Propensity Score Matching and Variations on the Balancing Test

Wang-Sheng Lee*

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

March 10, 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.

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. 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 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 across groups using *t*-tests (Rosenbaum and Rubin 1985), and (iv) testing for the joint equality of covariate means across 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

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

these different balancing tests can give different results. In section 5, we discuss the setup of our Monte Carlo simulations and results. Finally, section 6 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_A, we have

$$D \perp X_A$$

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

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

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 identically 0.5. 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,

$$D \perp X_A \mid B$$

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,

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

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_A.

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

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

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

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

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

and

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

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

[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.

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

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

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 are perfectly homogeneous in $p(X)$, theorem 2 of

[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).

[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.

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:

$$prob(X_i \mid D_i=1, p(X_i) = p) = prob(X_i \mid D_i=0, p(X_i) = p) = prob(X_i \mid p)$$

and

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

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.

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

[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.

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:

$$D \perp X \mid p(X) \tag{1}$$

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 across 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.

---

[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.

[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) and Lee (2004) for further clarification.

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:

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

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

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

Conceptually, verifying balance involves checking if (2) holds, usually after invoking the common support assumption.[10]

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* across 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

[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.

[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*.

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. The DW test as described in Dehejia and Wahba (2002) provided no formal guidance in how the strata are chosen. This criticism has been somewhat mitigated by Michalopoulos, Bloom and Hill (2004), who provide more detail in how the DW test can be implemented in practice.[12] 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 literature. Appendix A describes the intuition behind the Bonferroni correction – one of the more common ways of dealing with multiple comparisons.

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. Typically, the data are grouped into deciles, although fewer groups can be used if the sample size is smaller. 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. Finally, Tsiatis (1980) introduced a score test for the effect of indicator functions for subsets of the covariate space.

The Hosmer and Lemeshow (1980) test is known to be unreliable (low in power) if the number of observations in some of the groups are small. Similarly, a weakness of the score test of Tsiatis (1980) is that Tsiatis did not specify a method for choosing the number of subsets and the way of grouping. The DW test appears to possess both these weaknesses. But we are unaware of any work that has focused on estimating the properties of the DW test.

A careful reading and comparison of Dehejia and Wahba (1999) and Dehejia and Wahba (2002) reveals an important point not yet picked up by the literature – although the DW test is justifiable as a heuristic specification check when stratifying on propensity scores (because balance is checked for within the subclasses of the exact *same* sample to be used for estimating the average treatment effect), it is less appropriate as a specification check for the adequacy of the estimated propensity scores when matching approaches other than stratification are used. This is because the sample changes considerably when matching approaches other than stratification are used. For example, suppose there are n treatment units and nR comparison units (with $R \geq 1$). 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

[12] The DW test has been implemented using both the *t*-test (for the test that covariate mean differences in each strata are zero across groups) and the *F*-test or Hotelling test (for the test that covariate mean differences in each strata are jointly zero across groups). The Stata ado program *pscore* written by Becker and Ichino (2002) has an algorithm for implementing the DW test.

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

[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 full matching) and matching structure (one-to-one, one-to-*k*, kernel matching, optimal matching etc.).

might be necessary to conclude that propensity score matching methods cannot solve the selection problem.

In summary, a before matching balance test is really more of a specification test when used in conjunction with stratification on the propensity score. 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.

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:

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

In contrast, in Dehejia and Wahba (2002), the specification used is:

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

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

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

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 is trying to make is that there are many possible specifications that can pass the DW test.

[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.

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 across groups using *t*-tests, and (iii) testing for the joint equality of covariate means across 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 neigbour 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]

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:

[15] When using a *t*-test level of $\alpha = 0.005$, 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.

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

where for each covariate, $\bar{X}_T$ and $\bar{X}_C$ are the sample means for the full treatment and comparison groups, $\bar{X}_{TM}$ and $\bar{X}_{CM}$ 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. 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.

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)

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.

4.5 Summary of Balancing Tests on the NSW-PSID Data

Here is the summary of the four balancing tests we conducted on the NSW-PSID data. The first test was based on a sample of $n = 1{,}331$ (the common support region of $n = 2{,}675$, 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.

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

[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.

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 across 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).

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

[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:

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

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

$\sum_{k=1}^{K} X_k$ 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. 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

[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.

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]

5.2 Results for Outcomes based on Generated Data

We also analysed the differences in the bias on the outcome by whether the balancing test is passed or not. The outcome equation was specified using exactly the same *X*s as the selection equation for simplicity (no instrument is required in matching), even though theory does not require that it be so. The data were generated so that the true treatment effect is 5.

[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.

As we were only able to obtain good test sizes for the DW test and not for the after matching tests, we focus our analysis on using the DW test in this section. When we analysed the bias on the outcomes based on whether the DW test indicated balance or imbalance (table 12), based on the Bonferroni adjusted test level that appears to give the correct test size, we found that it appears to offer some protection in terms of being able to weed out the biased estimates. Imbalanced covariates appear to lead to more biased estimates than balanced covariates. For example, in the case of 12 N(0, 6) covariates and n = 2,000 (bottom panel), the bias on the outcome when the DW test was passed (4.82 – 5 = -0.18) was smaller than the bias when the DW test was not passed (8.32 – 5 = 3.32). But in the majority of cases considered, there did not appear to be any gain from using the DW test as the average treatment effect was often close to the true effect of 5, irrespective of whether the DW test suggested balance or imbalance.

6. Conclusions

This paper was motivated by Smith and Todd's (2005a) 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.

An important point made in this paper is that the DW test is best used as a specification test or balancing test when stratification on the propensity score is done. When used in this context, the DW test combined with a Bonferroni adjustment appears to have the correct size and also appears to be somewhat useful in weeding out potentially biased estimates of the average treatment effect. But more work is needed to better understand the power of the DW test when there is imbalance due to misspecification of the propensity score model (so that it is not true that $X \perp D \mid p(X)$ because a misspecified $p(X)$ is being used), as this is a situation that is likely to occur in practice.

Applying the DW test does not make much sense if any other matching algorithm other than stratification is to be used. This is because the matched sample is different from the original sample. Checking for balance in the matched sample should involve after matching balancing tests, like the test for standardised differences, the *t*-test, and the Hotelling test. These check for differences in the covariate means in the weighted data which have implicitly conditioned on the propensity score, and attempt to replicate the check for covariate balance often done in single-factor experiments. The use of rigid rules when applying these after matching balancing tests do not appear to work well when the number of covariates is greater than 2. This recommends against using the Hotelling test as an after matching balancing test. Rather than rejecting the notion of balance as long as one covariate is found to be imbalanced, it is perhaps better to use the test for standardised differences and the *t*-test to help see the reduction in covariate imbalance before and after matching. Judgement is then required by the analyst to determine if the two groups are similar enough in terms of observed covariates.

Although the four balancing tests studied in this paper capture some aspect of the balancing property of propensity scores, none of the tests really check for the distributional balance that is required in theory. This leaves scope for better balancing tests to be developed as diagnostics when using propensity score methods. Furthermore, while simulation results based on artificially generated data can be useful, it would be 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).

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

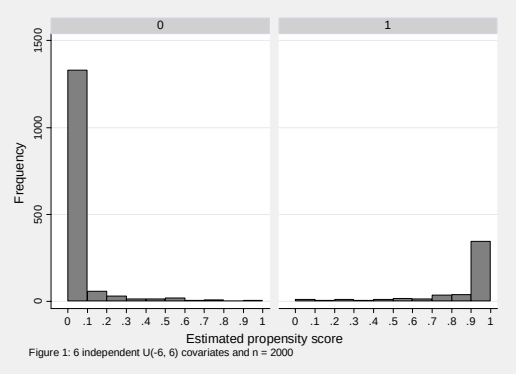

Figure 1: 6 independent U(-6, 6) covariates and n = 2000

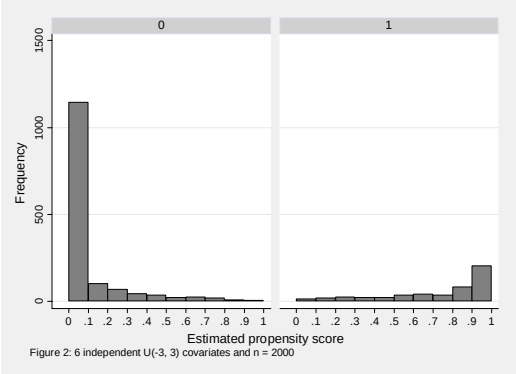

Figure 2: 6 independent U(-3, 3) covariates and n = 2000

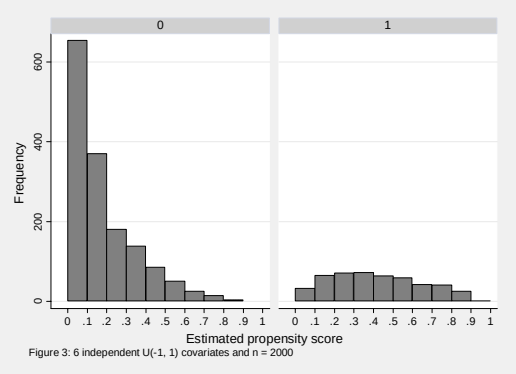

Figure 3: 6 independent U(-1, 1) covariates and n = 2000

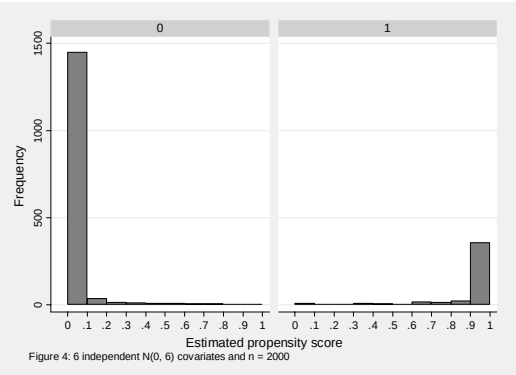

Figure 4: 6 independent N(0, 6) covariates and n = 2000

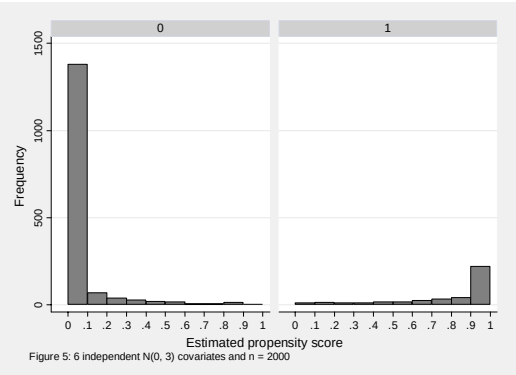

Figure 5: 6 independent N(0, 3) covariates and n = 2000

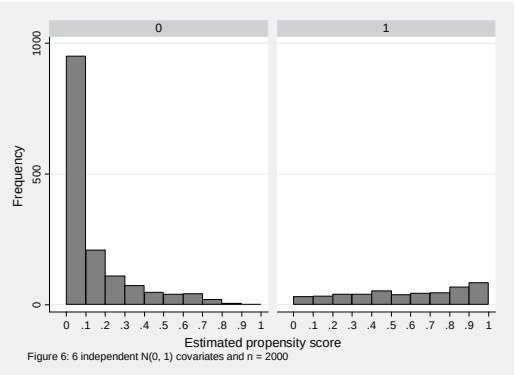

Figure 6: 6 independent N(0, 1) covariates and n = 2000

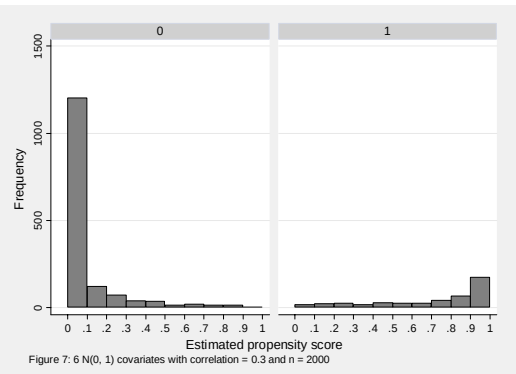

Figure 7: 6 N(0, 1) covariates with correlation = 0.3 and n = 2000

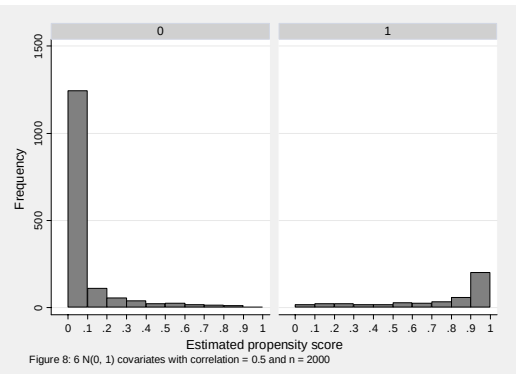

Figure 8: 6 N(0, 1) covariates with correlation = 0.5 and n = 2000

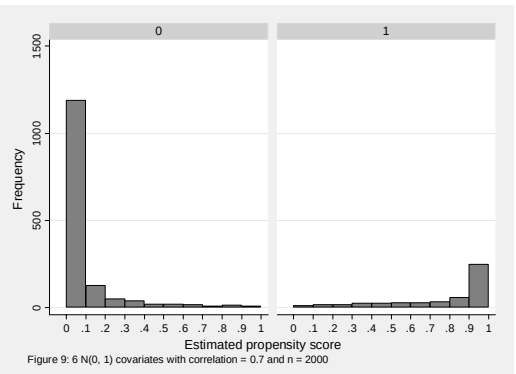

Figure 9: 6 N(0, 1) covariates with correlation = 0.7 and n = 2000

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 ($\alpha = 0.05$)			Simulated Percent Rejection of Balance Based on the DW test ($\alpha = 0.01$)			Simulated Percent Rejection of Balance Based on the DW test ($\alpha = 0.05/10 = 0.005$)		
	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 500$	$n = 1{,}000$	$n = 2{,}000$
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 ($\alpha = 0.05$)			Simulated Percent Rejection of Balance Based on the DW test ($\alpha = 0.01$)			Simulated Percent Rejection of Balance Based on the DW test ($\alpha = 0.05/30 = 0.0016$)		
	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 500$	$n = 1{,}000$	$n = 2{,}000$
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 ($\alpha = 0.05$)			Simulated Percent Rejection of Balance Based on the DW test ($\alpha = 0.01$)			Simulated Percent Rejection of Balance Based on the DW test ($\alpha = 0.05/60 = 0.0008$)		
	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 500$	$n = 1{,}000$	$n = 2{,}000$
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 $\alpha = 0.05$ 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 ($\alpha = 0.05$)			Simulated Percent Rejection of Balance Based on the Hotelling test ($\alpha = 0.01$)		
	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 500$	$n = 1{,}000$	$n = 2{,}000$
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 ($\alpha = 0.05$)			Simulated Percent Rejection of Balance Based on the Hotelling test ($\alpha = 0.01$)		
	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 500$	$n = 1{,}000$	$n = 2{,}000$
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 ($\alpha = 0.05$)			Simulated Percent Rejection of Balance Based on the Hotelling test ($\alpha = 0.01$)		
	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 500$	$n = 1{,}000$	$n = 2{,}000$
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 ($\alpha = 0.05$)			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)	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 ($\alpha = 0.05$)			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)	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 ($\alpha = 0.05$)			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)	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 $\alpha = 0.05$ 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 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 ($\alpha = 0.05$)			Simulated Percent Rejection of Balance Based on the Hotelling test ($\alpha = 0.01$)		
	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 500$	$n = 1{,}000$	$n = 2{,}000$
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 ($\alpha = 0.05$)			Simulated Percent Rejection of Balance Based on the Hotelling test ($\alpha = 0.01$)		
	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 500$	$n = 1{,}000$	$n = 2{,}000$
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 ($\alpha = 0.05$)			Simulated Percent Rejection of Balance Based on the Hotelling test ($\alpha = 0.01$)		
	$n = 500$	$n = 1{,}000$	$n = 2{,}000$	$n = 500$	$n = 1{,}000$	$n = 2{,}000$
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: Analysis of Bias of the Treatment Effect for the Stratification Method based on the DW Test with Bonferroni Adjustment

Two Covariates

Covariate distribution	Simulated Treatment Effect when passes the DW test			Simulated Treatment Effect when fails the DW test		
	n = 500	n = 1,000	n = 2,000	n = 500	n = 1,000	n = 2,000
Independent U(-6, 6)	4.88	5.19	5.01	3.91	5.01	5.05
Independent U(-3, 3)	5.01	5.00	5.01	5.06	5.09	5.03
Independent U(-1, 1)	5.01	4.99	4.99	4.94	4.98	4.98
Independent N(0, 6)	5.76	5.09	5.08	5.09	3.31	2.97
Independent N(0, 3)	4.74	4.97	5.02	4.18	4.29	4.42
Independent N(0, 1)	5.01	5.00	5.01	4.96	4.98	5.03
N(0,1) correlation = 0.3	5.00	5.00	5.00	5.08	4.97	4.99
N(0,1) correlation = 0.5	4.98	4.99	5.01	5.03	4.97	4.99
N(0,1) correlation = 0.7	4.99	4.99	4.99	4.98	4.93	4.99

Six Covariates

Covariate distribution	Simulated Treatment Effect when passes the DW test			Simulated Treatment Effect when fails the DW test		
	n = 500	n = 1,000	n = 2,000	n = 500	n = 1,000	n = 2,000
Independent U(-6, 6)	4.59	5.48	4.83	6.28	4.65	4.94
Independent U(-3, 3)	4.87	5.06	5.06	4.37	4.66	5.08
Independent U(-1, 1)	5.01	5.00	5.00	5.04	4.96	4.98
Independent N(0, 6)	4.21	5.76	5.16	6.99	8.98	4.05
Independent N(0, 3)	4.92	5.12	5.06	4.53	3.79	5.94
Independent N(0, 1)	5.01	5.00	5.01	5.02	5.06	5.06
N(0,1) correlation = 0.3	5.01	4.99	5.02	4.85	4.92	5.11
N(0,1) correlation = 0.5	5.06	5.04	5.01	5.38	4.88	4.99
N(0,1) correlation = 0.7	5.11	5.02	5.01	5.11	4.94	5.03

Twelve Covariates

Covariate distribution	Simulated Treatment Effect when passes the DW test			Simulated Treatment Effect when fails the DW test		
	n = 500	n = 1,000	n = 2,000	n = 500	n = 1,000	n = 2,000
Independent U(-6, 6)	5.26	5.88	4.50	7.84	5.89	7.90
Independent U(-3, 3)	4.81	5.14	4.99	5.78	6.52	5.13
Independent U(-1, 1)	4.97	4.99	5.00	4.89	4.95	5.00
Independent N(0, 6)	-	8.17	4.82	-	4.09	8.32
Independent N(0, 3)	6.00	5.47	5.02	5.38	6.81	7.64
Independent N(0, 1)	4.99	5.01	4.94	5.09	5.31	4.99
N(0,1) correlation = 0.3	4.96	5.04	4.88	4.99	5.05	5.65
N(0,1) correlation = 0.5	4.98	4.96	4.93	5.04	4.67	5.15
N(0,1) correlation = 0.7	4.94	4.90	4.93	4.27	5.29	5.09

Notes: Based on 500 replications. The true treatment effect is 5. For the case of 12 N(0, 6) covariates and n = 500, there was a lack of common support for many of the replications. The numbers in this table correspond to the Bonferroni adjusted test levels in the last three columns of table 5.

References

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

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.

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

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.

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

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.

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

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.

Lee, W. (2004). "Propensity Score Matching, Conditional Independence, the Pre-Program Test, and Graphical Models." Manuscript. University of Melbourne.

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.

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.

Tsiatis, A. (1980). "A Note on a Goodness of Fit Tests for the Logistic Regression Model." *Biometrika*, 67, pp. 250-251.

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:

$$\alpha_{\text{joint}} = 1-(1-.01)^{35} = .297$$

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.

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

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

$$\begin{aligned}\alpha &= 1-(1-\alpha_{\text{joint}})^{1/m} \\ &= 1-(1-.01)^{1/35} \\ &= 0.00029\end{aligned}$$

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 = 0.00029.