Universität St.Gallen

Testing for covariate balance using quantile regression and resampling methods

Martin Huber

June 2010 Discussion Paper no. 2010-18

Department of Economics University of St. Gallen

Editor: Martina Flockerzi
University of St. Gallen
Department of Economics
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35
Email vwaabtass@unisg.ch

Publisher: Department of Economics
University of St. Gallen
Varnbüelstrasse 19
CH-9000 St. Gallen
Phone +41 71 224 23 25
Fax +41 71 224 31 35

Electronic Publication: http://www.vwa.unisg.ch

Testing for covariate balance using quantile regression and resampling methods[1]

Martin Huber

Author's address:

Martin Huber, Ph.D.
SEW, University of St. Gallen
Varnbüelstrasse 14
9000 St. Gallen
Phone +41 71 224 2299
Fax +41 71 224 2302
Email Martin.Huber@unisg.ch
Website www.sew.unisg.ch

[1] I have benefited from comments by Eva Deuchert, Bernd Fitzenberger, Michael Lechner, Enno Mammen, Blaise Melly, Conny Wunsch, and conference/seminar participants in Linz (Annual Meeting of the Austrian Economic Association), London (cemmap conference on quantile regression), Geneva (Annual Meeting of the Swiss Society of Economics and Statistics), and Freiburg i.B. (seminar in empirical economics).

Abstract

Consistency of propensity score matching estimators hinges on the propensity score's ability to balance the distributions of covariates in the pools of treated and nontreated units. Conventional balance tests merely check for differences in covariates' means, but cannot account for differences in higher moments. Specification tests constitute an alternative, but might reject misspecified, but yet balancing propensity score models. This paper proposes balance tests based on (i) quantile regression to check for differences in the distributions of continuous covariates and (ii) resampling methods to estimate the distributions of the proposed Kolmogorov-Smirnov and Cramer-von-Mises-Smirnov test statistics. Simulations suggest that the tests capture imbalances related to higher moments when conventional balance tests fail to do so and correctly keep misspecified, but balancing propensity scores when specification tests reject the null.

Keywords

Balancing property, balance test, propensity score matching.

JEL Classification

C12, C15, C21

1 Introduction

Propensity score matching (Rosenbaum and Rubin, 1983, 1985) has become an increasingly popular estimation method in many fields of empirical research concerned with the evaluation of treatment effects in a conditional independence or selection on observables framework (see Imbens, 2004). Applications include the evaluation of active labor market policies (Heckman, Ichimura, and Todd, 1997, 1998), the estimation of the health effects of unemployment (Böckerman and Ilmakunnas, 2009), the evaluation of trade gains due to a common currency (Persson, 2001), and many others.

Propensity score matching (PSM) is attractive because it does not rely on tight functional form assumptions as parametric estimators, nor is it prone to the curse of dimensionality issue inherent in matching on a high dimensional covariate vector directly. However, one condition for consistency of PSM is the balancing property of the presumed propensity score model. It states that conditional on the propensity score, the distributions of the covariates in the pools of treated and nontreated units must be equal, i.e., balanced.

Most of the balancing tests suggested in the literature, as the DW test (see Dehejia and Wahba, 1999, 2002), the regression test of Smith and Todd (2005), or the two-sample t-test for matched samples, merely check for differences in the means of covariates. Thus, they might lack power when imbalances affect distributional features other than the mean. Specification tests as suggested by Shaikh, Simonsen, Vytlacil, and Yildiz (2009) constitute an alternative to balancing tests. Yet, there might exist propensity score models that balance the covariates despite the fact that they are misspecified, e.g., when they are only a monotonic transformation of the true propensity score. Such models would be unnecessarily rejected by powerful specification tests.

This paper contributes to the literature on balancing tests by suggesting test procedures for continuous covariates which account for differences in the entire quantile regression functions. In contrast to commonly applied mean difference tests, the proposed methods also capture distributional imbalances related to higher moments. The procedures are based on (i) quantile regression, (ii) the computation of Kolmogorov-Smirnov (KS) and Cramer-von-Mises-Smirnov (CMS) test statistics on the empirical inference process, and (iii) bootstrapping to estimate the distributions of the KS and CMS statistics in order to compute critical values and p-values. Furthermore, the paper discusses how to implement the tests as full sample tests (based on the entire sample) and after matching tests (based on the sample of matched units alone) and points to differences in the interpretation of the results. It also provides simulation evidence on the performance of the tests relative to existing balancing and specification tests. It therefore complements the analysis of the finite sample properties of balancing tests by Lee (2006) and extends the range of tests investigated.

The simulations show that the KS and CMS resampling procedures appear to be very competitive when implemented as full sample tests. They capture imbalances related to higher moments when conventional balancing tests fail to do so and correctly keep misspecified, but balancing propensity scores when specification tests reject the null. When implemented as after matching tests, they are again very powerful but reject the null too often when the balancing property holds, such that their relative performance to other tests is ambiguous.

The remainder of this paper is organized as follows. Section 2 motivates PSM and more formally discusses the condition to be tested. Section 3 reviews the DW test and the regression test of Smith and Todd (2005) and introduces full sample balancing tests for continuous covariates based on resampling and quantile regression. Section 4 discusses conventional after matching tests and introduces the procedures for matched samples. Section 5 presents simulation results about the finite sample properties of KS and CMS resampling tests and their competitors. Section 6 presents two empirical applications of full sample and after matching tests. Section 7 concludes.

2 Propensity score matching and testable conditions

In the treatment evaluation literature, identification strategies based on 'selection on observables' rely on the assumption that all factors jointly affecting the treatment probability and the outcome are observed and thus, can be controlled for. Hence, hypothetical outcomes that would have been realized under alternative treatment states are assumed to be independent of the actual treatment status conditional on the observed covariates. This is known as the conditional independence assumption (CIA), see for instance Imbens (2004) for an in-depth discussion. It implies that the effect of the treatment on the outcome is conditionally unconfounded. Let Y denote the outcome variable, D a binary treatment taking either the value 1 (treated) or 0 (nontreated),[1] and X a vector of observed covariates with the parameter space $\mathcal{X}$. The CIA states that

$$Y^1, Y^0 \perp D | X = x \quad \forall\, x \in \mathcal{X}, \tag{1}$$

[1] In contrast, Imbens (2000) and Lechner (2001) discuss effect evaluation for multiple treatments. The discussion in this paper could be easily extended to their framework.

where Y^1, Y^0 are the hypothetical outcomes for $D = 0, 1$ and $\perp$ denotes independence.

From a practitioner's perspective, conditioning on a high dimensional X may be problematic, as the number of possible combinations of elements in X increases exponentially in the dimension of X such that a precise estimation quickly becomes exorbitantly data hungry. In the literature, this problem is referred to as curse of dimensionality. Let $p^*(X) \equiv \Pr(D = 1|X)$ denote the unknown probability of being treated conditional on X, henceforth referred to as true propensity score. Rosenbaum and Rubin (1983) have shown that conditioning on the true propensity score is equivalent to conditioning on the covariates directly, as both X and $p^*(X)$ are balancing scores in the sense that they adjust the distributions of covariates in the treatment and in the control (or nontreated) group. Thus, if (1) is satisfied, it also holds that the hypothetical outcomes are independent of the treatment conditional on the propensity score:

$$Y^1, Y^0 \perp D|p^*(X). \tag{2}$$

Conditioning on the one dimensional propensity score rather than on the multidimensional vector of covariates circumvents the practical issues related to the curse of dimensionality, e.g., the occurrence of empty cells for particular combinations of covariates. For this reason, propensity score matching is frequently used in empirical applications. If (2) is satisfied, average treatment effects (ATEs) and quantile treatment effects (QTEs) can be consistently estimated, given that there is sufficient common support with respect to $p^*(X)$ among treated and nontreated units. The balancing property of $p^*(X)$ implies that

$$X \perp D|p^*(X). \tag{3}$$

Note that (3) is a mechanical result related to the balancing property and holds even if the CIAs (1) and (2) do not (such that the effect of D on Y is confounded). In the real world the structural form of the true propensity score is usually unknown to the researcher. In empirical applications it is most commonly modeled parametrically using probit or logit specifications. Let $p(X)$ denote the presumed specification of the true $p^*(X)$. Whereas the balancing property of $p^*(X)$ follows from the proof in Rosenbaum and Rubin (1983), it is a priori not clear whether $p(X)$ balances X in the pools of treated and nontreated units. However, the balancing property of $p(X)$ is testable by verifying whether

$$F_{X|D=1,p(X)=\rho}\left(x|D = 1, p(X) = \rho\right) = F_{X|D=0,p(X)=\rho}(x|D = 0, p(X) = \rho) \tag{4}$$
$$\forall\ x \in \mathcal{X},\ \forall\ \rho \in (0, 1),$$

where $F_{X|D=d,p(X)}(\cdot|D=d,p(X))$ denotes the conditional cdf of X given $D=d$ and $p(X)$. If (5) is satisfied, it holds that

$$X \perp D | p(X). \tag{5}$$

Instead of building tests for equality of the conditional distribution functions given $p(X)$[2] it is equally valid to test for differences in the conditional quantile functions for $D=1,0$, as the quantile function is simply the inverse of the distribution function. Let Q_A^τ represent the quantile at rank $\tau \in (0,1)$ for some variable A, $Q_A^\tau = \inf\{a : F_A(a) \geq \tau\}$. Then, $F_A(a) = Q_A^{\tau^{-1}}$. For $Q_X^\tau(d,\rho))$ denoting the τth conditional quantile of X given $D=d$ and $p(X)=\rho$, the balancing property implies that

$$Q_X^\tau(1,\rho) = Q_X^\tau(0,\rho), \quad \forall\ \tau,\rho\ \in\ (0,1),\ \forall\ x \in \mathcal{X}. \tag{6}$$

However, conventional balancing tests merely capture differences in means by verifying whether

$$E[X|D=1,p(X)=\rho] = E[X|D=0,p(X)=\rho],\ \forall\ x \in \mathcal{X},\ \forall\ \rho \in (0,1), \tag{7}$$

which is necessary, but not sufficient for (5). Therefore, these tests do not account for distributional differences related to higher moments and ignore valuable information that might point to the violation of covariate balance, see also the discussion in Sekhon (2007a). Furthermore, Lee (2006) provides simulation evidence that conventional balancing tests have poor size properties in their original forms where inference is based on asymptotic theory. He suggests to compute p-values using permuted test statistics by randomly shuffling the treatment and control labels 1,0 a large number of times in order to estimate the distribution of the respective test statistic nonparametrically. Even though permutation improves the finite sample properties, the permuted versions are as incapable to account for differences in higher moments as the original tests.

Specification tests for the propensity score model constitute an alternative to balancing tests. However, it is the balancing property and not the fit of $p(X)$ that is of interest when using PSM. Misspecification is innocuous as long as (5) is satisfied.[3] Beside the correct, but unknown model, there might exist a misspecified, but easy to estimate model that equally satisfies the balancing property and is chosen by the practitioner for the sake of econometric feasibility. This is the

[2] Testing for equality of conditional distributions is discussed in Li, Maasoumic, and Racine (2009), although for discrete conditioning variables, whereas we need to condition on a continuous $p(X)$.

[3] In contrast, estimators based on inverse probability weighting generally rely on the correctness of the propensity score model such that specification tests are relevant for this class of estimators.

case whenever the misspecified model is only a monotonic transformation of the true model, such that the order of the individual propensity scores remains unchanged. Simulation results in Zhao (2008) suggest that ATE estimates based on matching are hardly affected by misspecified, but balancing propensity scores as long as the CIA holds.

Thus, balancing tests appear to be more attractive than specification tests. However, for the reasons discussed it seems more appropriate to use procedures that capture imbalances in the entire distributions rather than in the means alone. The following sections will propose such tests for continuously distributed covariates that can be applied to full and matched samples.

3 Full sample tests

3.1 Balancing tests suggested in the literature

Balancing tests can be categorized into methods testing the balancing property (i) in the entire sample (thereafter referred to as full sample tests) or, after having applied the matching algorithm, (ii) in the sample of matched units alone (henceforth after matching tests). Two tests of the former kind are the DW test used in Dehejia and Wahba (1999, 2002), which is based on a process originally proposed by Rosenbaum and Rubin (1984) and Rubin (1997), and the regression test of Smith and Todd (2005).

Smith and Todd (2005) suggest regressing the kth element in the covariate vector X, denoted as X_k, on a quartic of the estimated propensity score $\hat{p}(X)$, the treatment state D, and interaction terms:

$$X_k = \beta_0 + \sum_{j=1}^{4} \beta_j \hat{p}(X)^j + \gamma_0 D + \sum_{j=1}^{4} \gamma_j D \hat{p}(X)^j + \epsilon, \tag{8}$$

where β, γ denote the coefficients and ϵ is the error term. After the regression a Wald-test is used to test whether the coefficients on the treatment dummy and the interaction terms, $\gamma_0, ..., \gamma_4$, are jointly equal zero. This would imply that D did not provide further information about the conditional mean of X_k given $\hat{p}(X)$, which is a necessary, albeit not sufficient condition for the balancing property.

The DW test is based on stratification on the propensity score. It uses t-tests to test for mean differences in (elements of) X across treated and nontreated units within strata with the same mean values of estimated propensity scores. The DW algorithm is provided in Dehejia and Wahba (2002):

1. Start with a parsimonious logit specification to estimate the score.

2. Sort data according to estimated propensity score (ranking from lowest to highest).

3. Stratify all observations such that estimated propensity scores within a stratum for treated and comparison units are close (no significant difference); for example, start by dividing observations into strata of equal score range ($0 - 0.2, \ldots, 0.8 - 1$).

4. Statistical test: for all covariates, differences in means across treated and comparison units within each stratum are not significantly different from zero.

a. If covariates are balanced between treated and comparison observations for all strata, stop.

b. If covariates are not balanced for some stratum, divide the stratum into finer strata and reevaluate.

c. If a covariate is not balanced for many strata, modify the logit by adding interaction terms and/or higher-order terms of the covariate and reevaluate.

Lee (2006) provides simulation evidence that the standard DW test has rather poor size properties. He suggests to estimate the distribution of the test statistic nonparametrically instead of approximating it by the asymptotic t-distribution. This is done by randomly shuffling the treatment and control labels 1,0 in the full sample a large number of rounds in order to compute the t-statistics in each round. Using the distribution of permuted test statistics allows computing the p-values of the test statistic obtained for the original sample. Lee (2006) demonstrates that permutation tests (see Pitman, 1937) have considerably better size properties in finite samples than those relying on asymptotic theory. However, one shortcoming of all conditional mean difference tests as the DW test and the regression test of Smith and Todd (2005) is that they do not account for differences in higher moments of the covariates. For this reason, we propose test statistics that capture differences in the entire covariate distribution.

3.2 Estimation

Let us assume that we would like to test whether some continuously distributed covariate is balanced conditional on $p(X)$. We denote the covariate by X_k, indicating that it is the k^{th} element in the covariate vector X. The null hypothesis is

$$H_0 : Q^{\tau}_{X_k}(1,\rho) = Q^{\tau}_{X_k}(0,\rho), \quad \forall\ \tau,\rho \ \epsilon\ (0,1), \tag{9}$$

i.e., that the conditional quantiles of X_k given $p(X)$ are equal across treatment states $D = 1,0$ at all ranks and for all values of the propensity score. This would imply that (5) holds.

The proposed test procedure can be divided into 3 steps. Prior to testing, we predict the propensity scores for the units in the sample based on the presumed model $p(X)$. The first step of testing consists of estimating the covariate's conditional quantiles. In the second step, Kolmogorov-Smirnov (KS) and Cramer-Von-Mises-Smirnov (CMS) statistics are computed based on the differences in the conditional quantiles across treatment states. Finally, we estimate the distributions of the test statistics based on bootstrapping, i.e., we draw a large number of bootstrap samples out of the original sample and compute the recentered test statistics for every draw. P-value are computed as the the share of bootstrapped statistics being larger or equal to the respective statistic for the original sample.

In the first step, we estimate $Q^{\tau}_{X_k}(1,\rho), Q^{\tau}_{X_k}(0,\rho)$ by regressing X on a constant and a polynomial of the propensity score estimate, e.g., on the score itself, its square and its cubic. Let $\hat{p}(X_i)$ denote the propensity score estimate for unit i and specification $p(X)$. For treatment state $d=1,0$, the quantile coefficients β^{τ}_{d} are estimated by solving the following minimization problem:

$$\hat{\beta}^{\tau}_{d} = \min_{\beta} \frac{1}{n_d} \sum_{i:D=d}^{n} \eta_{\tau} \left(X_{k,i} - \sum_{l=0}^{L} (\hat{p}^{l}(X_i))\beta \right), \tag{10}$$

where n_d is the number of observations with $D=d$. $\eta_{\tau}(v) = v(\tau - I\{v \leq 0\})$ is the check function, an asymmetric loss function, suggested by Koenker and Bassett (1978) in their seminal paper on quantile regression. By setting $L=3$ we regress X_k on a constant and the third order polynomial of the propensity score estimate. While we suspect this specification to be sufficiently flexible for a univariate regression, we also try lower orders in our simulations presented in Section 5. The conditional quantile of X_k given $D=d$ at $p(X)=\rho$ is predicted by

$$\hat{Q}^{\tau}_{X_k}(d,\rho) = \sum_{l=0}^{L} (\rho^{l})\hat{\beta}^{\tau}_{d}. \tag{11}$$

3.3 Testing

We would like to infer whether the process $Q^{\tau}_{X_k}(1,\rho) - Q^{\tau}_{X_k}(0,\rho)$, which is not observed, is different from zero. Instead, we observe the empirical inference process

$$\hat{Q}^{\tau}_{X_k}(1,\rho) - \hat{Q}^{\tau}_{X_k}(0,\rho), \tag{12}$$

i.e., the difference between the conditional quantile estimates. We use these differences to compute KS and CMS test statistics, denoted as T_n, which account for differences in the conditional quantile estimates across ranks (τ) of the covariate distribution and across propensity scores (ρ). Let n, n_1, n_0 denote the total sample size, the number of treated, and the number of nontreated observations, respectively. The KS statistic is based on the supremum of the difference across ranks and scores, the CMS statistic on the integration over the squared differences:

$$\begin{aligned} T_n^{KS} &= \sup_{\tau \in \mathcal{T}, p \in \mathcal{P}} \sqrt{\frac{n_1 \cdot n_0}{n}} ||\hat{Q}^{\tau}_{X_k}(1,\rho) - \hat{Q}^{\tau}_{X_k}(0,\rho)||_{\hat{\Lambda}}, \\ T_n^{CMS} &= \frac{n_1 \cdot n_0}{n} \int_{\mathcal{T}} \int_{\mathcal{P}} ||\hat{Q}^{\tau}_{X_k}(1,\rho) - \hat{Q}^{\tau}_{X_k}(0,\rho)||^2_{\hat{\Lambda}} d\tau d\rho. \end{aligned} \tag{13}$$

$\mathcal{T}, \mathcal{P}$ denote the parameter spaces of τ and $p(X)$ and are naturally bounded between 0 and 1. $||a||_{\hat{\Lambda}_\tau}$ denotes $\sqrt{a'\hat{\Lambda}a}$ and $\hat{\Lambda}$ is a positive weighting matrix satisfying $\hat{\Lambda} = \Lambda + o_p(1)$. Λ is positive definite, continuous and symmetric.

Two weighting schemes are considered in the simulations and applications. Firstly, we use the inverse of the variances of $\hat{Q}^{\tau}_{X_k}(1,\rho) - \hat{Q}^{\tau}_{X_k}(0,\rho)$, which we compute as a by-product of the resampling procedure described below. This seems to be a natural choice, as it gives more weight to differences in conditional quantiles that are precisely estimated. However, this choice need not be optimal with respect to the testing problem at hand, where we are primarily worried about differences that largely affect the estimation of the ATE (or other parameters of interest). We therefore also consider weighting by the densities of the predicted propensity scores. This gives more weight to differences in areas with high propensity score densities which supposedly have a higher impact on the estimation of the ATE. The optimal weighting matrix to be used is nevertheless an open issue that might be addressed in future research.

T_n^{KS}, T_n^{CMS} are non-pivotal and distribution-free in the sense that their distributions do not converge to any known distribution. For linear quantile regression processes such as the one considered in this paper, Chernozhukov and Fernandez-Val (2005) show in Theorem 1 that the distributions of T_n^{KS}, T_n^{CMS} can be consistently estimated by resampling the recentered test statistics under their Assumptions A.1-A.3. These assumptions state that the data are stationary and strongly mixing (which is satisfied in i.i.d. samples) and that the uniformly consistent parameters entering the null hypothesis, in our case the quantile coefficient estimates, are asymptotically Gaussian under local and global alternatives. Following their approach, we draw J samples of size n with replacement from the original sample. For each bootstrap sample we estimate the propensity scores and the conditional quantiles to compute the bootstrapped inference process

$$\hat{Q}^{\tau}_{X_k,j}(1,\rho) - \hat{Q}^{\tau}_{X_k,j}(0,\rho). \tag{14}$$

$\hat{Q}^{\tau}_{X_k,j}(1,\rho), \hat{Q}^{\tau}_{X_k,j}(0,\rho)$ denote the conditional quantile estimates for sample draw j, where $(1 \leq j \leq J)$. The corresponding KS and CMS statistics of the bootstrapped and recentered inference processes are

$$\begin{aligned}
T_{n,j}^{KS} &= \sup_{\tau\in\mathcal{T},p\in\mathcal{P}} \sqrt{\frac{n_1\cdot n_0}{n}} ||\hat{Q}^{\tau}_{X_k,j}(1,\rho) - \hat{Q}^{\tau}_{X_k,j}(0,\rho) - (\hat{Q}^{\tau}_{X_k}(1,\rho) - \hat{Q}^{\tau}_{X_k}(0,\rho))||_{\hat{\Lambda}}, \\
T_{n,j}^{CMS} &= \frac{n_1\cdot n_0}{n} \int_{\mathcal{T}}\int_{\mathcal{P}} ||\hat{Q}^{\tau}_{X_k,j}(1,\rho(x))) - \hat{Q}^{\tau}_{X_kj}(0,\rho)) - (\hat{Q}^{\tau}_{X_k}(1,\rho) - \hat{Q}^{\tau}_{X_k}(0,\rho))||^2_{\hat{\Lambda}} d\tau d\rho.
\end{aligned} \tag{15}$$

Note that these statistics differ slightly to Chernozhukov and Fernandez-Val (2005) in that $\frac{n1\cdot n0}{n}$ is used instead of n as we consider a two samples testing problem. Finally, we compute the p-values by $1/J\sum_{j=1}^{J} I\{T_{n,j} > T_n\}$ which is a consistent estimator of $\Pr[T(\hat{Q}^{\tau}_{X_k}(1,\rho) - \hat{Q}^{\tau}_{X_k}(0,\rho) - (Q^{\tau}_{X_k}(1,\rho) - Q^{\tau}_{X_k}(0,\rho))) > T_n]$.

One important difference to Chernozhukov and Fernandez-Val (2005) (where the regressors are known) is that we need to estimate the propensity score, which serves as the regressor in our test procedure. To the best of our knowledge no analytical results for resampling methods of statistics on quantile regression processes exist when the regressor is estimated. However, simulation results in Section 5 suggest that the test procedures perform well at least when the propensity score is estimated parametrically.

3.4 Balancing tests versus specification tests

The proposed balancing tests are conceptually easy to implement as they rely on regression with a single regressor, namely the propensity score estimate. In contrast, when using parametric specification tests comparing two models against each other, it is a priori not clear which models should be considered at all. The reset test proposed by Ramsey (1969) belongs to this class of tests and is based on the estimation of both models to be compared. One can theoretically choose from an infinite number of alternative specifications and the tests are only powerful if the alternative model is correct, or at least more accurate than the model assumed to be true under the null.

This concern does, however, not apply to all specification tests. Lagrange multiplier (LM) (see Breusch and Pagan, 1980, for implementations of LM specification tests in econometrics), do not require the estimation of the alternative model, some of them (e.g., the JarqueBera test) not even its specification. Also omnibus tests like the information matrix test (see White, 1982) avoid the issue of choosing an alternative model to be tested against the initial propensity score specification. The same applies to the test of Shaikh, Simonsen, Vytlacil, and Yildiz (2009) who provide an alternative to parametric specification tests. They show that if the propensity score is correctly specified, it holds that

$$f_{p(X)|D=1}(\rho|D=1) = \frac{\Pr(D=1)}{\Pr(D=0)} \frac{\rho}{1-\rho} f_{p(X)|D=0}(\rho|D=0) \quad \forall\ \rho \in (0,1), \tag{16}$$

where $f_{p(X)|D=d}(\cdot|D=d)$ denotes the pdf of $p(X)$ conditional on $D=d$. The authors provide a test based on kernel density estimation of the propensity score that is asymptotically normally distributed under null hypothesis:

$$T_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{1}{h} K\left(\frac{\hat{p}(X_i) - \hat{p}(X_j)}{h}\right) \hat{\epsilon}_i \hat{\epsilon}_j, \tag{17}$$

with K denoting the kernel function, h the bandwidth, and $\hat{\epsilon_i} \equiv D_i - \hat{p}(X_i)$.

However, one concern applying to any specification test is that we are merely interested in the validity of the balancing property, and not in the fit of $p(X)$. Thus, relying on specification tests rather than balancing tests may be overly restrictive. An incorrect, but easy to estimate (parametric) specification might be similarly accurate with respect to covariate balance as the true model, see Zhao (2008). Yet, it would most likely be rejected by a (powerful) specification test. Section 5 presents simulation results for a scenario where specification tests unnecessarily reject a misspecified, but balancing score, whereas the balancing tests do not.

4 After matching tests

Full sample tests verify whether the balancing property can be rejected for the population the entire sample is drawn from. After matching tests assess the balance in the matched sample, i.e., whether the matched comparisons are appropriate counterfactuals. Unlike in full sample tests, we need not condition on the propensity score, as this is done by the matching algorithm prior to testing. Poor balance in matched samples is either due to the propensity score's lacking balancing property, or to the matching algorithm's failure to establish common support in the propensity score, i.e., the occurrence of bad matches, or both. Thus, after matching tests are sensitive to the common support restrictions imposed in the PSM procedure, while full sample tests are not. Therefore, both full sample and after matching balancing tests appear to be useful tools that complement each other. Although the researcher is ultimately interested in the balance of the matched sample, both types of tests may be applied to trace the reasons for imbalance in order to take the right measures. A rejection by full sample tests indicates that the propensity score specification fails to balance the distributions of covariates which will most likely lead to imbalances in the matched sample. This suggests the modification of the propensity score specification. In contrast, imbalance attested by after matching tests when using a propensity score specification that has passed the full sample tests suggests the application of a different matching algorithm.

Full sample and after matching tests also differ with respect to the interpretation of the test statistics. Firstly, after matching tests are asymptotically not valid for testing the balancing property with respect to the population, because the matched sample is a nonrandom draw that depends on the matching algorithm. Therefore, judgements about balance strictly refer to the matched sample at hand. Secondly, predefined significance levels at which the null hypothesis has to be rejected are irrelevant stopping rules for covariate balancing in matched samples. After all, the researcher seeks to maximize balance without limit and would ideally obtain the same distributions of covariates in the pools of treated and nontreated matches. Therefore, Imai, King, and Stuart (2006) and Sekhon (2007a) criticize the use of hypothesis tests along with predefined levels of significance as stopping rules which nevertheless frequently appears in empirical work.

4.1 Balancing tests suggested in the literature

Several after matching tests have been proposed and applied in the literature. The most popular method among practitioners appears to be the two sample t-test which simply tests for mean differences in a particular covariate across treated and nontreated units. As for the DW test, Lee (2006) suggests to use permutation to improve the finite sample properties of the t-test. Conceptually similar to the t-test, the hotelling test checks for joint equality in the means of all elements in X in the matched sample. Diamond and Sekhon (2006) recommend to use (in addition to t-tests) univariate permuted KS tests which test for equality in distributions. The KS distribution test is the only after matching test which accounts for imbalances in higher moments just as the quantile based KS and CMS tests suggested below.

Imai, King, and Stuart (2006) argue that standard hypothesis tests as the t-test are inaccurate for assessing balance in matched samples, because the test statistic can be distorted by randomly dropping observations, even though the balance is unaffected. This suggests the use of methods that are robust to variations in the sample size, such as quantile-quantile plots of the covariates after matching. Rosenbaum and Rubin (1985) propose a test of standardized differences (see also

Imbens and Wooldridge, 2009, for a different version of the test) which is not affected by the sample size, but merely checks for differences in means. It is based on the mean differences of the covariate across treated and nontreated matches, scaled by the square root of the variances of the covariate in the full sample. Rosenbaum and Rubin consider a standardized difference greater than 20 as ‘large’, i.e., pointing to imbalances. In the Monte Carlo simulations and the application we will consider several hypothesis tests as well as the test of standardized differences.

4.2 Testing

It is straightforward to implement the KS and CMS resampling procedures as after matching tests to check for imbalances in the entire distributions of matched treated and nontreated units. As mentioned before, we need not condition on the propensity score any more as this task is performed by a (hopefully accurate) matching algorithm prior to testing. The after matching test consists of three steps: The estimation of the unconditional quantiles in the pools of treated and nontreated matched units, the computation of the test statistics, and the resampling procedure to compute p-values.

Let $\hat{Q}^{\tau}_{X^m_k}(d)$ the τth unconditional quantile in the sample of matched units with $D = d$, which is now estimated without regressing on the propensity score. Analogous to the full sample test, the KS and CMS statistics for the empirical inference process $\hat{Q}^{\tau}_{X^m_k}(1) - \hat{Q}^{\tau}_{X^m_k}(0)$ are

$$\begin{aligned} T^{KS}_{n^m} &= \sup_{x\in\mathcal{X}^m} \sqrt{\frac{n^m_1 \cdot n^m_0}{n^m}} ||\hat{Q}^{\tau}_{X^m_k}(1) - \hat{Q}^{\tau}_{X^m_k}(0)||_{\hat{\Lambda}}, \qquad (18) \\ T^{CMS}_{n^m} &= \frac{n^m_1 \cdot n^m_0}{n^m} \int_{\mathcal{X}^m} ||\hat{Q}^{\tau}_{X^m_k}(1) - \hat{Q}^{\tau}_{X^m_k}(0)||^2_{\hat{\Lambda}} dx. \qquad (19) \end{aligned}$$

We draw J bootstrap samples from the matched sample, estimate the quantiles $\hat{Q}^{\tau}_{X^m_k,j}(1), \hat{Q}^{\tau}_{X^m_k,j}(0)$ an compute the statistics on the bootstrapped and recentered inference processes:

$$\begin{aligned} T^{KS}_{n^m,j} &= \sup_{x\in\mathcal{X}^m} \sqrt{\frac{n^m_1 \cdot n^m_0}{n^m}} ||\hat{Q}^{\tau}_{X^m_k,j}(1) - \hat{Q}^{\tau}_{X^m_k,j}(0) - (\hat{Q}^{\tau}_{X^m_k}(1) - \hat{Q}^{\tau}_{X^m_k}(0))||_{\hat{\Lambda}}, \qquad (20) \\ T^{CMS}_{n^m,j} &= \frac{n^m_1 \cdot n^m_0}{n^m} \int_{\mathcal{X}_m} ||\hat{Q}^{\tau}_{X^m_k,j}(1) - \hat{Q}^{\tau}_{X^m_k,j}(0) - (\hat{Q}^{\tau}_{X^m_k}(1) - \hat{Q}^{\tau}_{X^m_k}(0))||^2_{\hat{\Lambda}} dx. \qquad (21) \end{aligned}$$

P-values are obtained by $1/J \sum_{j=1}^{J} I\{T_{n^m,j} > T_{n^m}\}$. The unconditional quantiles are estimated at parametric rates and are, in contrast to the full sample tests, not regressed on any estimated parameter.

As mentioned above, the p-values do not bear the same interpretation as in classic hypothesis tests (e.g., when testing the balancing property using the full sample tests). They are not to be used as stopping rules, but should be maximized without limit in order to maximize covariate balance in the matched sample. Note that the after matching tests might also be applied as *before matching* tests to investigate balance in the full sample (without conditioning on the propensity score). Differences in before and after matching p-values indicate the balance gains due to propensity score matching. Furthermore, the tests can be used to assess covariate balance in randomized experiments.

In addition to conventional bootstrapping, one may estimate the distribution of the test statistics by permutation, i.e., by randomly shuffling treatment and control labels among matched observations without replacement. Permutation tests are valid when shuffling the labels does not affect the results under the null hypothesis, see Good (2001). This condition is satisfied in the context of balancing tests where the distribution of the covariate is independent of the treatment label under null. Lee (2006) applies permutation to the t-test and Diamond and Sekhon (2006) to the KS distribution test. The latter was originally proposed by Abadie (2002) who uses the permuted KS test in a different context, namely to test for distributional treatment effects in an IV framework. Abadie shows that the procedure has correct asymptotic size under the weak condition that the variable (in our case X_k^m) has a nondegenerate distribution with bounded support. In the Monte Carlo simulations both resampling- and permutation-based versions of the tests are investigated.

As a final remark, it is important to note that the proposed methods cannot be easily applied to matching algorithms that do not create an explicit matched sample. E.g., kernel matching as discussed in Heckman, Ichimura, and Todd (1997, 1998) merely provides weights which balance the propensity scores of treated and nontreated units and allow predicting the counterfactual outcomes. These weights do not reveal the value of the counterfactual covariate, as there is no one-to-one correspondence between the propensity score and the covariate. The same value of the propensity score can in principle be obtained by many combinations of the covariates. For mean difference tests, it suffices that the weights allow estimating the conditional mean of the counterfactual covariate given the propensity score, as the tests average over the covariates in the 'matched' sample anyway. This is neither the case for the proposed CMS and KS procedures, nor for the KS distribution test, which require the knowledge of the distributions of the covariates in the matched sample. With this respect, full sample tests appear to be more generally applicable than the after matching tests considered in this section.

5 Monte Carlo simulations

In this section, we present Monte Carlo evidence on the finite sample properties of KS and CMS full sample and after matching tests and run a horse race with other tests proposed in the literature. Concerning the propensity score model, we consider three different scenarios: Correct specification of the propensity score, misspecification but satisfaction of the balancing property, and misspecification and violation of the balancing property. The motivation for these particular scenarios is that we would expect different classes of tests to behave differently. Accurate balancing tests should keep the null in the first and second scenario and reject it in the third, whereas specification tests should only keep the null in the first scenario. The scenarios give an intuition about the strengths and weaknesses of different (classes of) tests, but of course, they do not claim completeness, as many more data generating processes could be considered.

5.1 Full sample tests

Starting with the full sample tests, we compare the performance of our procedures to the DW test[4] (see Dehejia and Wahba 1999, 2002) with and without Bonferroni adjustment, the regression test of Smith and Todd (2005), a specification test related to Shaikh, Simonsen, Vytlacil, and Yildiz (2009), and the Ramsey (1969) reset test. We first consider the results for a correctly specified (and thus, balancing) propensity score model. The data generating process (DGP) is

$$\begin{aligned}
D_i &= I\{\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon > 0\},\\
Y_i &= \gamma_1 X_{1,i}^2 + \gamma_2 X_{2,i} + \gamma_3 D_i + U_i\\
X_1, X_2 &\sim \text{unif}(0,3), \quad \varepsilon \sim N(0,2), \quad U \sim N(0,1)\\
\beta_0 &= -1.5, \quad \beta_1 = \beta_2 = 0.5, \quad \gamma_1 = \gamma_2 = \gamma_3 = 1.
\end{aligned}$$

Treatment effects are homogenous w.r.t. X and equal to 1. The constant in the treatment equation (β_0) is chosen such that the unconditional probability to receive the treatment is about

[4] We test for equality in mean propensity scores among treated and nontreated units within a stratum at the 10% level of significance.

50 %, and the same applies to the other scenarios considered further below. The propensity score is correctly specified as a probit model,

$$p(X) = \Pr(D = 1|X) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2),$$

where $\Phi(\cdot)$ denotes the normal cdf.

We test whether the continuous covariate X_1 is balanced conditional on the propensity score for two sample sizes, $n = 1000, 4000$. These sample sizes are comparable to the data analyzed in several recent empirical studies using PSM estimators, e.g., Berger and Hill (2005), Blundell, Dias, Meghirs, and Reenen (2004), Jalan and Ravallion (2003), and Loecker (2007). Table 1 reports the rejection frequencies of the null hypothesis at the 5% and 10% significance levels, i.e., the share of p-values that lie at or below 0.05 and 0.10, respectively, for 1000 Monte Carlo replications. The propensity scores are estimated using the correct probit model. Inference for the KS and CMS balancing tests is based on 499 bootstrap draws. The conditional quantiles are evaluated at $\tau \in \mathcal{T}_{[0.25,0.75]} = \{0.25, 0.30, 0.35, ..., 0.75\}$. The propensity score $p(X)$ is evaluated on an equidistant grid consisting of 10 values between the 0.25th and 0.75th quantile of the estimated propensity score, which ensures that boundary regions with sparse data are not used in the test procedures.

We consider different combinations of smoothing and weighting schemes Λ for the KS and CMS balancing tests: We weight differences in conditional quantiles (i) by the inverse of their respective variance (CMS balancing (var), KS balancing (var)), which gives more weight to differences that are precisely estimated, and (ii) by the densities of the predicted propensity scores (CMS resampling (dens), CMS resampling (dens)), which gives more weight to differences in areas with large densities of the propensity score. Furthermore, smoothing is varied by using only the propensity score or 2nd and 3rd order polynomials of the propensity score in the quantile regressions, respectively.

Table 1 also reports the rejection frequencies of the regression test of Smith and Todd (2005), henceforth ST test, and the DW test. Following Lee (2006), whose simulations suggest that the DW test has very poor size properties and rejects the null much too often, we also consider a modified DW test with an approximation of the Bonferroni adjustment (‘DW Bonferroni adj.’). Testing for balance with respect to X_1, the Bonferroni adjustment implies that the significance level (i.e., 5 or 10%) is divided by the number of intervals such that the chance of rejection for each t-test in a particular interval is adjusted downwards to keep the overall probability of incorrect rejection constant as the number of intervals increases.

As discussed in Section 2, Shaikh, Simonsen, Vytlacil, and Yildiz (2009) show that a correctly specified propensity score model implies condition (16) and propose a test based on kernel density estimation of the propensity score. We define the smoothing parameter h to be the optimal bandwidth according to the maximum likelihood (ML) cross validation criterion for kernel density estimation, see Hayfield and Racine (2008). We have also tried versions of the test based on under- and oversmoothing (using half of and twice the optimal bandwidth, respectively), but as the results are not too different to optimal smoothing they are omitted in the tables. Finally, we run a Ramsey (1969) reset test where the correct model is tested against the alternative

$$p(X) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \beta_4 X_2 + \beta_5 X_2^2 + \beta_6 X_2^3).$$

As expected, all tests correctly keep the null in most Monte Carlo replications. The CMS test is very conservative and rejects the balancing hypothesis substantially less frequently than the theoretical rates of 5 and 10% and even more so when using propensity score density weighting. However, when using a second order polynomial of the propensity score, the empirical size of the test improves as the sample size increases. The rejection frequencies of the KS test are generally closer to the theoretical size, again in particular when using a 2nd order polynomial. Note that the rejection rates of either test are non-monotone in the order of the propensity score. The Shaikh et al. specification test, the DW test with Bonferroni adjustment, and the ST test are conservative for both sample sizes whereas the standard DW test rejects the null somewhat too often for $n = 4000$. Empirical sizes of the Ramsey reset test are among the most accurate. With the exception of the DW test without Bonferroni adjustment, the empirical size of which deteriorates in the sample size, no class of tests seems to do strikingly better or worse than any other.

To check the accuracy of propensity score methods under the correct specification we apply two nearest neighbors caliper matching and inverse probability weighting (IPW) estimators to the simulated data. For matching we use the Match command by Sekhon (2007b) and set the caliper to 0.1 standard deviations of the propensity score. The ATE estimate is $\hat{\Delta} = 1.004$ for $n = 1000$ and the mean squared error (MSE) is 0.008. For $n = 4000$, $\hat{\Delta} = 1.002$ and MSE$= 0.002$. The IPW estimator, see for instance Horvitz and Thompson (1952) and Hirano, Imbens, and Ridder (2003), performs similarly well. $\hat{\Delta} = 0.998, 1.002$ and MSE$= 0.007, 0.002$ for $n = 1000, 4000$.

We now turn to a more interesting scenario where the propensity score is misspecified, but yet

Table 1: Full sample tests: Rejection frequencies under correct specification

	n=1000		n=4000	
rejection rates at	5%	10%	5%	10%
CMS balancing (var) order 1*	0.000	0.000	0.000	0.000
CMS balancing (var) order 2*	0.010	0.037	0.021	0.084
CMS balancing (var) order 3*	0.000	0.001	0.000	0.000
CMS balancing (dens) order 1*	0.000	0.002	0.000	0.000
CMS balancing (dens) order 2*	0.009	0.032	0.020	0.078
CMS balancing (dens) order 3*	0.000	0.000	0.000	0.000
KS balancing (var) order 1*	0.011	0.034	0.033	0.061
KS balancing (var) order 2*	0.033	0.076	0.044	0.109
KS balancing (var) order 3*	0.009	0.025	0.013	0.036
KS balancing (dens) order 1*	0.008	0.021	0.006	0.025
KS balancing (dens) order 2*	0.037	0.090	0.064	0.132
KS balancing (dens) order 3*	0.004	0.019	0.012	0.031
DW	0.044	0.047	0.134	0.155
DW Bonferroni adj.	0.007	0.010	0.018	0.032
Smith and Todd	0.015	0.045	0.027	0.067
Shaikh et al. spec. test**	0.006	0.009	0.006	0.011
Ramsey reset	0.037	0.079	0.043	0.087

Note: 1000 Monte Carlo replications.

*: 499 bootstrap draws per replication.

**: bandwidth according to ML cross validation.

balancing. We investigate the performance of the tests when data are drawn from the following DGP:

$$\begin{aligned}
D_i &= I\{\beta_0 + \beta_1 X_{1,i}^3 + \beta_2 X_{2,i} + \varepsilon > 0\}, \\
Y_i &= \gamma_1 X_{1,i}^2 + \gamma_2 X_{2,i} + \gamma_3 D_i + U_i \\
X_1, X_2 &\sim \text{unif}(0,3), \quad \varepsilon \sim N(0,5), \quad U \sim N(0,1) \\
\beta_0 &= -3, \quad \beta_1 = 0.3, \quad \beta_2 = 0.5, \quad \gamma_1 = \gamma_2 = \gamma_3 = 1.
\end{aligned}$$

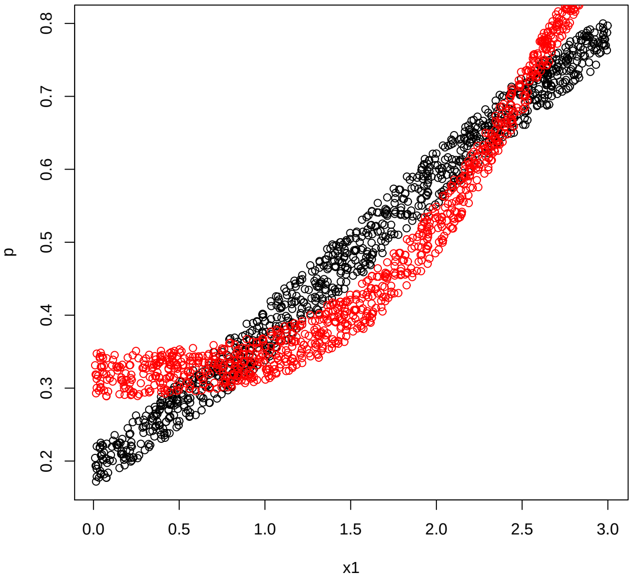
We incorrectly use the same propensity score model as before, $p(X) = \Pr(D = 1|X) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$, such that β_1 is estimated with respect to X_1 instead of X_1^3. Thus, it is assumed that the index model that underlies the treatment probability is linear in X_1, whereas the true relationship is cubic. Yet, the incorrect model satisfies the balancing property for variable X_1, as it is only a monotonous transformation of the true model such that the order of the propensity scores is preserved under misspecification. Even though the propensity scores themselves are poorly estimated, the treated are matched to nontreated units with similar $p^*(X)$ when using propensity score matching.

To gain some intuition, Figure 1 displays 1000 simulated values of X_1 along with propensity score estimates (i) using the misspecified probit model (dark bubbles) and (ii) based on the correct specification $p^*(X)$ (light bubbles). As the rank of each observation on average remains the same in either case such that observations with similar $p^*(X)$ are matched even when using the wrong specification, estimation is consistent.[5]

Table 2 reports the rejection frequencies under the misspecified, but balancing scenario where the propensity score is estimated based on the misspecified probit model. All versions of the CMS test are either on the conservative side or have rejection frequencies that are not too far from the theoretical sizes. Note that there seems to be no clear relationship between the empirical size and the order or the weighting scheme. Also the results for the KS test are quite satisfactory, with the exception of the test versions using a third order polynomial under the larger sample size which rejects the null too often. The standard DW test is quite accurate for $n = 1000$, but its performance deteriorates in the sample size. The Bonferroni adjustment considerably improves the size properties of the DW test for $n = 4000$. The rejection frequencies of the ST test are already too high for $n = 1000$ and severely increase in the sample size. This is

[5] It is, however, less efficient than estimation based on the true propensity score model.

Figure 1: Misspecified and balancing scenario



Propensity scores under misspecification (dark bubbles) and correct specification (light).

somewhat surprising, as the ST procedure should theoretically test for covariate balance, not for misspecification. Still, it seems to have power into the wrong direction. As expected, the rejection rates of the Shaikh et al. specification test increase in the sample size. It rejects the misspecified, but balancing model in all replications for $n = 4000$. The reset test is most 'powerful' in inappropriately rejecting the null. Its rejection frequencies are above 90% even for the smaller sample size. We conclude that only the CMS and KS procedures as well as the DW test with Bonferroni adjustment yield satisfactory results under the misspecified, but balancing scenario.

Again, we investigate the finite sample properties of two nearest neighbors caliper matching on the propensity score. $\hat{\Delta} = 1.033$ for $n = 1000$ and the MSE is equal to 0.008. For $n = 4000$, $\hat{\Delta} = 1.031$ and MSE$= 0.003$. Similar to the results in Zhao (2008), the misspecification of the propensity score does not much affect PSM. This is, however, not true for IPW estimators, as consistency of this class of estimators is contingent on the correctness of the propensity score specification. Indeed, the IPW estimates are substantially biased ($\hat{\Delta} = 1.293, 1.295$) and the MSEs are large $(0.096, 0.090)$ for 1000 and 4000 observations, respectively. Therefore, PSM seems to be more robust to propensity score misspecification.

Thirdly, we consider a DGP under which the probit specification is misspecified and not

Table 2: Full sample tests: Rejection frequencies under misspecification and balance

	n=1000		n=4000	
rejection rates at	5%	10%	5%	10%
CMS balancing (var) order 1*	0.007	0.013	0.001	0.004
CMS balancing (var) order 2*	0.056	0.079	0.005	0.010
CMS balancing (var) order 3*	0.009	0.038	0.039	0.105
CMS balancing (dens) order 1*	0.008	0.015	0.000	0.003
CMS balancing (dens) order 2*	0.049	0.080	0.005	0.008
CMS balancing (dens) order 3*	0.011	0.030	0.037	0.107
KS balancing (var) order 1*	0.019	0.037	0.003	0.011
KS balancing (var) order 2*	0.073	0.123	0.063	0.103
KS balancing (var) order 3*	0.047	0.115	0.121	0.227
KS balancing (dens) order 1*	0.035	0.070	0.043	0.077
KS balancing (dens) order 2*	0.047	0.081	0.026	0.062
KS balancing (dens) order 3*	0.051	0.113	0.176	0.296
DW	0.074	0.082	0.265	0.301
DW Bonferroni adj.	0.021	0.030	0.047	0.063
Smith and Todd	0.182	0.274	0.747	0.850
Shaikh et al. spec. test**	0.508	0.588	1.000	1.000
Ramsey reset	0.915	0.951	1.000	1.000

Note: 1000 Monte Carlo replications.

*: 499 bootstrap draws per replication.

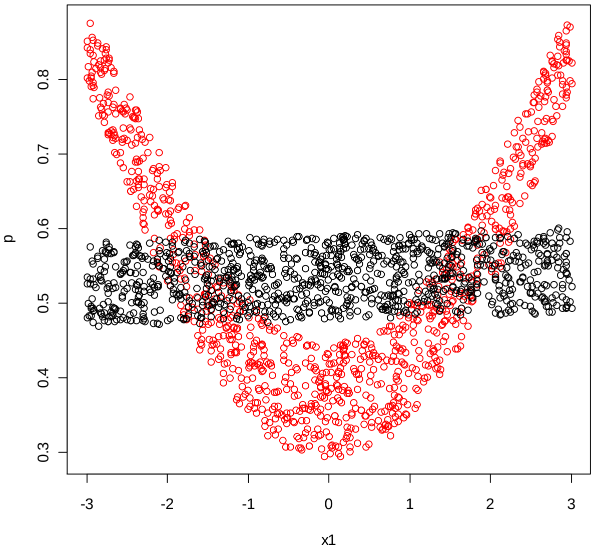
**: bandwidth according to ML cross validation.

balancing:

$$\begin{aligned} D_i &= I\{\beta_0 + \beta_1 X_{1,i}^2 + \beta_2 X_{2,i} + \varepsilon > 0\}, \\ Y_i &= \gamma_1 X_{1,i}^2 + \gamma_2 X_{2,i} + \gamma_3 D_i + U_i \\ X_1, X_2 &\sim \text{unif}(-3,3), \quad \varepsilon \sim N(0,5), \quad U \sim N(0,1) \\ \beta_0 &= -3, \quad \beta_1 = 1, \quad \beta_2 = 0.5, \quad \gamma_1 = \gamma_2 = \gamma_3 = 1. \end{aligned}$$

To clarify the issues of misspecification *and* imbalance, Figure 2 displays 1000 simulated realizations of X_1 along with propensity score estimates under misspecification (dark bubbles) and under the correct specification (light bubbles).

Figure 2: Misspecified and non-balancing scenario



Propensity scores under misspecification (dark bubbles) and correct specification (light).

Imbalance is due to the fact that observations with high absolute values in X_1 are more likely to be treated than those with values close to zero. Only treated and nontreated with the same or similar $p^*(X)$ should be compared to each other. It is obvious that matching on estimates of $p(X)$ fails to do. The reason is that the incorrect model $p(X)$ cannot handle the U-shaped non-monotonicity in the relation between X_1 and the true propensity score. $p^*(X)$ is minimized at the mean of X_1, which is zero, and increases in either direction. Due to this symmetric relationship, the expected value of the slope coefficient estimate β_1 is zero. Therefore, the expected values of the propensity score estimates are independent of X_1, implying that $E(X_1|D = d, p(X)) = E(X_1|D = d)$. Hence, matching is random with respect to the true propensity score such that observations with fairly different X_1 are incorrectly compared to each other.

Table 3 reports the results under the misspecified, non-balancing scenario. Already for $n = 1000$, the CMS and KS tests are quite powerful and even more so when using inverse variance weighting. In the latter case, the null is always rejected in more than 90 % of the simulations. For $n = 4000$, the rejection rates amount to 100 % for any test version, independent of the order and the weighting scheme. In contrast, the power of balancing test based on mean differences is low. Note that for the DGP considered, the expected value of X_1 is zero for the treated and for the nontreated. Hence, $E(X|D = d, p(X)) = E(X|D = d)$ and $E(X|D = 1) = E(X|D = 0) = 0$ together imply that conventional balancing tests have no power to reject the null. This explains the poor performance of the DW test (with and without Bonferroni adjustment) and the ST test. Interestingly, the specification tests considered do no better. For the reset test, a zero coefficient on X_1 is on average as likely as non-zero coefficients on X_1 and higher order terms and therefore, it has little power. But also the Shaik et al. test keeps the null most of the time. The only tests that have power in this particular scenario with misspecification and imbalance are the CMS and KS tests.

How is the PSM estimator affected by the imbalance? For $n = 1000$, the ATE estimate is severely biased ($\hat{\Delta} = 3.038$) and the MSE (4.191) is huge. For $n = 4000$, $\hat{\Delta} = 3.071$ and MSE= 4.297. The IPW estimator yields $\hat{\Delta} = 3.094, 3.097$ and MSE= $4.414, 4.404$ for $n = 1000, 4000$, respectively. Thus, the imbalance is not innocuous and entails severe biases and inconsistency. We conclude that the KS and CMS balance tests appear to be quite competitive when compared to existing specification and balancing tests. Their rejection frequencies are low when the balancing property holds and very high when it is violated, at least in the scenarios considered.

5.2 After matching tests

This section presents simulations on the finite sample properties of after matching tests and considers the same DGPs as for the full sample tests. We compare our CMS and KS tests based on resampling (in our case bootstrapping) and permutation to the permuted KS distribution test (Diamond and Sekhon, 2006), the permuted and conventional (i.e., relying on asymptotic theory)

Table 3: Full sample tests: Rejection frequencies under misspecification and imbalance

	n=1000		n=4000	
rejection rates at	5%	10%	5%	10%
CMS balancing (var) order 1*	0.930	0.953	1.000	1.000
CMS balancing (var) order 2*	0.975	0.991	1.000	1.000
CMS balancing (var) order 3*	0.997	1.000	1.000	1.000
CMS balancing (dens) order 1*	0.904	0.949	1.000	1.000
CMS balancing (dens) order 2*	0.896	0.970	1.000	1.000
CMS balancing (dens) order 3*	0.955	0.992	1.000	1.000
KS balancing (var) order 1*	0.996	0.997	1.000	1.000
KS balancing (var) order 2*	0.999	0.999	1.000	1.000
KS balancing (var) order 3*	1.000	1.000	1.000	1.000
KS balancing (dens) order 1*	0.982	0.995	1.000	1.000
KS balancing (dens) order 2*	0.997	0.998	1.000	1.000
KS balancing (dens) order 3*	0.997	1.000	1.000	1.000
DW	0.062	0.070	0.173	0.184
DW Bonferroni adj.	0.009	0.013	0.033	0.046
Smith and Todd	0.037	0.088	0.137	0.219
Shaikh et al. spec. test**	0.010	0.013	0.001	0.001
Ramsey reset	0.138	0.202	0.085	0.149

Note: 1000 Monte Carlo replications.

*: 499 bootstrap draws per replication.

**: bandwidth according to ML cross validation.

two sample t-tests, and the test of standardized differences of Rosenbaum and Rubin (1985).

For the CMS and KS resampling tests, we again consider two different weighting schemes Λ: We weight differences in quantiles (i) by the inverse of their respective variances (CMS resampling (var), KS resampling (var)) and (ii) by the densities of the predicted propensity scores (CMS resampling (dens), CMS resampling (dens)). To be precise, we weight the differences in quantiles by the product of the densities at the respective quantiles in the samples of treated and nontreated matches. The quantiles are evaluated at $\tau \in \mathcal{T}_{[0.1,0.9]} = \{0.10, 0.11, 0.12, ..., 0.90\}$ and inference relies on 499 bootstrap draws or permutations, respectively.

Table 4 displays the results for the correctly specified (and balancing) scenario. Even though the balancing property holds, the rejection frequencies of the CMS and KS tests, including the KS distribution test, are much higher than the theoretical sizes and increase with the sample size. The tests seem to detect the slightest imbalances not eliminated by the matching algorithm. This is unsatisfactory, as the caliper matching procedure yields estimates which are close to the true value even without perfect balance. Note that the empirical sizes of the CMS and KS resampling tests are more accurate when weighting by the propensity score densities, but are still far from being acceptable. The KS distribution test used by Diamond and Sekhon (2006) performs even worse. In contrast, the rejection frequencies of permuted and standard t-tests are not too far from the theoretical sizes, whereas the test of standardized differences is very conservative.

In the misspecified but balancing scenario (see Table 5), the CMS and KS resampling tests with propensity score density weighting have accurate sizes for $n = 1000$, but reject the null much too often for $n = 4000$. Again, they perform better than the CMS and KS tests based on inverse variance weighting. Also the KS distribution test rejects the null much too often whereas the t-tests and the test of standardized differences are overly conservative for both sample sizes.

Under misspecification and imbalance all CMS and KS procedures are very powerful and reject the null all the time, see Table (6). In contrast, mean difference tests fail to detect the imbalance related to higher moments. The rejection frequencies of the t-tests are fairly low and the test of standardized differences has no power at all. Summing up, simulation evidence on after matching tests is ambiguous about the relative performance of the proposed tests. Even though the CMS and KS tests are very powerful under imbalance, they reject the null much too often when the balancing property holds. This suggests that we should have more confidence in the CMS and KS full sample tests than in the after matching versions. Using the density of the

Table 4: After matching tests: Rejection frequencies under correct specification

	n=1000		n=4000	
rejection rates at	5%	10%	5%	10%
CMS resampling (var)*	0.239	0.371	0.624	0.783
CMS resampling (dens)*	0.150	0.274	0.560	0.721
KS resampling (var)*	0.366	0.490	0.788	0.871
KS resampling (dens)*	0.168	0.271	0.616	0.752
CMS permutation*	0.218	0.355	0.622	0.772
KS permutation*	0.385	0.531	0.807	0.885
KS distribution*	0.695	0.828	0.989	0.998
permuted t-test*	0.015	0.042	0.068	0.119
standard t-test	0.010	0.024	0.066	0.118
test of standardized differences**	0.000		0.000	

Note: 1000 Monte Carlo replications,

*: 499 bootstrap draws/permutations per replication.

**: rejection if absolute standardized difference > 20.

Table 5: After matching tests: Rejection frequencies under misspecification and balance

	n=1000		n=4000	
rejection rates at	5%	10%	5%	10%
CMS resampling (var)*	0.082	0.160	0.410	0.573
CMS resampling (dens)*	0.047	0.097	0.335	0.481
KS resampling (var)*	0.160	0.234	0.603	0.722
KS resampling (dens)*	0.054	0.111	0.446	0.588
CMS permutation*	0.093	0.158	0.411	0.579
KS permutation*	0.184	0.251	0.653	0.749
KS distribution*	0.418	0.561	0.940	0.973
permuted t-test	0.000	0.000	0.000	0.001
standard t-test	0.000	0.000	0.000	0.000
test of standardized differences**	0.000		0.000	

Note: 1000 Monte Carlo replications,

*: 499 bootstrap draws/permutations per replication.

**: rejection if absolute standardized difference > 20.

propensity score estimates as weights in the after matching tests partly alleviates the problem of over-rejection. Therefore, more research is required with regard to the optimal choice of the weighting matrix in balancing tests.

Table 6: After matching tests: Rejection frequencies under misspec. and imbalance

	n=1000		n=4000	
rejection rates at	5%	10%	5%	10%
CMS resampling (var)*	1.000	1.000	1.000	1.000
CMS resampling (dens)*	1.000	1.000	1.000	1.000
KS resampling (var)*	1.000	1.000	1.000	1.000
KS resampling (dens)*	1.000	1.000	1.000	1.000
CMS permutation*	1.000	1.000	1.000	1.000
KS permutation*	1.000	1.000	1.000	1.000
KS distribution*	1.000	1.000	1.000	1.000
permuted t-test	0.050	0.110	0.115	0.167
standard t-test	0.064	0.106	0.144	0.213
test of standardized differences**	0.000		0.000	

Note: 1000 Monte Carlo replications,

*: 499 bootstrap draws/permutations per replication.

**: rejection if absolute standardized difference > 20.

6 Empirical applications

In this section, we apply full sample and after matching tests to labor market data previously analyzed by Ichino, Mealli, and Nannicini (2008).

6.1 Full sample tests

Ichino, Mealli, and Nannicini (2008) use PSM to evaluate the effects of job placements by temporary work agencies (TWAs) on the probability to find permanent employment later on in the two Italian regions of Sicily and Tuscany. The data were collected by phone interviews. The treatment period (having or not having a temporary job by TWA assignment) covers the first semester of 2001, the outcome (permanent employment) was measured in November 2002. Pre-treatment covariates X include detailed information about demographic characteristics, educational attainment, family background and the recent employment history of treated and nontreated individuals. While Ichino, Mealli, and Nannicini (2008) are interested in the robustness of the estimated effects with respect to omitted unobserved factors that would violate the CIA, we use their data to investigate the balancing property of their propensity score specification, which is based on a probit model.

We restrict our attention to the sample drawn in Tuscany, which consists of 281 treated and 628 nontreated individuals. We test the balancing property of the propensity score specification used in Ichino, Mealli, and Nannicini (2008) for the variable 'fraction of the school-to-work period that the worker spent as unemployed' (in %), which characterizes the relative time spent in unemployment after finishing eduction. Before matching, the fraction is 37.9 % for the treated and 47.7 % for the nontreated individuals in the sample. We apply the CMS and KS full sample tests to the region of common support in the predicted propensity scores $\hat{p}(X_i)$. Therefore, observations in any treatment group with $\hat{p}(X_i)$ higher than the maximum and lower than the minimum in the other treatment group are discarded from the sample. This leaves us with 255 treated and 519 nontreated individuals. We test the null hypothesis at ranks $\tau \in \mathcal{T}_{[0.25,0.75]} = \{0.25, 0.25, 0.30, ..., 0.75\}$ and $p(x) \in \mathcal{P}_{[0.20,0.80]} = \{0.20, 0.25, 0.30, ..., 0.80\}$ using 999 bootstrap replications.

Table 7 presents the test results. All CMS and KS balancing tests keep the null at the 5% level, irrespective of the order of the propensity score and the weighting scheme. Ichino, Mealli, and Nannicini (2008) use the DW test algorithm for Stata provided by Becker and Ichino (2002) and do not reject the balancing property either. Note, however, that the significance level chosen by the authors is 0.1 %. Setting the significance level to just 1% would reject the null, but one has to bear in mind that this result comes without the Bonferroni adjustment. This example highlights the arbitrariness of the standard DW test with respect to the significance level to be chosen when there are many propensity score intervals. The ST test, which uses a quartic of the propensity score in the regression, rejects the null at the 1% level. However, the test is very sensitive to the choice of the order. Versions based on squared and cubic expansions of the propensity score yield p-values larger than 5 %. Whereas the CMS and KS balancing tests unanimously keep the null under various propensity score polynomials, the conclusions drawn from the ST and DW tests depend on the choice of the functional form and the level of significance that is considered to be appropriate in the light of stratification, respectively.

Table 7: Application of full sample tests

	'fraction unemployed' p-value
CMS balancing (var) order 1*	0.615
CMS balancing (var) order 2*	0.724
CMS balancing (var) order 3*	0.562
CMS balancing (dens) order 1*	0.590
CMS balancing (dens) order 2*	0.638
CMS balancing (dens) order 3*	0.382
KS balancing (var) order 1*	0.624
KS balancing (var) order 2*	0.768
KS balancing (var) order 3*	0.644
KS balancing (dens) order 1*	0.676
KS balancing (dens) order 2*	0.630
KS balancing (dens) order 3*	0.704
DW**	0.009
Smith and Todd	0.003
Shaikh et al. spec. test+	0.189

Note: *: 999 bootstrap draws.

**: minimum p-value of all intervals.

+: bandwidth according to ML cross validation.

6.2 After matching tests

We apply the CMS and KS after matching tests based on resampling and permutation, the permuted KS distribution test, the standard and permuted t-tests, and the test of standardized differences to the same variable after the application the two nearest neighbors caliper matching algorithm.[6] Figure 3 presents the distributions of the variables 'fraction of the school-to-work period that the worker spent as unemployed' for treated and nontreated matches. The distributions appear to be similar and also the sample means are quite close, namely 43.072 % for the treated and 43.626 % for the nontreated individuals.

Figure 3: Fraction in unemployment (in %) of treated (left) and nontreated (right) matches

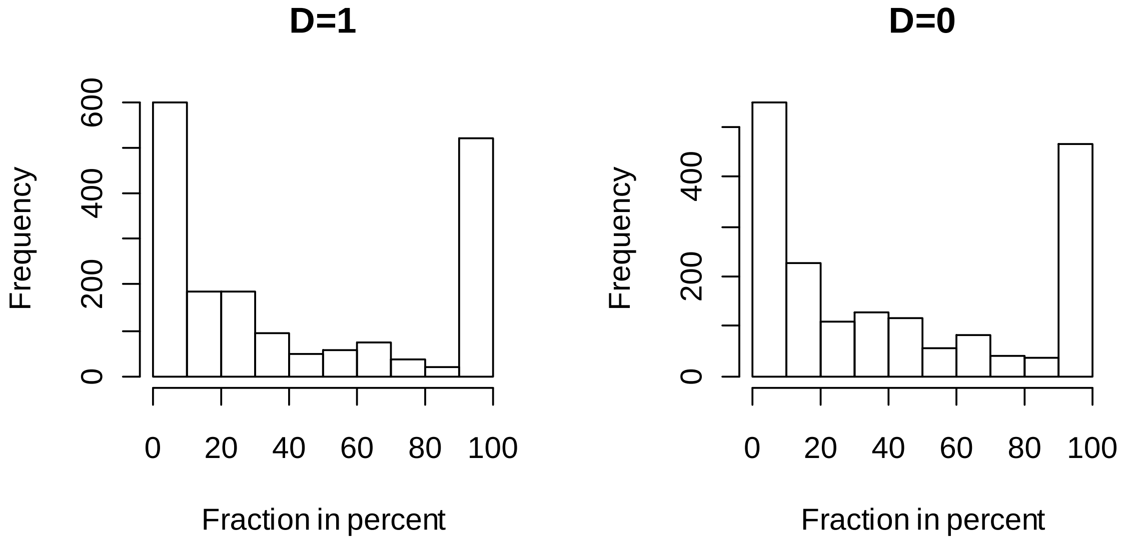


Table 8 reports the results of the CMS and KS tests, which evaluate the quantiles at ranks $\tau \in \mathcal{T}_{[0.1,0.9]} = \{0.1, 0.11, 0.12, ..., 0.9\}$ and are based on 999 bootstrap samples. Most of the CMS and KS tests yield p-values larger than 5% for balance of the variable 'fraction of period in unemployment', which is in line with the CMS and KS full sample tests. Only the KS distribution test is highly significant, whereas the t-tests and the test of standardized difference suggest that the variable is well balanced. Summing up, both the full sample and after matching versions of the quantile-based CMS and KS tests do not provide evidence that the balancing property fails

[6] The caliper is set to 0.1 standard deviations of the propensity score and 59 observations (6.5%) are dropped due to a lack of common support.

for the variable considered.

Table 8: Application of after matching tests

	'fraction unemployed' p-value
CMS resampling (var)*	0.270
CMS resampling (dens)*	0.303
KS resampling (var)*	0.348
KS resampling (dens)*	0.553
CMS permutation*	0.075
KS permutation*	0.046
KS distribution*	0.000
permuted t-test*	0.664
standard t-test	0.675
test of standardized differences**	-1.434

Note: *: 999 bootstrap draws/permutations per replication.

**: rejection if absolute standardized difference > 20.

7 Conclusion

The balancing property of the propensity score is key to the consistency of propensity score matching estimators. Thus, the attractiveness of this class of estimators over parametric alternatives with respect to model flexibility is lost when using a propensity score specification that is incapable of balancing the distributions of the covariates in the groups of treated and nontreated units.

In this paper, we propose balancing tests for continuous covariates based on quantile regression and bootstrapping Kolmogorov-Smirnov and Cramer-von-Mises-Smirnov statistics. These tests account for differences in the entire distributions of the covariates. If distributional differences affect higher moments, they are likely to be more powerful than conventional balancing tests as the DW test used in Dehejia and Wahba (1999, 2002), the regression test by Smith and Todd (2005), and the two sample t-test for matched samples, which merely check for differences in means. In contrast to specification tests such as suggested by Shaikh, Simonsen, Vytlacil, and Yildiz (2009), the proposed methods do not reject misspecified, but yet balancing propensity scores. This is beneficial from a practitioner's point of view who might prefer a misspecified, but easy to estimate model over the unknown true model for the sake of econometric feasibility. As long as the incorrect specification balances, propensity score matching is consistent, such that specification tests seem overly restrictive.

The proposed tests can either be applied in full or in matched samples. Implemented as full sample tests, they test balancing conditional on the propensity score. Similar to the DW test, a rejection of the null implies the use of a different, typically more flexible propensity score specification. Monte Carlo results suggest that the power and size properties are satisfactory in scenarios where conventional balancing tests fail to detect imbalances and specification tests incorrectly reject a misspecified, but balancing propensity score model. Implemented as after matching tests, the tests apply to unconditional quantiles in the pools of treated and nontreated units, as the matching algorithm should eliminate differences in the common support of the propensity score prior to testing. The proposed tests are very powerful when the matched sample is not balanced, but reject the null too often when the balancing property actually holds. This suggests that we should have more confidence in the CMS and KS full sample tests than in the after matching versions.

There are several caveats of the CMS and KS tests that future research might want to address. In contrast to specification tests and other balancing tests, the methods apply to continuous covariates alone and are not suitable to assess the balance of discrete variables with few mass points Also for this reason, they are not useful to build joint tests on all covariates. Finally, it remains to be clarified which are the optimal weighting functions to be used in the test procedures.

References

Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97, 284–292.

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

Berger, L. M., and J. Hill (2005): "Maternity leave, early maternal employment and child health and development in the US," *The Economic Journal*, 115, F29–F47.

Blundell, R., M. C. Dias, C. Meghirs, and J. V. Reenen (2004): "Evaluating the Employment Impact of a Mandatory Job Search Program," *Journal of the European Economic Association*, 2, 569–606.

Böckerman, P., and P. Ilmakunnas (2009): "Unemployment and self-assessed health: Evidence from panel data," *Journal of Health Economics*, 18, 161–179.

Breusch, T. S., and A. R. Pagan (1980): "The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics," *The Review of Economic Studies*, 47, 239–253.

Chernozhukov, V., and I. Fernandez-Val (2005): "Subsampling inference on quantile regression processes," *Sankhya: The Indian Journal of Statistics*, 67, 253–276.

Dehejia, R. H., and S. Wahba (1999): "Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programmes," *Journal of American Statistical Association*, 94, 1053–1062.

——— (2002): "Propensity-score-matching methods for nonexperimental causal studies," *The Review of Economics and Statistics*, 84, 151–161.

Diamond, A., and J. S. Sekhon (2006): "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies," *Institute of Governmental Studies Working Paper*.

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

Hayfield, T., and J. Racine (2008): "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27, 1–32.

Heckman, J. 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, 605–654.

——— (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Horvitz, D. G., and D. J. Thompson (1952): “A Generalization of Sampling without Replacement from a Finite Universe,” *Journal of the American Statistical Association*, 47, 663–685.

Ichino, A., F. Mealli, and T. Nannicini (2008): “From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity?,” *Journal of Applied Econometrics*, 23, 305–327.

Imai, K., G. King, and E. Stuart (2006): “The Balance Test Fallacy in Matching Methods for Causal Inference,” *unpublished manuscript.*

Imbens, G. W. (2000): “The Role of the Propensity Score in Estimating Dose-Response Functions,” *Biometrika*, 87, 706–710.

——— (2004): “Nonparametric estimation of average treatment effects under exogeneity: a review,” *The Review of Economics and Statistics*, 86, 4–29.

Imbens, G. W., and J. M. Wooldridge (2009): “Recent Developments in the Econometrics of Program Evaluation,” *Journal of Economic Literature*, 47, 5–86.

Jalan, J., and M. Ravallion (2003): “Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching,” *Journal of Business and Economic Statistics*, 21, 19–30.

Koenker, R., and G. Bassett (1978): “Regression quantiles,” *Econometrica*, 46, 33–50.

Lechner, M. (2001): “Identification and estimation of causal effects of multiple treatments under the conditional independence assumption,” in *Econometric Evaluations of Active Labor Market Policies in Europe*, ed. by M. Lechner, and F. Pfeiffer. Heidelberg: Physica.

Lee, W. (2006): “Propensity Score Matching and Variations on the Balancing Test,” *unpublished manuscript.*

Li, Q., E. Maasoumic, and J. Racine (2009): “A nonparametric test for equality of distributions with mixed categorical and continuous data,” *Journal of Econometrics*, 148, 186–200.

Loecker, J. D. (2007): “Do exports generate higher productivity? Evidence from Slovenia,” *Journal of International Economics*, 73, 69–98.

Persson, T. (2001): “Currency Unions and Trade: How Large Is the Treatment Effect?,” *Economic Policy*, 16, 435–448.

Pitman, E. J. G. (1937): “Significance Tests Which May be Applied to Samples From any Populations,” *Supplement to the Journal of the Royal Statistical Society*, 4, 119–130.

Ramsey, J. B. (1969): “Tests for specification errors in classical linear least squares regression analysis,” *Journal of the Royal Statistical Society, series B*, 31, 350–371.

Rosenbaum, P. R., and D. B. Rubin (1983): “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, 70, 41–55.

——— (1984): “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score,” *Journal of the American Statistical Association*, 79, 516–524.

——— (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *The American Statistician*, 39, 33–38.

Rubin, D. B. (1997): "Estimating Causal Effects from Large Data Sets Using Propensity Scores," *Annals of Internal Medicine*, 127, 757–763.

Sekhon, J. S. (2007a): "Alternative Balance Metrics for Bias Reduction in Matching Methods for Causal Inference," *unpublished manuscript.*

——— (2007b): "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R," *forthcoming in the Journal of Statistical Software.*

Shaikh, A. M., M. Simonsen, E. J. Vytlacil, and N. Yildiz (2009): "A specification test for the propensity score using its distribution conditional on participation," *Journal of Econometrics*, 151, 33–46.

Smith, J., and P. Todd (2005): "Rejoinder," *Journal of Econometrics*, 125, 365–375.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.

Zhao, Z. (2008): "Sensitivity of propensity score methods to the specifications," *Economics Letters*, 98, 309–319.