set based on the NSW-PSID data can be generated by assuming that the error term in each simulation is random and independent of the covariates.

We first examine the results of the most promising test as suggested from the simulations on artificial data. Unfortunately, when the DW test with a Bonferroni adjustment is simulated using the real world data, the test size is no longer close to the desired size of 5% and increases to 23.75% (Table 12, first row). As a robustness check, the same exercise was repeated using the DW 2002 and DW 2005 specifications to define the data generating process underlying the simulations. The same results (Table 12, rows 2 and 3) emerged. One possible reason that the modified DW test fails to achieve the correct size is the presence of difficult to balance variables. To test this possibility, the simulation exercise was repeated using the DW 1999 specification but excluding the variables *Nodeg*, *Black* and *Hisp*, the variables that were found to be difficult to balance using the after matching balancing tests in section 4. Using such a specification, a close to correct size of 5.9% was obtained.

In general, however, an approach that involves dropping variables in order to obtain balance cannot be recommended as this would lead to departures from the CIA, assuming the CIA holds on the full set of *X*s. What is clear is that the DW test that is based on using multiple *t*-tests on mean covariate differences between groups will tend to reject balance too often when there are 'problematic' covariates.

The *t*-test performs well when conditions underlying it are met, especially the assumption that the distribution of data in the underlying population from which each of the samples is derived is normal. However, when one cannot make the assumption of normality, Westfall and Young (1993, section 2.5.1) have illustrated using simulation how non-normal variables can adversely affect the sampling distribution of the *t*-statistic. In considering the minimum of ten independent *t*-statistics when sampling from a lognormal population, they show how the effect of skewness is greatly amplified. They find in their simulations that values in the extreme low tails are much more likely than under the corresponding *t*-distribution (see their Figure 2.3) and that the possibility of observing a statistically significant result can be much larger than under the assumption of normal data. They make the point that this can cause false significances to occur even if correction for multiple tests like the Bonferroni adjustment are used. Although they expect the *t*-test to be more robust when testing contrasts between two or more groups, especially if the skewness of the variables in the groups are roughly comparable, we see in our Monte Carlo simulations using the NSW-PSID data that even with the Bonferroni correction, we were obtaining more false significant results than we would expect.

Before discussing a possible further refinement of the DW test in the next section, we briefly present the results of the three after matching balancing tests using the NSW-PSID data (Tables 13 to 15). When performing the test for standardised differences (Table 13), it can be seen that for most of the variables (with the exception of *Age* and *Educ*), the average standardised difference decreases after nearest neighbour or kernel matching has been done. However, the rule of rejecting balance as long as any variable has a standardised difference > 20 leads to extremely high rejection rates (close to 100%) on the balanced data set, and resembles the findings based on artificial data (Tables 8 and 11).

The after matching version of the *t*-test seems to perform well, as can be seen from the relatively high average *p*-values after nearest neighbour matching or kernel matching (Table 14). However, 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 that once again mirror the simulation results based on artificial data (last two rows of Table 14).

Finally, the results of the Hotelling test are shown in Table 15. Like the other two after matching tests, the null of balance is rejected too often.

7. Conducting Non-Parametric Balancing Tests

A systematic investigation of four commonly employed balancing tests has to this point not yielded any encouraging results. There is little Monte Carlo evidence to support these four balancing tests as they are currently employed in the literature. In this section, we offer suggestions on how some of these tests might be improved in terms of obtaining better test sizes. The suggested remedy essentially involves conducting non-parametric versions of the balancing tests described. In particular, for the DW test, a permutation distribution of the *t*-statistic is used in place of the *t*-distribution. Similarly, for the other two 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.

In the next few sections, we first describe the theory 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.

7.1 The Theory of Permutation Tests

Permutation tests are a computer intensive statistical technique that was introduced by R.A. Fisher in the 1930s. 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).

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 **y** and **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 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. For example, the *t*-test is used in the DW test as well as an after matching test. 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, labelling 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.[21]

Good (2001) outlines a simple five-step procedure for conducting permutation tests.

[21] Abadie (2002) notes in his ‘summary and discussion of possible extensions’ section (p.290) 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.

(1) Analyse the problem.

(2) Choose a test statistic.

(3) Compute the test statistic for the original labelling of the observations.

(4) Rearrange (permute) the labels and recompute the test statistic for the rearranged labels. Repeat a large number of times to obtain the distribution of the test statistic. (In small samples, it is possible to perform all possible permutations to obtain an exact test).

(5) Accept or reject the hypothesis using this permutation distribution as a guide.

In our context, to form a two-tailed test for whether a covariate mean in the treatment group is larger than it is for the control group at the 5% level, we would use the 2.5% and 97.5% values in the permuted distribution of the test statistic as our cut-off values. We focus on the case where matching is done with replacement. Hansen (2006) discusses the use of permutation tests when matching is done without replacement.

Assuming the assumption of exchangeability of the observations is plausible, an appealing aspect of permutations tests is that they are exact in finite samples and unbiased. According to Good (2001), the key question to ask if one's observations are exchangeable and a permutation test applicable is whether under the null hypothesis of no differences among the various groups, we can exchange the labels on the observations without significantly affecting the results. It is quite clear that in the context of a balancing test, this condition is fulfilled.

Below, we discuss in more detail how we implement the permutation versions of the DW test, as well as the permuted versions of the after matching *t*-test and Hotelling test.

7.2 The Permutation Version of the DW Test

The procedure for implementing the DW test is described by Dehejia and Wahba (2002) in their appendix, and provided in more detail in Becker and Ichino (2002). We repeat it here so that the differences between it and the permutation version of the DW test are clear.

Algorithm for 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 *t*-test to test that the mean *p(X)* values for treated and comparison units do not differ. If the test fails, spilt the interval in half and test again. The 'optimal' number of intervals is found when the mean *p(X)* values for treated and comparison units do not differ in all intervals.
4. Within each interval, use a *t*-test to test that for all covariates, the mean differences 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.

Sections 5 and 6 have shown that this algorithm does not work well in Monte Carlo simulations. Note that with regard to the point mentioned earlier about the *t*-test performing poorly when variables are non-normal, in the case of the DW algorithm, as the *t*-test is used in steps 3 and 4, there are two possible junctures where problems might arise. We therefore modify the algorithm in two respects: (i) using a permuted version of the *t*-test in both selecting for the 'optimal' number of intervals to use and in testing for the equality of covariates within each interval; (ii) correcting for multiple testing using the Bonferroni adjustment. The modified algorithm is as follows.

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 1000 random permutations) to test that the mean *p(X)* values for treated and comparison units do not differ. The test level to use is α (e.g., 0.05, 0.01). If the test fails, spilt the interval in half and test again. The 'optimal' number of intervals is found when the mean *p(X)* values for treated and comparison units do not differ in all intervals.
4. Within each interval, use a permuted *t*-test (with at least 1000 random permutations) to test that for all covariates, the mean differences treated and comparison units are not significantly different from zero. To adjust for multiple covariates, the Bonferroni adjusted test level to use is $\alpha/(kv)$, where *k* is the number of intervals and *v* is the number of covariates (e.g., if α is chosen to be 0.05, with 5 intervals and 8 variables, the Bonferroni adjusted test level is 0.05/(5x8) = 0.00125).
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.

This modified version of the DW test achieves a good size (1.2%), based on simulations using the DW 1999 specification, where 500 Monte Carlo simulations based on 1000 random permutations of the *t*-test were performed (Table 16).[22] This result should be directly compared with the size results in Table 12, as both Tables 12 and 16 use the same data set and specification for the true propensity score. The only difference is that the original DW algorithm is used in Table 12 while the permutation version of the algorithm is used in Table 16.

7.3 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. Implicitly, conditioning on *p(X)* occurs through the weighting of the comparison group observations, where 'distant' observations are given little or no weight. As a result, the *t*-test used in this context implicitly checks for the balancing property of propensity scores given in (1). As the standard *t*-test did not fare well in the Monte Carlo simulations in sections 5 and 6, we modify the test by using the permutation version of the *t*-test. In addition, to account for multiple testing, we use the Bonferroni

[22] This simulation is computationally expensive as in each Monte Carlo simulation, the number of random permutations can be quite large. With 13 variables, ≥ 5 intervals in each simulation and 1000 random permutations, a Pentium 4, 3.2 Ghz computer can do approximately 100 Monte Carlo simulations in 24 hours.

adjustment to divide the chosen test level by the number of variables (e.g., if α is chosen to be 0.05, with 8 variables, the Bonferroni adjusted test level is 0.05/(8) = 0.00625).

The after matching Hotelling test is the multivariate analogue 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 sections 5 and 6. However, unlike the *t*-test, because it tests for a joint null hypothesis, no correction for multiple testing needs to be done.[23]

Using 1000 random permutations of the *t*-test and Hotelling test, we see in Table 17 that much better test sizes are achieved. Using a Bonferroni adjusted test size of 0.05/13 = 0.0038, the *t*-test rejects the null of balance only 1.8% of the time under nearest neighbour matching, and only 2.2% of the time under kernel matching. The fact that it is less than the nominal size of 5% could be due to the conservative nature of the Bonferroni adjustment. Furthermore, in our 500 Monte Carlo simulations, the Hotelling test hardly or never rejects the null at the 1% and 5% level, implying a very low type 1 error rate.

7.4 Power of the Tests and Bias on Outcomes

Having shown via simulation 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 a power simulation by introducing a small variation to the setup. This allows us to determine the utility of balancing tests *when we do not estimate the propensity score correctly*. In other words, the true data generating process (DGP) for the propensity score differs from the specification we use to estimate the propensity score. As the true DGP is unknown in observational studies, these power results are potentially more important in practice than the size results, as tests with good sizes but low power can be of limited use.

As a researcher attempting to do propensity score matching, one of the main choices that needs to be made is the set of *X* variables to include when estimating the propensity score. In other words, *X* needs to be chosen so that the CIA is fulfilled. These variables are chosen as they affect both the selection into treatment and the outcome variable. As there is no separate equation specifying the outcome unlike other selection models in econometrics, when one says that variable ***z*** is incorrectly omitted from the set *X*, it could be because it is not included in the true DGP for the propensity score, the outcome, or both. Hence, in the context of propensity score matching, when one estimates *p(X)* without ***z***, the nature of the omitted variable problem is not immediately clear.

One of the key reasons put forth by Dehejia and Wahba (1999, 2002) as to why they were able to successfully replicate the NSW experimental findings was that they used the variable that measured earnings history going further back (*RE74*). In order to accommodate this, they restricted the sample from Lalonde (1986) to those observations with non-missing values for *RE74*. As the inclusion of this variable

[23] See Blair et al. (1994) for more discussion on the permutation version of Hotelling's test.

has been subject to much discussion (see Smith and Todd 2005a), we base our power simulations on the NSW-PSID data and focus on scenarios where we include *RE74* in the true DGP for the propensity score and outcome, but neglect to include it when we estimate the propensity score. By repeatedly estimating the incorrect propensity score (i.e., not including *RE74* when we should be), we can assess how often the balancing tests find imbalance when there should be imbalance, and also how findings of balance or imbalance are related to bias in the estimated treatment effect.

In order to determine the effect of omitting *RE74* when we shouldn't be, we consider three cases of the true DGP: (1) *p(X)* contains *RE74* and *Y* contains RE74; (2) *p(X)* contains *RE74* and *Y* does not contain *RE74*; and (3) *p(X)* does not contain *RE74* and *Y* contains *RE74*.

In the first DGP, the true propensity score and the outcome variable contain the variable *RE74*. Based on the NSW-PSID data, we let the true DGP for the propensity score be

$$\begin{aligned}\Pr(D=1 \mid X) = F(&-3.929 + 0.2833\text{age} - 0.0054\text{age}^2 - 0.0237\text{educ} - 1.788\text{ married} + \\ &0.6739\text{nodegree} + 2.1331\text{black} + 2.2192\text{hisp} - 0.000116\text{RE74} - 0.000263\text{RE75})\end{aligned} \tag{4}$$

and the true DGP for the outcome variable be

$$\begin{aligned}Y = &-2872.058 + 1000\text{D} + 106.30\text{age} - 2.62\text{age}^2 + 586.52\text{educ} + 629.18{*}\text{nodegree} + 975.46\text{married} \\ &- 518.28\text{black} + 2248.67\text{hisp} + 0.2757\text{RE74} + 0.5659\text{RE75} + \varepsilon\end{aligned} \tag{5}$$

where ε is N(0, 1000). 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 \$1000. That is, when *p(X)* is estimated using the specification in (4) and the outcome is generated according to (5), the estimated treatment effect should have a distribution that is centred around \$1000. This is indeed the case when we perform simulations (results not shown and available on request).

In the second DGP, we use the same true DGP for the propensity score in (4) but exclude *RE74* from the true DGP for the outcome variable.

$$\begin{aligned}Y = &-4684.848 + 1000\text{D} + 198.08\text{age} - 3.53\text{age}^2 + 646.86\text{educ} + 527.84{*}\text{nodegree} + 1320.29\text{married} \\ &- 608.57\text{black} + 2608.62\text{hisp} + 0.7850\text{RE75} + \eta\end{aligned} \tag{6}$$

where η is N(0, 1000). Here, compared to the first DGP, the omission of *RE74* is considered less serious as it plays no role in the process that determines the outcome.

Finally, in the third DGP, we omit *RE74* from the true DGP for the propensity score

$$\begin{aligned}\Pr(D=1 \mid X) = F(&-3.669 + 0.2476\text{age} - 0.0049\text{age}^2 - 0.0295\text{educ} - 1.938\text{ married} + \\ &0.8449\text{nodegree} + 2.1182\text{black} + 2.0111\text{hisp} - 0.000350\text{RE75})\end{aligned} \tag{7}$$

while using the true DGP for the outcome variable as in (5), where *RE74* is included. In this DGP, the omission of *RE74* is again less serious than it is for the first DGP, as it does not determine selection into treatment.

For all DGPs, the estimated propensity score in each simulation is estimated as

$$\Pr(D = 1 \mid X) = F(\text{age, age}^2\text{,educ, married, nodegree, black, hisp, RE75}) \qquad (8)$$

and the estimated average treatment effect is the mean difference $(Y \mid D = 1) - (Y \mid D = 0)$.

In the three DGPs, interest centres around the two questions of how often balance is rejected and how different the estimated treatment effect is from the true treatment effect of $1000 when we incorrectly omit *RE74* and use (8) to estimate the propensity score.

7.5 Power Simulation Results for the Stratification Method

The results of the power simulations for the three DGPs when matching is done based on the stratification method is shown in Table 18. In the first DGP, we see that the permuted version of the DW test has a rejection rate of the null of balance of 3.6%. This suggests that even though the test has good size properties, it has low power to detect an omission of a relevant variable that should be included in the propensity score. Next, when we compare the average treatment effects when we find balance and imbalance, we see that there are very little differences in the estimated average treatment effect ($349.50 vs. $351.97). Perhaps what is more noticeable is the fact that both these estimates are more than two standard deviations away from the true average treatment effect of $1000. In short, for this DGP, as *RE74* is omitted so that the CIA is not fulfilled, balancing tests are of little utility as they do not help distinguish between good and bad estimates of the average treatment effect. If no balancing tests were employed, the estimated average treatment effect is $349.59, $0.09 more than the effect when balance is found.

In the second DGP, recall that the omission of *RE74* is considered less serious as it plays no role in the process that determines the outcome. This makes a key difference in the estimates of the average treatment effect. When balance is found, the estimated effect is now $886.58, much closer to the true value of $1000. However, balancing tests again appear to be of little utility as the estimated effects when imbalance is found ($897.24) when no balancing test is employed ($886.96) are essentially the same as the effect when balance is found.

The third DGP has results similar to those of the second DGP. While the estimated average treatment effects are again closer to the true effect because of a less severe case of omitted variables, balancing tests do not seem to be of much use. When no balancing test is employed, the estimated effect is $877.82, as compared to $878.54 when balance is found.

7.6 Power Simulation Results for Nearest Neighbour Matching

The results for after nearest neighbour matching balancing tests are slightly more encouraging (Table 19). Although the power of the tests to detect misspecification are still low (e.g., balance is rejected less than 15% of the time), the average treatment effects when balanced are now generally closer to \$1000 than they are when imbalanced. In the first DGP, for example, using the Bonferroni corrected test size of 0.00625, the estimated average treatment effect is \$558.19 as compared to \$428.91 when imbalance is found. However, because the percent of times balance is rejected is small (about 1%), not using a balancing test also gives rise to an average treatment effect of \$557.16 that is close to the effect when balanced.

The second and third DGPs provide similar results in terms of finding average treatment effects when balanced to be closer to \$1000 than when imbalanced, and in terms of finding that one could have done equally well not employing a balancing test at all.

As compared to the stratification method, it appears that nearest neighbour matching under these three DGPs provide average treatment effects that are in general closer to the true treatment effect. As the permutation version of the Hotelling test (Table 20) hardly rejects the null of balance (it has very low power), it is not meaningful to compare the bias of average treatment effects when balance or imbalance is found.

7.7 Power Simulation Results for Kernel Matching

When permutation versions of after kernel matching balancing tests are done (Table 21), the results largely mirror those for nearest neighbour matching. The chief exception is that kernel matching appears to perform less well for these three DGPs, as the estimated average treatment effects are more biased than they were for nearest neighbour matching. For the permutation version of the Hotelling test (Table 22), the null of balance is never rejected in all three DGPs, implying that the test has no power to detect the omission of *RE74*.

8. Conclusions and Recommendations for Practice

This paper was motivated by Smith and Todd's (2005b) observation that multiple versions of the balancing test exist, with little known about their properties or how they compare to one another. An example based on the NSW-PSID data was provided, illustrating how different balancing tests gave rise to different results regarding balance. We argue that this difference can mainly be explained by the difference between the use of a before matching balancing test and an after matching balancing test, as these tests for balance in different samples.

The failure of the three after matching balancing tests on simulations based on artificial data and all the balancing tests on simulations based on the real world NSW-PSID data suggests a need for an overhaul of balancing tests. Balancing tests currently employed in the literature are shown by Monte Carlo simulations in this paper to provide overly large test sizes. 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.

When the CIA is fulfilled, simulations in sections 5 and 6 show the unreliability of parametric based balancing tests as balance is often rejected more often than it should be. Section 7 shows how permutation versions of the DW test, after matching *t*-test and after matching Hotelling test can achieve better test sizes.

However, when the CIA is not fulfilled, it appears that even non-parametric balancing tests that have low type 1 error rates are of limited use. This underscores the importance of attempting to fulfil the CIA when using propensity score matching estimators.

Figures 1-9: The Nine Different Covariate Distributions used in the Monte Carlo Simulations and their Distribution of Propensity Scores

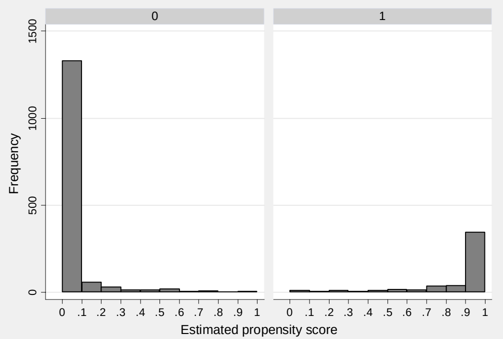

Figure 1: 6 independent U(-6, 6) covariates and n = 2000

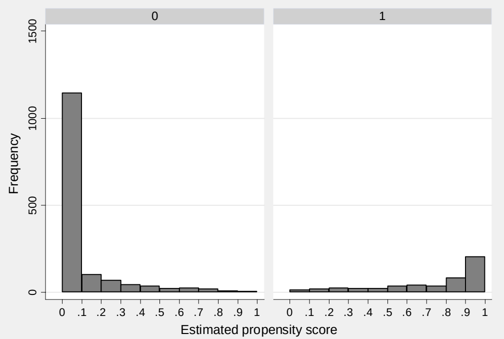

Figure 2: 6 independent U(-3, 3) covariates and n = 2000

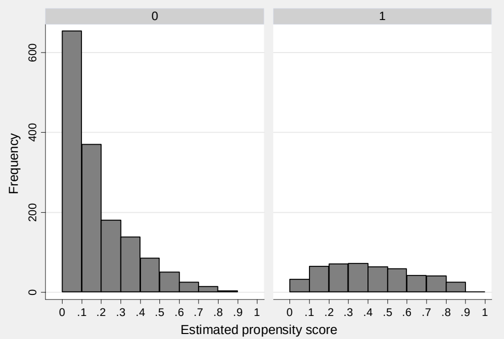

Figure 3: 6 independent U(-1, 1) covariates and n = 2000

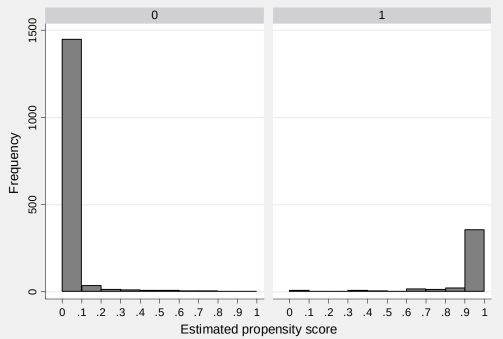

Figure 4: 6 independent N(0, 6) covariates and n = 2000

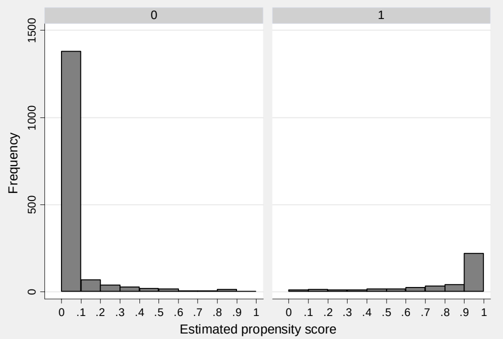

Figure 5: 6 independent N(0, 3) covariates and n = 2000

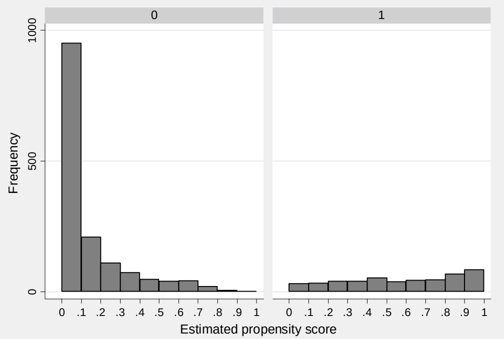

Figure 6: 6 independent N(0, 1) covariates and n = 2000

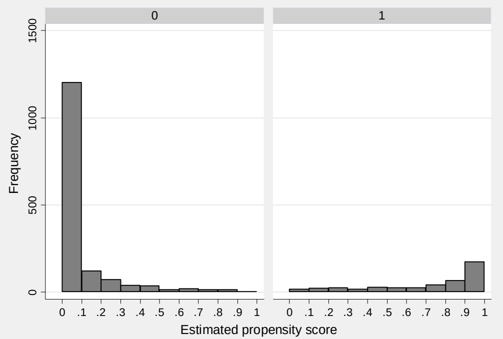

Figure 7: 6 N(0, 1) covariates with correlation = 0.3 and n = 2000

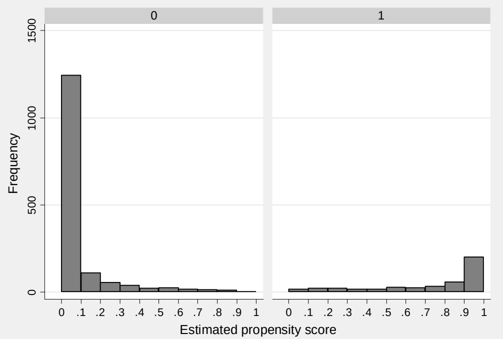

Figure 8: 6 N(0, 1) covariates with correlation = 0.5 and n = 2000

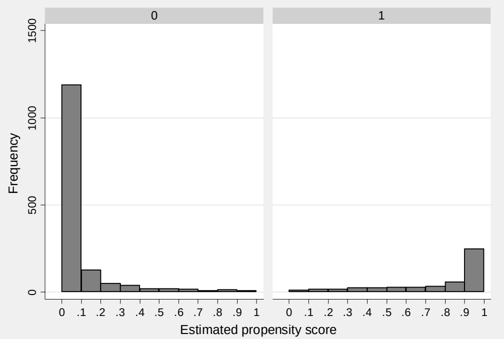

Figure 9: 6 N(0, 1) covariates with correlation = 0.7 and n = 2000

Notes: The cases illustrated are when there are 6 covariates and *n* = 2000, which correspond to the middle panel and third, sixth and ninth columns of Tables 5-11.

Table 5: Stratification Method and using the DW Test as a Check for Balance Before Matching

Two Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the DW test (α = 0.05)			Simulated Percent Rejection of Balance Based on the DW test (α = 0.01)			Simulated Percent Rejection of Balance Based on the DW test (α = 0.05/10 = 0.005)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	23.4	32.4	43.4	6.2	9.2	13.8	3.6	6.2	8.8
Independent U(-3, 3)	25.2	33.4	46.2	6.4	8.2	13.0	4.2	4.8	10.0
Independent U(-1, 1)	19.2	27.8	35.0	3.4	9.0	11.6	1.8	4.4	5.8
Independent N(0, 6)	17.4	22.4	33.6	3.4	4.0	6.6	2.0	2.6	3.0
Independent N(0, 3)	22.4	30.8	35.8	3.8	7.4	9.0	1.4	4.4	5.8
Independent N(0, 1)	24.0	30.2	40.4	6.0	6.8	10.2	3.0	3.6	5.8
N(0,1) correlation = 0.3	26.4	32.6	40.8	5.0	8.0	10.0	3.4	3.8	5.2
N(0,1) correlation = 0.5	29.0	35.0	42.4	5.8	7.2	13.4	3.6	4.2	5.0
N(0,1) correlation = 0.7	29.4	36.2	48.8	9.8	10.4	18.0	2.8	4.2	9.2

Six Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the DW test (α = 0.05)			Simulated Percent Rejection of Balance Based on the DW test (α = 0.01)			Simulated Percent Rejection of Balance Based on the DW test (α = 0.05/30 = 0.0016)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	60.6	71.2	85.6	17.2	22.4	25.6	3.6	4.6	6.2
Independent U(-3, 3)	66.6	79.8	89.2	15.8	23.2	30.2	2.2	3.8	5.8
Independent U(-1, 1)	64.8	77.4	87.4	11.8	25.2	27.8	1.4	4.8	5.8
Independent N(0, 6)	53.2	66.2	74.0	10.4	14.4	18.4	0.8	2.0	3.2
Independent N(0, 3)	56.8	74.4	81.2	12.4	16.6	21.0	2.2	3.2	1.8
Independent N(0, 1)	68.0	78.6	87.8	13.0	20.4	22.4	1.2	3.0	3.8
N(0,1) correlation = 0.3	67.4	78.6	88.2	12.8	23.0	28.8	1.2	3.4	8.4
N(0,1) correlation = 0.5	64.4	82.8	90.4	16.8	24.4	29.4	1.8	4.8	4.4
N(0,1) correlation = 0.7	66.4	86.8	89.0	14.0	26.2	34.2	1.8	8.0	7.6

Twelve Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the DW test (α = 0.05)			Simulated Percent Rejection of Balance Based on the DW test (α = 0.01)			Simulated Percent Rejection of Balance Based on the DW test (α = 0.05/60 = 0.0008)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	79.0	86.0	94.4	24.8	26.4	34.8	2.4	1.4	2.6
Independent U(-3, 3)	82.8	93.0	96.8	25.0	34.0	43.6	2.4	2.0	2.6
Independent U(-1, 1)	88.0	96.6	98.6	28.0	37.6	44.0	3.0	2.4	4.0
Independent N(0, 6)	-	79.2	91.2	-	22.2	32.0	-	2.8	2.0
Independent N(0, 3)	77.0	90.8	95.2	24.2	32.8	34.4	2.2	2.0	1.8
Independent N(0, 1)	89.0	94.0	98.8	27.0	35.4	46.4	2.6	1.8	3.4
N(0,1) correlation = 0.3	85.2	90.2	97.4	28.6	33.8	40.8	2.4	2.4	3.2
N(0,1) correlation = 0.5	81.6	91.0	96.2	29.2	34.2	43.8	2.4	2.6	4.4
N(0,1) correlation = 0.7	77.2	91.6	94.6	21.8	34.2	43.0	1.4	3.2	2.0

Notes: Based on 500 replications. The smaller level of α in the third set of columns is meant to be an approximate Bonferroni adjustment for multiple comparisons within each block at the α = 0.05 level, assuming the average number of blocks in each simulation is five. The actual number of blocks used in any simulation actually varies depending on the particular distribution of *p(X)* in each simulation as each block that is unbalanced needs to be divided into smaller blocks until no imbalance in covariates is found. Hence, even though the overall average number of blocks used in all the simulations is approximately five, the Bonferroni correction used here is only an approximation with the point being to illustrate the importance of correcting for multiple comparisons. For the case of 12 N(0, 6) covariates and *n* = 500, there was a lack of common support for many of the replications.

Table 6: Nearest Neighbour Matching and using *t*-Tests as a Check for Balance After Matching

Two Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the t-test ($\alpha = 0.05$)			Simulated Percent Rejection of Balance Based on the t-test ($\alpha = 0.01$)			Simulated Percent Rejection of Balance Based on the t-test ($\alpha = 0.05/2 = 0.025$)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	67.4	69.2	72.0	50.8	57.0	60.6	60.0	65.4	66.2
Independent U(-3, 3)	9.4	5.0	5.0	2.0	1.2	1.0	5.2	2.6	1.8
Independent U(-1, 1)	0	0	0.2	0	0	0	0	0	0
Independent N(0, 6)	98.2	99.0	99.8	94.0	96.4	99.4	97.4	98.4	99.6
Independent N(0, 3)	70.4	77.8	78.0	57.0	64.4	64.8	64.2	72.0	72.4
Independent N(0, 1)	3.6	3.2	1.8	1.0	0.4	0.2	2.0	1.0	0.8
N(0,1) correlation = 0.3	3.8	1.2	2.8	0.4	0.2	0.2	1.2	0.8	1.2
N(0,1) correlation = 0.5	3.2	1.8	3.0	0.8	0.6	0.2	1.8	1.0	1.2
N(0,1) correlation = 0.7	1.6	0.4	2.8	0.4	0	1.0	0.6	0	1.4

Six Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the t-test ($\alpha = 0.05$)			Simulated Percent Rejection of Balance Based on the t-test ($\alpha = 0.01$)			Simulated Percent Rejection of Balance Based on the t-test ($\alpha = 0.05/6 = 0.0083$)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	100	100	100	99.8	100	100	99.8	100	100
Independent U(-3, 3)	95.0	99.0	99.6	87.8	94.6	96.4	87.4	94.0	96.0
Independent U(-1, 1)	16.6	16.8	20.6	3.6	4.0	3.6	2.6	3.2	2.4
Independent N(0, 6)	100	100	100	100	100	100	100	100	100
Independent N(0, 3)	99.6	100	100	98.8	99.6	100	98.2	99.6	99.8
Independent N(0, 1)	56.4	70.6	80.6	30.0	44.2	57.6	29.0	42.6	56.4
N(0,1) correlation = 0.3	91.2	97.0	97.8	77.8	91.6	94.8	76.2	91.0	94.4
N(0,1) correlation = 0.5	97.2	99.4	99.4	91.8	97.2	98.2	91.0	96.6	97.8
N(0,1) correlation = 0.7	99.6	100	99.8	99.0	99.4	99.8	98.8	99.2	99.0

Twelve Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the t-test ($\alpha = 0.05$)			Simulated Percent Rejection of Balance Based on the t-test ($\alpha = 0.01$)			Simulated Percent Rejection of Balance Based on the t-test ($\alpha = 0.05/12 = 0.00416$)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	100	100	100	100	100	100	100	100	100
Independent U(-3, 3)	100	100	100	99.4	100	100	99.0	100	100
Independent U(-1, 1)	57.0	78.8	77.8	28.4	45.2	44.4	15.6	31.0	30.4
Independent N(0, 6)	-	100	100	-	100	100	-	100	100
Independent N(0, 3)	100	100	100	100	100	100	100	100	100
Independent N(0, 1)	94.6	99.8	99.6	76.2	97.4	98.8	65.2	95.0	96.6
N(0,1) correlation = 0.3	100	100	100	100	100	100	99.8	100	100
N(0,1) correlation = 0.5	100	100	100	100	100	100	100	100	100
N(0,1) correlation = 0.7	100	100	100	100	100	100	100	100	100

Notes: Based on 500 replications. The smaller level of α in the third set of columns is meant to be an approximate Bonferroni adjustment for multiple comparisons at the $\alpha = 0.05$ level. For the case of 12 N(0, 6) covariates and *n* = 500, there was a lack of common support for many of the replications.

Table 7: Nearest Neighbour Matching and using Hotelling Tests as a Check for Balance After Matching

Two Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the Hotelling test (α = 0.05)			Simulated Percent Rejection of Balance Based on the Hotelling test (α = 0.01)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	69.4	64.4	69.0	53.4	54.8	60.4
Independent U(-3, 3)	5.6	3.4	3.2	1.6	0.6	0.6
Independent U(-1, 1)	0	0	0.4	0	0	0
Independent N(0, 6)	99.6	99.4	99.8	98.6	99.2	99.8
Independent N(0, 3)	68.0	75.8	82.8	54.4	62.8	67.6
Independent N(0, 1)	2.4	1.4	1.6	0.6	0.4	0.2
N(0,1) correlation = 0.3	3.0	1.4	3.0	0.4	0.2	0.6
N(0,1) correlation = 0.5	5.2	3.6	6.0	1.0	1.2	0.6
N(0,1) correlation = 0.7	6.4	5.2	6.4	2.6	1.0	2.6

Six Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the Hotelling test (α = 0.05)			Simulated Percent Rejection of Balance Based on the Hotelling test (α = 0.01)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	100	100	100	99.8	100	100
Independent U(-3, 3)	89.4	97.4	98.6	81.6	93.6	96.6
Independent U(-1, 1)	2.6	2.8	2.8	0.4	0.6	1.2
Independent N(0, 6)	100	100	100	100	100	100
Independent N(0, 3)	99.4	99.8	100	98.6	99.8	99.8
Independent N(0, 1)	29.4	45.8	57.0	16.4	31.0	42.4
N(0,1) correlation = 0.3	82.6	94.2	94.8	71.8	89.4	92.8
N(0,1) correlation = 0.5	93.6	98.2	99.2	89.4	96.8	97.6
N(0,1) correlation = 0.7	100	99.8	100	98.4	99.2	99.8

Twelve Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the Hotelling test (α = 0.05)			Simulated Percent Rejection of Balance Based on the Hotelling test (α = 0.01)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	100	100	100	100	100	100
Independent U(-3, 3)	100	100	100	99.2	100	100
Independent U(-1, 1)	16.0	38.2	36.8	7.2	19.8	20.6
Independent N(0, 6)	-	100	100	-	100	100
Independent N(0, 3)	100	100	100	100	100	100
Independent N(0, 1)	73.4	97.4	98.2	61.2	96.0	96.4
N(0,1) correlation = 0.3	100	100	100	100	100	100
N(0,1) correlation = 0.5	100	100	100	100	100	100
N(0,1) correlation = 0.7	100	100	100	100	100	100

Notes: Based on 500 replications. For the case of 12 N(0, 6) covariates and *n* = 500, there was a lack of common support for many of the replications.

Table 8: Nearest Neighbour Matching and using Standardised Differences as a Check for Balance After Matching

Two Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on Any Standardised Differences > 20			Simulated Percent Rejection of Balance Based on Any Standardised Differences > 40		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	67.4	52.2	34.4	20.0	12.2	3.6
Independent U(-3, 3)	11.6	1.2	0	0.2	0	0
Independent U(-1, 1)	1.8	0	0	0	0	0
Independent N(0, 6)	99.0	97.4	98.2	80.8	78.2	67.2
Independent N(0, 3)	81.0	73.0	53.6	43.6	29.8	12.6
Independent N(0, 1)	8.2	0.6	0.2	0	0	0
N(0,1) correlation = 0.3	12.0	1.4	0	0.4	0	0
N(0,1) correlation = 0.5	9.6	1.4	0	0	0	0
N(0,1) correlation = 0.7	7.2	0.4	0.6	0.2	0	0

Six Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on Any Standardized Differences > 20			Simulated Percent Rejection of Balance Based on Any Standardised Differences > 40		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	100	100	100	99.2	98.4	99.6
Independent U(-3, 3)	96.6	96.6	90.6	70	63.6	42.8
Independent U(-1, 1)	36.0	9.2	1.2	0.8	0	0
Independent N(0, 6)	100	100	100	99.8	100	98.6
Independent N(0, 3)	99.8	100	99.8	93.8	93.8	93.4
Independent N(0, 1)	73.8	60.8	36.6	17.2	8.2	0.6
N(0,1) correlation = 0.3	97.2	95.8	87.8	67.2	52.8	26.2
N(0,1) correlation = 0.5	99.4	98.6	95.4	82.8	73.6	52.8
N(0,1) correlation = 0.7	100	99.6	99.0	95.2	88.2	75.4

Twelve Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on Any Standardised Differences > 20			Simulated Percent Rejection of Balance Based on Any Standardised Differences > 40		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	100	100	100	100	100	100
Independent U(-3, 3)	100	100	100	98.6	99.2	96.8
Independent U(-1, 1)	86.4	72.8	27.0	13.4	3.8	0.2
Independent N(0, 6)	-	100	100	-	100	100
Independent N(0, 3)	100	100	100	100	100	100
Independent N(0, 1)	98.8	99.6	96.2	65.0	70.8	47.0
N(0,1) correlation = 0.3	100	100	100	99.4	100	98.8
N(0,1) correlation = 0.5	100	100	100	100	100	99.8
N(0,1) correlation = 0.7	100	100	100	100	100	99.8

Notes: Based on 500 replications. For the case of 12 N(0, 6) covariates and *n* = 500, there was a lack of common support for many of the replications.

Table 9: Kernel Matching and using *t*-tests as a Check for Balance After Matching

Two Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the t-test (α = 0.05)			Simulated Percent Rejection of Balance Based on the t-test (α = 0.01)			Simulated Percent Rejection of Balance Based on the t-test (α = 0.05/2 = 0.025)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	65.2	86.0	98.6	44.6	65.2	93.2	53.6	75.6	97.4
Independent U(-3, 3)	3.4	7.0	17.2	0.6	0.4	0.6	1.6	1.8	6.0
Independent U(-1, 1)	0	0	0	0	0	0	0	0	0
Independent N(0, 6)	98.6	99.4	100	90.4	98.0	100	96.0	99.2	100
Independent N(0, 3)	65.2	74.8	97.0	45.0	51.2	79.8	54.0	62.8	92.6
Independent N(0, 1)	0.6	0.2	1.8	0.4	0	0	0.4	0	0
N(0,1) correlation = 0.3	0.6	0.8	7.2	0.2	0	0.4	0.4	0.2	2.2
N(0,1) correlation = 0.5	1.6	3.0	16.0	0	0.4	1.2	0.2	0.8	5.2
N(0,1) correlation = 0.7	0.8	3.2	32.8	0	0	2.6	0	1.0	11.4

Six Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the t-test (α = 0.05)			Simulated Percent Rejection of Balance Based on the t-test (α = 0.01)			Simulated Percent Rejection of Balance Based on the t-test (α = 0.05/6 = 0.0083)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	99.6	100	100	97.8	99.6	100	97.8	99.4	100
Independent U(-3, 3)	79.4	92.2	97.0	56.2	70.8	90.2	53.4	68.2	88.0
Independent U(-1, 1)	0.2	0.4	0	0.2	0.2	0	0.2	0.2	0
Independent N(0, 6)	100	100	100	99.8	100	100	99.8	100	100
Independent N(0, 3)	97.6	100	100	90.8	97.4	99.0	89.8	97.2	99.0
Independent N(0, 1)	21.2	30.4	48.2	5.8	9.0	17.2	4.8	7.6	15.4
N(0,1) correlation = 0.3	81.8	94.2	99.8	59.8	81.0	95.6	57.2	79.4	95.2
N(0,1) correlation = 0.5	96.8	100	100	87.4	97.8	99.8	85.4	97.4	99.8
N(0,1) correlation = 0.7	99.8	99.8	100	98.8	99.8	100	98.6	99.8	100

Twelve Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the t-test (α = 0.05)			Simulated Percent Rejection of Balance Based on the t-test (α = 0.01)			Simulated Percent Rejection of Balance Based on the t-test (α = 0.05/12 = 0.00416)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	100	100	100	100	100	100	100	100	100
Independent U(-3, 3)	99.6	100	100	96.0	99.2	99.4	94.8	98.0	98.8
Independent U(-1, 1)	9.6	26.2	18.4	1.6	6.4	2.8	0.4	2.4	0.8
Independent N(0, 6)	-	100	100	-	100	100	-	100	100
Independent N(0, 3)	100	100	100	100	100	100	99.8	100	100
Independent N(0, 1)	70.6	94.2	96.6	39.6	75.2	80.2	29.0	62.8	67.2
N(0,1) correlation = 0.3	100	100	100	99.0	100	100	98.4	100	100
N(0,1) correlation = 0.5	100	100	100	100	100	100	100	100	100
N(0,1) correlation = 0.7	100	100	100	100	100	100	100	100	100

Notes: Based on 500 replications. The smaller level of α in the third set of columns is meant to be an approximate Bonferroni adjustment for multiple comparisons at the α = 0.05 level. For the case of 12 N(0, 6) covariates and *n* = 500, there was a lack of common support for many of the replications.

Table 10: Kernel Matching and using Hotelling-Tests as a Check for Balance After Matching

Two Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the Hotelling test (α = 0.05)			Simulated Percent Rejection of Balance Based on the Hotelling test (α = 0.01)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	73.2	90.6	99.6	57.2	78.0	98.2
Independent U(-3, 3)	1.4	0.8	13.4	0.2	0	0.6
Independent U(-1, 1)	0	0	0	0	0	0
Independent N(0, 6)	100	100	100	99.2	100	100
Independent N(0, 3)	73.0	91.6	99.4	51.8	77.2	97.2
Independent N(0, 1)	0.4	0	0	0.2	0	0
N(0,1) correlation = 0.3	0.4	0	1.2	0.2	0	0.2
N(0,1) correlation = 0.5	0.2	0.6	2.6	0.2	0	0.2
N(0,1) correlation = 0.7	0.6	1.2	6.8	0	0	0.4

Six Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the Hotelling test (α = 0.05)			Simulated Percent Rejection of Balance Based on the Hotelling test (α = 0.01)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	100	100	100	100	100	100
Independent U(-3, 3)	74.4	94.2	100	57.8	86.2	99.0
Independent U(-1, 1)	0	0	0	0	0	0
Independent N(0, 6)	100	100	100	100	100	100
Independent N(0, 3)	98.8	100	100	96.8	99.8	100
Independent N(0, 1)	5.8	14.4	25.6	0.8	6.4	10.0
N(0,1) correlation = 0.3	63.8	88.0	99.0	49.8	76.2	95.0
N(0,1) correlation = 0.5	89.4	97.8	100	79.6	94.6	99.6
N(0,1) correlation = 0.7	98.8	99.8	100	96.4	99.2	100

Twelve Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on the Hotelling test (α = 0.05)			Simulated Percent Rejection of Balance Based on the Hotelling test (α = 0.01)		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	100	100	100	100	100	100
Independent U(-3, 3)	98.8	100	100	97.0	100	100
Independent U(-1, 1)	0.6	2.0	1.0	0.2	0.4	0.2
Independent N(0, 6)	-	100	100	-	100	100
Independent N(0, 3)	100	100	100	100	100	100
Independent N(0, 1)	35.6	88.4	94.2	22.2	75.4	84.8
N(0,1) correlation = 0.3	100	100	100	98.8	100	100
N(0,1) correlation = 0.5	100	100	100	100	100	100
N(0,1) correlation = 0.7	100	100	100	100	100	100

Notes: Based on 500 replications. For the case of 12 N(0, 6) covariates and *n* = 500, there was a lack of common support for many of the replications.

Table 11: Kernel Matching and using Standardised Differences as a Check for Balance After Matching

Two Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on Any Standardised Differences > 20			Simulated Percent Rejection of Balance Based on Any Standardised Differences > 40		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	67.2	62.4	50.0	14.0	7.4	1.0
Independent U(-3, 3)	4.8	0.2	0	0	0	0
Independent U(-1, 1)	0	0	0	0	0	0
Independent N(0, 6)	99.8	99.4	100	77.4	69.6	61.8
Independent N(0, 3)	83.4	67.4	62.4	32.0	15.4	2.4
Independent N(0, 1)	2.8	0	0	0.2	0	0
N(0,1) correlation = 0.3	6.2	0.6	0.4	0.2	0	0
N(0,1) correlation = 0.5	9.2	2.0	0.4	0	0	0
N(0,1) correlation = 0.7	9.0	1.4	0.6	0	0	0

Six Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on Any Standardised Differences > 20			Simulated Percent Rejection of Balance Based on Any Standardised Differences > 40		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	100	99.8	99.0	92.6	84.8	63.0
Independent U(-3, 3)	89.4	82.8	65.0	32.2	16.0	2.2
Independent U(-1, 1)	1.4	0.2	0	0	0	0
Independent N(0, 6)	100	100	100	99.0	97.2	91.6
Independent N(0, 3)	100	99.8	98.2	84.0	76.2	46.8
Independent N(0, 1)	39.4	19.4	4.6	2.0	0.6	0
N(0,1) correlation = 0.3	93.2	91.4	84.8	43.2	17.4	2.8
N(0,1) correlation = 0.5	99.2	99.6	99.0	72.4	55.8	23.6
N(0,1) correlation = 0.7	100	99.8	100	94.8	86.6	72.0

Twelve Covariates

Covariate distribution	Simulated Percent Rejection of Balance Based on Any Standardised Differences > 20			Simulated Percent Rejection of Balance Based on Any Standardised Differences > 40		
	n = 500	*n* = 1,000	*n* = 2,000	*n* = 500	*n* = 1,000	*n* = 2,000
Independent U(-6, 6)	100	100	100	100	99.8	95.8
Independent U(-3, 3)	100	99.8	98.2	94.6	83.6	40.4
Independent U(-1, 1)	33.2	19.8	0.4	0.6	0.2	0
Independent N(0, 6)	-	100	100	-	100	99.8
Independent N(0, 3)	100	100	100	99.8	98.8	91.8
Independent N(0, 1)	92.6	92.4	66.2	31.0	22.6	2.2
N(0,1) correlation = 0.3	100	100	100	97.2	98.6	86.2
N(0,1) correlation = 0.5	100	100	100	99.8	99.8	99.8
N(0,1) correlation = 0.7	100	100	100	100	100	100

Notes: Based on 500 replications. For the case of 12 N(0, 6) covariates and *n* = 500, there was a lack of common support for many of the replications.

Table 12: Monte Carlo Results for the DW test on the NSW-PSID Data – Percentage of Times Balance is Not Achieved (2000 replications)

Specification	Number of variables	Bonferroni adjusted *t*-test level	Percentage of Times Balance is Not Achieved
DW 1999	13	0.05/(13x5) = 0.00076	23.8
DW 2002	15	0.05/(15x5) = 0.00066	19.4
DW 2005	9	0.05/(9x5) = 0.0011	22.9
DW 1999 without *Nodeg*, *Black* and *Hisp*	9	0.05/(9x5) = 0.0011	5.9

Note: The specification without the difficult to balance variables *Nodeg*, *Black* and *Hisp* is: $prob(D = 1 \mid X) = F(\text{age, age}^2\text{, educ, educ}^2\text{, married, RE74, RE74}^2\text{, RE75, RE75}^2)$, where *F* is the cumulative logistic distribution.

Table 13: Monte Carlo Results for the Test for Standardised Differences using the DW 1999 specification (2000 replications)

Variable	Standardised Difference Before Matching	Average Standardised Difference After Nearest Neighbour Matching	Average Standardised Difference After Kernel Matching (Gaussian kernel)
Age	-54.35	-34.66	-30.99
Educ	-26.70	-8.92	-9.28
Married	-123.78	-61.88	-63.24
Nodeg	52.07	27.40	28.72
Black	90.16	47.88	47.76
Hisp	0.81	-8.58	-5.17
RE74	-109.31	-64.36	-70.05
RE75	-111.59	-55.94	-63.22
Decision rule		Percent of times balance is rejected using nearest neighbour matching	Percent of times balance is rejected using kernel matching
Reject if any SD > 20		99.9	100
Reject if any SD > 40		98.7	99.9

Table 14: Monte Carlo Results for the *t*-Test After Matching using the DW 1999 specification (2000 replications)

Variable	*p*-Value of *t*-Test Before Matching	Average *p*-Value of *t*-Test After Nearest Neighbour Matching	Average *p*-Value of *t*-Test After Kernel Matching (Gaussian kernel)
Age	0.000	0.282	0.334
Age^2	0.000	0.332	0.378
Educ	0.019	0.245	0.286
$Educ^2$	0.005	0.251	0.293
Married	0.000	0.513	0.485
Nodeg	0.000	0.234	0.249
Black	0.000	0.491	0.540
Hisp	0.663	0.360	0.342
RE74	0.000	0.621	0.613
RE75	0.000	0.529	0.473
$RE74^2$	0.012	0.572	0.665
$RE75^2$	0.018	0.589	0.543
*Black*U74*	0.000	0.568	0.588
Decision rule		Percent of times balance is rejected after nearest neighbour matching	Percent of times balance is rejected after kernel matching
Reject if any p-value < 0.05		81.9	75.7
Reject if any p-value < 0.0038		52.6	42.5

Note: The second critical p-value used is a Bonferroni adjusted p-value (0.05/13 = 0.0038).

Table 15: Monte Carlo Results for the Hotelling Test After Matching using the DW 1999 specification (2000 replications)

Decision rule	Percent of times balance is rejected after nearest neighbour matching	Percent of times balance is rejected after kernel matching
Reject if p-value < 0.05	76.90	58.35
Reject if p-value < 0.01	65.95	44.65

Table 16: Monte Carlo Results for the Permutation Version of the DW test on the NSW-PSID Data – Percentage of Times Balance is Not Achieved (500 replications)

Specification	Number of variables	Bonferroni adjusted *t*-test level	Percentage of Times Balance is Not Achieved
DW 1999	13	0.05/(13x5) = 0.00076	1.2

Table 17: Monte Carlo Results for the Permutation Versions of After Matching Tests using the DW 1999 specification (500 replications)

Decision rule	Percent of times balance is rejected after nearest neighbour matching	Percent of times balance is rejected after kernel matching
Permuted t-test		
Reject if any p-value < 0.05	26.2	26.0
Reject if any p-value < 0.01	5.6	5.2
Reject if any p-value < 0.0038	2.2	1.8
Permuted Hotelling test		
Reject if p-value < 0.05	0.2	0
Reject if p-value < 0.01	0	0

Note: The third critical p-value used for the permuted *t*-test is a Bonferroni adjusted p-value (0.05/13 = 0.0038).

Table 18: Analysis of Power and Bias of the Treatment Effect for the Before Matching DW Test (Permutation Version)

	Percentage of Simulations Balance Rejected	Simulated Treatment Effect when Balanced	Simulated Treatment Effect when not Balanced
DGP 1: RE74 in *p(X)*, RE74 in Y			
No balancing test used	-	\$349.59 (306.30)	-
p-value < 0.0125	3.6%	\$349.50 (308.44)	\$351.97 (249.26)
DGP 2: RE74 in *p(X)*, RE74 not in Y			
No balancing test used	-	\$886.96 (261.99)	-
p-value < 0.0125	3.6%	\$886.58 (263.53)	\$897.24 (223.22)
DGP 3: RE74 not in p(X), RE74 in Y			
No balancing test used	-	\$877.82 (283.32)	-
p-value < 0.0125	1.8%	\$878.54 (282.78)	\$838.57 (327.43)

Notes: Standard deviation in parentheses. Based on 500 replications with 1000 permutations of the test statistic. In each instance, *p(X)* is estimated without RE74. The true treatment effect is \$1000. The critical p-value used for the permuted *t*-test is a Bonferroni adjusted p-value (0.05/(5x8)) = 0.00125.

Table 19: Analysis of Power and Bias of the Treatment Effect for the After Nearest Neighbour Matching t-test (Permutation Version)

	Percentage of Simulations Balance Rejected	Simulated Treatment Effect when Balanced	Simulated Treatment Effect when not Balanced
DGP 1: RE74 in *p(X)*, RE74 in Y			
No balancing test used	-	$557.16 (389.76)	-
p-value < 0.05	14.2%	$533.56 (380.56)	$699.73 (416.39)
p-value < 0.01	1.0%	$558.47 (390.83)	$427.14 (255.52)
p-value < 0.00625	0.8%	$558.19 (390.49)	$428.91 (295.02)
DGP 2: RE74 in p(X), RE74 not in Y			
No balancing test used	-	$1017.33 (335.64)	-
p-value < 0.05	14.2%	$997.27 (325.92)	$1138.52 (369.04)
p-value < 0.01	1.0%	$1019.51 (336.07)	$801.35 (214.92)
p-value < 0.00625	0.8%	$1019.49 (335.73)	$749.45 (208.88)
DGP 3: RE74 not in p(X), RE74 in Y			
No balancing test used	-	$1065.11 (360.88)	-
p-value < 0.05	12.8%	$1050.61 (359.72)	$1163.89 (355.93)
p-value < 0.01	0.2%	$1063.54 (359.53)	$1848.79 (-)
p-value < 0.00625	0.2%	$1063.54 (359.53)	$1848.79 (-)

Notes: Standard deviation in parentheses. Based on 500 replications with 1000 permutations of the test statistic. In each instance, *p(X)* is estimated without RE74. The true treatment effect is $1000. The third critical p-value used for the permuted *t*-test is a Bonferroni adjusted p-value (0.05/8) = 0.00625.

Table 20: Analysis of Power and Bias of the Treatment Effect for the After Nearest Neighbour Matching Hotelling Test (Permutation Version)

	Percentage of Simulations Balance Rejected	Simulated Treatment Effect when Balanced	Simulated Treatment Effect when not Balanced
DGP 1: RE74 in *p(X)*, RE74 in Y			
p-value < 0.05	0.2%	$557.91 (389.79)	183.15 (-)
p-value < 0.01	0%	$557.16 (389.76)	-
DGP 2: RE74 in *p(X)*, RE74 not in Y			
p-value < 0.05	0.2%	$1017.28 (335.97)	1042.48 (-)
p-value < 0.01	0%	$1017.33 (335.64)	-
DGP 3: RE74 not in *p(X)*, RE74 in Y			
p-value < 0.05	0%	$1065.11 (360.88)	-
p-value < 0.01	0%	$1065.11 (360.88)	-

Note: Standard deviation in parentheses.

Table 21: Analysis of Power and Bias of the Treatment Effect for the After Kernel Matching t-test (Permutation Version)

	Percentage of Simulations Balance Rejected	Simulated Treatment Effect when Balanced	Simulated Treatment Effect when not Balanced
DGP 1: RE74 in *p(X)*, RE74 in Y			
No balancing test used	-	\$218.97 (296.63)	-
p-value < 0.05	11.2%	\$229.02 (279.71)	\$139.28 (401.06)
p-value < 0.01	0.8%	\$219.94 (294.88)	\$98.86 (518.26)
p-value < 0.00625	0.4%	\$220.41 (295.19)	-\$140.16 (584.28)
DGP 2: RE74 in p(X), RE74 not in Y			
No balancing test used	-	\$754.78 (256.28)	-
p-value < 0.05	11.2%	\$765.84 (244.16)	\$667.14 (326.99)
p-value < 0.01	0.8%	\$756.76 (254.55)	\$508.92 (390.44)
p-value < 0.00625	0.4%	\$756.08 (254.98)	\$432.48 (503.71)
DGP 3: RE74 not in p(X), RE74 in Y			
No balancing test used	-	\$779.43 (279.81)	-
p-value < 0.05	8.6%	\$781.83 (268.38)	\$753.93 (383.98)
p-value < 0.01	0.4%	\$779.94 (279.81)	\$652.75 (353.95)
p-value < 0.00625	0%	\$779.43 (279.81)	-

Notes: Standard deviation in parentheses. Based on 500 replications with 1000 permutations of the test statistic. In each instance, *p(X)* is estimated without RE74. The true treatment effect is \$1000. The third critical p-value used for the permuted *t*-test is a Bonferroni adjusted p-value (0.05/8) = 0.00625.

Table 22: Analysis of Power and Bias of the Treatment Effect for the After Kernel Matching Hotelling Test (Permutation Version)

	Percentage of Simulations Balance Rejected	Simulated Treatment Effect when Balanced	Simulated Treatment Effect when not Balanced
DGP 1: RE74 in *p(X)*, RE74 in Y			
p-value < 0.05	0%	\$218.97 (296.63)	-
p-value < 0.01	0%	\$218.97 (296.63)	-
DGP 2: RE74 in *p(X)*, RE74 not in Y			
p-value < 0.05	0%	\$754.78 (256.28)	-
p-value < 0.01	0%	\$754.78 (256.28)	-
DGP 3: RE74 not in *p(X)*, RE74 in Y			
p-value < 0.05	0%	\$779.43 (279.81)	-
p-value < 0.01	0%	\$779.43 (279.81)	-

Note: Standard deviation in parentheses.

References

Abadie, A. (2002). “Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models.” *Journal of the American Statistical Association*, 97, pp. 284-292.

Becker, S. and A. Ichino. (2002). “Estimation of Average Treatment Effects Based on Propensity Scores.” *Stata Journal*, 2(4), pp. 358-377

Blair, R., J. Higgins, W. Karniski and J. Kromrey. (1994). “A Study of Multivariate Permutation Tests Which May Replace Hotelling’s T^2 Test in Prescribed Circumstances.” *Multivariate Behavioral Research*, 29, pp. 141-163.

Cochran, W. (1968). “The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies.” *Biometrics*, 14, pp. 295-313.

Dehejia, R. (2005a). “Practical Propensity Score Matching: A Reply to Smith and Todd.” *Journal of Econometrics*, 125, pp. 355-364.

Dehejia, R. (2005b). “Does Matching Overcome Lalonde’s Critique of Non-Experimental Estimators? A Postscript.” Manuscript.

Dehejia, R. and S. Wahba. (1999). “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs.” *Journal of the American Statistical Association*, 94, pp. 1053-1062.

Dehejia, R. and S. Wahba. (2002). “Propensity Score Matching Methods for Nonexperimental Causal Studies.” *Review of Economics and Statistics*, 84(1), 151-161.

Diamond, A. and J. Sekhon. (2005). “Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies.” Unpublished manuscript, Dept. of Government, Harvard University.

Drake, C. (1993). “Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect.” *Biometrics*, 49, pp. 1231-1236.

Efron, B. and R. Tibshirani. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Frölich, M. (2004). “Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators.” *Review of Economics and Statistics*, 86, pp. 77-90.

Good, P. (2001). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York: Springer.

Hansen, B. (2006). “Appraising Covariate Balance After Assignment to Treatment by Groups.” University of Michigan, Technical Report #436.

Ham, J., Li Xianghong, and P. Reagan. (2003). “Propensity Score Matching, a Distance-Based Measure of Migration, and the Wage Growth of Young Men.” Unpublished manuscript, Dept. of Economics, Ohio State University.

Ho, D., K. Imai, G. King, and E. Stuart. (2005). “Matching as Nonparametric Preprocessing for Improving Parametric Causal Inference.” Unpublished manuscript, Dept. of Government, Harvard University.

Heckman, J. (1979). “Sample Selection Bias as a Specification Error.” *Econometrica*, 47, pp. 153-161.

Heckman, J. and R. Robb. (1986). “Alternative Identifying Assumptions in Econometric Models of Selection Bias,” in *Advances in Econometrics: Innovations in Quantitative Economics, Essays in Honor of Robert L. Basmann* (Vol. 5), ed. D. Slottje, Greenwich, CT: JAI Press, pp. 243-287.

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

Heckman, J., H. Ichimura, and P. Todd. (1997), “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme,” *Review of Economic Studies*, 64, pp. 605-654.

Heckman, J., H. Ichimura, J. Smith, and P. Todd. (1998). “Characterizing Selection Bias Using Experimental Data.” *Econometrica* 66(5), pp. 1017-1098.

Hosmer, D. and S. Lemeshow. (1980). “Goodness of Fit Tests for the Multiple Logistic Regression Model.” *Communications in Statistics – Theory and Methods*, A9, pp. 1043-1069

Hosmer, D., S. Lemeshow and J. Klar. (1988). “Goodness-of-fit Testing for the Logistic Regression Model when the Estimated Probabilities are Small.” Biometrical Journal, 30, pp. 911-924.

Imai, K., G. King and E. Stuart. (2006). “The Balance Test Fallacy in Matching Methods for Causal Inference.” Unpublished manuscript, Dept. of Government, Harvard University.

Imbens, G. (2004). “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review.” *Review of Economics and Statistics*, 86, pp. 4-29.

Kennedy, P. (1995). “Randomization Tests in Economics.” *Journal of Business and Economic Statistics*, 13, pp. 85-94.

Lalonde, R. (1986). “Evaluating the Econometric Evaluations of Training Programs.” *American Economic Review*, 76, pp. 604-620.

Landwehr, J., D. Pregibon, and A. Shoemaker. (1984). “Graphical Methods for Assessing Logistic Regression Models.” *Journal of the American Statistical Association*, 79, pp. 61-71.

Michalopoulos, C., H. Bloom, and C. Hill. (2004). “Can Propensity Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?” *Review of Economics and Statistics*, 86, pp. 156-179.

Rosenbaum, P. (1987). “Model-Based Direct Adjustment.” *Journal of the American Statistical Association*, 82, pp. 387-394.

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

Rosenbaum, P. and D. Rubin. (1984). “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score,” *Journal of the American Statistical Association*, 79, pp. 516-524.

Rosenbaum, P. and D. Rubin. (1985). “Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score,” *American Statistician*, 3, pp. 33-38.

Rubin, D. (1997). “Estimating Causal Effects from Large Data Sets using Propensity Scores.” *Annals of Internal Medicine*, 127, pp. 757-763.

Sekhon, J. (2006). "Alternative Balance Metrics for Bias Reduction in Matching Methods for Causal Inference." Unpublished manuscript, Dept. of Political Science, UC Berkeley.

Smith, J. and P. Todd. (2005a). "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, 125, pp. 305-353 (with discussion).

Smith, J. and P. Todd. (2005b). "Rejoinder" to Dehejia (2005a). *Journal of Econometrics*, 125, pp. 365-375.

Westfall, P. and S. Young. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* New York: John Wiley.

Zhao, Z. (2004). "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence." *Review of Economics and Statistics*, 86, pp. 91-107.

Appendix A: Accounting for Multiple Comparisons

When performing the DW test, which is a *t*-test for differences in each covariate mean within blocks of the propensity score, the true level of the test is no longer at the specified α level. Consider the case when there are *m* independent tests to be made at the level α, and that a joint decision is declared to be correct only if all its parts are correct. In order for the joint decision to be correct, all the null hypotheses have to be true. Thus, for the case of *m* decisions,

$$prob(\text{joint decision is correct}) = prob(\text{all H}_0\text{'s are true}) = (1-\alpha)^m$$

This probability is called the joint confidence. The joint level of significance is defined as:

$$\begin{aligned}\alpha_{\text{joint}} &= 1 - \text{joint confidence}\\ &= prob(\text{joint type 1 error})\\ &= 1-(1-\alpha)^m\end{aligned}$$

Therefore, it is clear that with multiple tests, the chance of finding at least one significant result due to chance fluctuation increases. For example, suppose the significance level is set at 0.01, there are seven covariates and five blocks, so that there are 35 *t*-tests altogether. Then the probability one of the tests rejects the balancing property due to chance fluctuations is:

$$\alpha_{\text{joint}} = 1-(1-.01)^{35} = 0.297$$

Conversely, if the number of tests to be conducted, *m*, is known, one is often interested in knowing how to set α in order make α_{joint} some specified value. To do this, we solve for α in the equation above.

$$\begin{aligned}\alpha_{\text{joint}} &= 1-(1-\alpha)^m\\ \alpha &= 1-(1-\alpha_{\text{joint}})^{1/m}\end{aligned}$$

For example, with 35 independent comparisons to be made, to maintain $\alpha_{\text{joint}} = 0.01$, one should set α to be:

$$\begin{aligned}\alpha &= 1-(1-\alpha_{\text{joint}})^{1/m}\\ &= 1-(1-.01)^{1/35}\\ &= 0.0002871\end{aligned}$$

This is the basis of the Bonferroni correction, which suggests that the chance of rejection of each individual test should be adjusted downwards to keep the overall chance of incorrect rejection at a predefined level. A quick Bonferroni approximation that can be used in practice is to divide the chosen α level by the number of comparisons to be made, which in the example given, is 0.01/35 = 0.0002857. See Westfall and Young (1993) for more discussion of the issues involved in multiple testing.