On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies

Daniel L. MILLIMET
Department of Economics, Southern Methodist University, Dallas, TX 75275 (*millimet@mail.smu.edu*)

Rusty TCHERNIS
Department of Economics, Indiana University, Bloomington, IN 47405 (*rtcherni@indiana.edu*)

The use of propensity score methods for program evaluation with nonexperimental data typically requires that the propensity score be estimated, often with a model whose specification is unknown. Although theoretical results suggest that estimators using more flexible propensity score specifications perform better, this has not filtered into applied research. Here we provide Monte Carlo evidence indicating benefits of *overspecifying* the propensity score that are robust across a number of different covariate structures and estimators. We illustrate these results with two applications, one assessing the environmental effects of General Agreement on Tariffs and Trade/World Trade Organization membership and the other assessing the impact of adopting the euro on bilateral trade.

KEY WORDS: Currency union; Environment; Program evaluation; Treatment effects; World Trade Organization.

1. INTRODUCTION

Estimation methods using the propensity score (i.e., the probability of an observation receiving a particular treatment conditional on covariates) are widely used in economics and other disciplines in evaluating programs and interventions. But in most applications, the *true* propensity score is unknown, and thus it must be estimated. Even in cases where the true propensity score is known, Hahn (1998) and Hirano, Imbens, and Ridder (2003) have shown that using the true propensity score is inefficient. Nonetheless, little research to date has assessed the properties of estimators of the *treatment effect* under various propensity score estimation procedures. Although a few studies have provided theoretical guidance with respect to the number of terms to be used in series estimators (which we use to estimate the propensity score) or the asymptotic properties of treatment effect estimators (Newey 1994; Hirano, Imbens, and Ridder 2003; Abadie 2005), our focus is on the impact of propensity score specification on the finite-sample performance of *treatment effect* estimators. In this respect, few guidelines are available to applied researchers. Specifically, two interrelated specification issues arise when estimating the propensity score in practice: determining which variables to include in the propensity score model and determining the functional form of the model (i.e., the number of higher-order and/or interaction terms to be included).

As noted by Smith and Todd (2005) and references cited therein, using an overly crude propensity score specification is likely to yield biased estimates of the causal effect of the treatment. The inclusion of irrelevant variables in the propensity score model may have a similar impact, however. Throughout this article, we refer to the inclusion of two types of irrelevant variables in models used to estimate the propensity score: higher-order (and/or interaction) terms involving variables that are relevant at a lower order, and terms of any order involving variables that are wholly irrelevant to the determination of outcomes (i.e., have no impact on outcomes). Similarly, we refer to the exclusion of two types of relevant variables: relevant higher-order (and/or interaction) terms involving variables that are included at a lower order, and relevant terms of any order involving variables that are excluded from the estimation.

Rubin and Thomas (1996) argued in favor of including variables in the propensity score model unless there is consensus that they do not belong. Bryson, Dorsett, and Purdon (2002) argued against including irrelevant variables on efficiency grounds; the variance of the treatment effect estimator is likely to increase. Brookhart et al. (2006) suggested that variables related to the outcome of interest should always be included in the propensity score specification, but variables only weakly related to the outcome—even if strongly related to treatment assignment—should be excluded, because including them results in a higher mean squared error of the treatment effect estimate. Zhao (2008) presented some evidence suggesting that including irrelevant variables is not harmful, whereas excluding relevant variables is harmful. Finally, the intuition behind the result of Hirano, Imbens, and Ridder (2003) that using the true propensity score is inefficient even when it is known suggests that overfitting the propensity score model may have little negative impact on the properties of treatment effect estimates in practice.

Currently, the most common approach used by applied researchers to address specification issues with respect to the propensity score is to conduct balancing tests in studies using the propensity score as a way to implement matching or stratification estimators. If a variable is found to be unbalanced, then the usual approach is respecify the propensity score model, adding (additional) higher-order terms and/or interactions (see, e.g., Dehejia and Wahba 1999). There is no universally accepted balancing test, however (Smith and Todd 2005). Recently, Shaikh et al. (2005) developed an alternative test for

Journal of Business & Economic Statistics
July 2009, Vol. 27, No. 3
DOI: 10.1198/jbes.2009.06045

misspecification of the propensity score (although investigating the consequences of misspecifying the propensity score due to overfitting is not a goal of the article). In related work, Todd (1996) assessed the finite-sample performance of alternative propensity score estimators in terms of *estimating the propensity score* via a Monte Carlo study, but did not examine how the performance of the treatment effect estimator is affected by the different propensity score estimation methods considered.

In light of this background, the present article has two aims. First, we assess the practical effects of misspecifying the propensity score model via a Monte Carlo study *on the estimation of the treatment effect*. Our baseline experiments examine specifications issues in the context of models in which where treatment assignment depends on only a single covariate. Focusing on two estimators of treatment effects, the (unnormalized) inverse probability weighted estimator of Horvitz and Thompson (1952) and the normalized inverse probability weighted estimator of Hirano and Imbens (2001), we examine the *exclusion* of relevant higher-order terms and the *inclusion* of irrelevant higher-order terms. The results indicate that in many cases, overfitting the propensity score model results in a more efficient estimator, and in the remaining cases, it does no worse than the correctly specified model. Similar results were reported by Ichimura and Linton (2005), who used a kernel estimator to estimate propensity scores. In addition, in cases where the estimated propensity score often takes values close to zero or unity, the normalized estimator of Hirano and Imbens (2001) outperforms the unnormalized estimator.

Second, given the popularity of propensity score-matching (PSM) estimators in applied work, we demonstrate that our conclusion regarding overfitting also holds when kernel PSM estimators are used. Finally, because models with only a single covariate are not characteristic of much applied research and do not allow us to address issues related to overfitting when *including* wholly irrelavant variables (i.e., variables not relevant at any order) or *excluding* relevant variables entirely, we run additional simulations with three potentially relevant covariates. Although the treatment effect estimators fare poorly in all cases where a relevant regressor is excluded, this performance is not worsened by overfitting among the included covariates. Moreover, overfitting is of little consequence when wholly irrelevant variables are included in the model. Consequently, because the penalty for overfitting is minimal, we recommend practitioners report a set of causal estimates corresponding to various models for the propensity score.

Our second goal is to illustrate these ideas while addressing two important questions in the international arena: (a) Does the World Trade Organization (WTO) harm the environment? and (b) do currency unions promote international trade? Using data from the 1990s, we find that the WTO is beneficial for environmental measures that are global in nature and less directly tied to the sale of products (such as timber) that are protected by the WTO. Moreover, these results are sensitive to the specification of the propensity score model in some cases, although not to the choice of estimator (unnormalized versus normalized). Using data from 1994–2002, we found a positive impact of adopting the euro on bilateral trade. Whereas statistical significance was unaffected by altering the specification of the propensity score model, the magnitude of the causal estimate was affected. Thus both applications illustrate the practical benefits of comparing results from many propensity score specifications.

The remainder of the article is organized as follows. Section 2 presents a brief review of the potential outcome framework and the weighting estimators analyzed. Section 3 describes the Monte Carlo study and results, and Section 4 presents applications. Section 5 concludes.

2. SETUP

Consider a random sample of N individuals from a large population indexed by $i = 1, \ldots, N$. Using the potential outcome framework (see, e.g., Neyman 1923; Fisher 1935; Roy 1951; Rubin 1974), let $Y_i(t)$ denote the potential outcome of individual i under treatment t, $t \in \mathrm{T}$. Here we consider only the case of binary treatments, $T = \{0, 1\}$. The causal effect of the treatment ($t = 1$) relative to the control ($t = 0$) is defined as the difference between the corresponding potential outcomes. Formally, this is written as

$$\tau_i = Y_i(1) - Y_i(0), \tag{1}$$

and the population average treatment effect (ATE) is given by

$$\tau = \mathrm{E}(\tau_i) = \mathrm{E}[Y_i(1) - Y_i(0)]. \tag{2}$$

For each individual, we observe the triple $\{Y_i, T_i, X_i\}$, where Y_i is the observed outcome, T_i is a binary indicator of the treatment received, and X_i is a vector of covariates. The only requirement of the covariates included in X_i is that they are predetermined (i.e., unaffected by T_i) and do not perfectly predict treatment assignment. The relationship between the potential and observed outcomes is given by

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0), \tag{3}$$

which clearly shows that only one potential outcome is observed for any individual. As such, estimating τ is not trivial, because there is an inherent missing-data problem; some assumptions are required to proceed.

One such assumption is *unconfoundedness* or selection on observables (Rubin 1974; Heckman and Robb 1985). Under this assumption, treatment assignment is said to be independent of potential outcomes conditional on the set of covariates, X. As a result, selection into treatment is random conditional on X, and the average effect of the treatment can be obtained by comparing outcomes of individuals in different treatment states with identical values of the covariates. To solve the dimensionality problem that is likely to arise if X is a lengthy vector, Rosenbaum and Rubin (1983) proposed using the propensity score, $P(X_i) = \Pr(T_i = 1|X_i)$, instead of X as a conditioning variable.

Given knowledge of the propensity scores and sufficient overlap between the distributions of the propensity scores across the $t = 1$ and $t = 0$ groups (typically referred to as the *common support* condition; see Dehejia and Wahba 1999; Smith and Todd 2005), the ATE can be estimated in various ways (see D'Agostino 1998 and Imbens 2004 for summaries). Here we focus on the inverse probability weighted estimator of Horvitz and Thompson (1952), given by

$$\tau = \mathrm{E}\left[\frac{Y \cdot T}{P(X)} - \frac{Y \cdot (1 - T)}{1 - P(X)}\right]. \tag{4}$$

A sample estimate of τ can be computed using estimated propensity scores as follows:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{Y_i T_i}{\hat{P}(X_i)} - \frac{Y_i(1-T_i)}{1-\hat{P}(X_i)}\right]. \quad (5)$$

We refer to (5) as the unnormalized estimator.

Alternatively, Hirano and Imbens (2001) proposed an estimator that assigns weights normalized by the sum of propensity scores for treated and untreated groups, instead of assigning equal weights of $1/N$ to each observation. Their estimator is given by

$$\hat{\tau}_{norm} = \left[\sum_{i=1}^{N}\frac{Y_i T_i}{\hat{P}(X_i)} \Big/ \sum_{i=1}^{N}\frac{T_i}{\hat{P}(X_i)}\right] - \left[\sum_{i=1}^{N}\frac{Y_i(1-T_i)}{1-\hat{P}(X_i)} \Big/ \sum_{i=1}^{N}\frac{(1-T_i)}{1-\hat{P}(X_i)}\right]. \quad (6)$$

The advantage of estimator in (6), which we call the normalized estimator, is that the weights sum to unity within each group (Imbens, Newey, and Ridder 2005). In addition, Imbens, Newey, and Ridder (2005) showed that the normalized estimator is asymptotically equivalent to the unnormalized estimator and thus is asymptotically efficient.

This exposition demonstrates that researchers face two issues in practice. First, because the propensity score is typically unknown, it must be estimated, and theory offers little guidance. In fact, even when the propensity score is known, using the *true* propensity score is inefficient in general (Robins and Rotnitzky 1995; Rubin and Thomas 1996; Hahn 1998; Hirano, Imbens, and Ridder 2003). Second, practitioners must decide between the unnormalized Horvitz and Thompson (1952) estimator and the normalized Hirano and Imbens (2001) estimator. Hirano, Imbens, and Ridder (2003) showed that the estimator in (6) achieves the semiparametric efficiency bound when the propensity scores are estimated using a series logit estimator (SLE) (see Geman and Hwang 1982) and proposed a feasible variance estimator. To help inform applied researchers, we assess the practical performance of the two estimators shown in (5) and (6), as well as various specifications of the propensity score model. With respect to the latter, we focus on whether it pays to overspecify the propensity score equation. We now turn to our Monte Carlo study.

3. MONTE CARLO STUDY

3.1 Baseline Experiments

Data Simulation. Our initial set of experiments abstracts from issues of covariate selection and interaction terms and instead focuses on the case of a single, known covariate. To compare the two weighting estimators, as well as assess the benefit of overspecifying the propensity score equation, we simulate the treatment assignment based on a number of different, and increasingly difficult to estimate, *true* models, and then compare the performance of the estimators in (5) and (6). For each of the true models, we simulate 1000 data sets. Each data set contains 3 variables for each of the 1000 observations: T_i, a binary indicator of the treatment received, Y_i, the outcome, and the single covariate, $X_i \sim \mathrm{U}[0, 1]$.

To simulate treatment assignment, we first simulate the true propensity score, P_i, for each observation, and then draw T_i from a Bernoulli distribution with parameter P_i. We simulate the propensity scores from nine distinct settings, ranging from continuous, smooth, and monotone to noncontinuous and nonmonotone. Thus we compare the performance of the causal estimators using models for which SLE would be expected to perform well (e.g., continuous and smooth propensity scores), as well as more difficult models (e.g., noncontinuous propensity scores). The nine settings are as follows:

I. Logit specification:

$$P_i = \frac{\exp(A_i)}{1+\exp(A_i)},$$

where A_i is a polynomial as follows:
   a. Flat: $A_i = 0$
   b. Linear: $A_i = -3 + 6X_i$
   c. Quadratic and symmetric: $A_i = 2.5 - [2.5(1-2X_i)]^2$
   d. Quadratic and nonsymmetric: $A_i = 2.5 - [2.5(1-1.5X_i)]^2$
   e. Fourth degree: $A_i = -2.5 + 4.5X_i^4$.

II. Peak:
   a. Symmetric:

$$P_i = \begin{cases} 0.05 + 1.8X_i & \text{if } X_i < 0.5 \\ 1.85 - 1.8X_i & \text{if } X_i \geq 0.5 \end{cases}$$

   b. Nonsymmetric:

$$P_i = \begin{cases} 0.05 + 1.125X_i & \text{if } X_i < 0.8 \\ 4.55 - 4.5X_i & \text{if } X_i \geq 0.8. \end{cases}$$

III. Step:
   a. Monotonic:

$$P_i = \begin{cases} 0.33 & \text{if } X_i < 0.33 \\ 0.50 & \text{if } X_i \in [0.33, 0.67) \\ 0.67 & \text{if } X_i \geq 0.67 \end{cases}$$

   b. Nonmonotonic:

$$P_i = \begin{cases} 0.50 & \text{if } X_i < 0.33 \\ 0.90 & \text{if } X_i \in [0.33, 0.67) \\ 0.10 & \text{if } X_i \geq 0.67. \end{cases}$$

Although the relationships between the propensity scores and covariate are continuous and smooth in all of the logit specifications, specifications Ia, Ib, and Ie represent monotonic relationships, and specifications Ic and Id are nonmonotonic. In addition, specification Ie also is asymmetric. Specifications IIa and IIb, on the other hand, represent nonsmooth but continuous relationships, whereas specifications IIIa and IIIb represent noncontinuous relationships. The propensity scores are presented in Figure 1. Note that all specifications satisfy the common support condition.

Outcomes are simulated using the following specification:

$$Y_i = \beta_1 + \beta_2 T_i + \beta_3 X_i + \beta_4 T_i X_i + \varepsilon_i, \quad (7)$$

where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. To determine the values of the parameters, we use the data from Lalonde (1986). Specifically, we estimate (7) using these data, where X represents the first principal component of the set of the covariates used by Deheija and

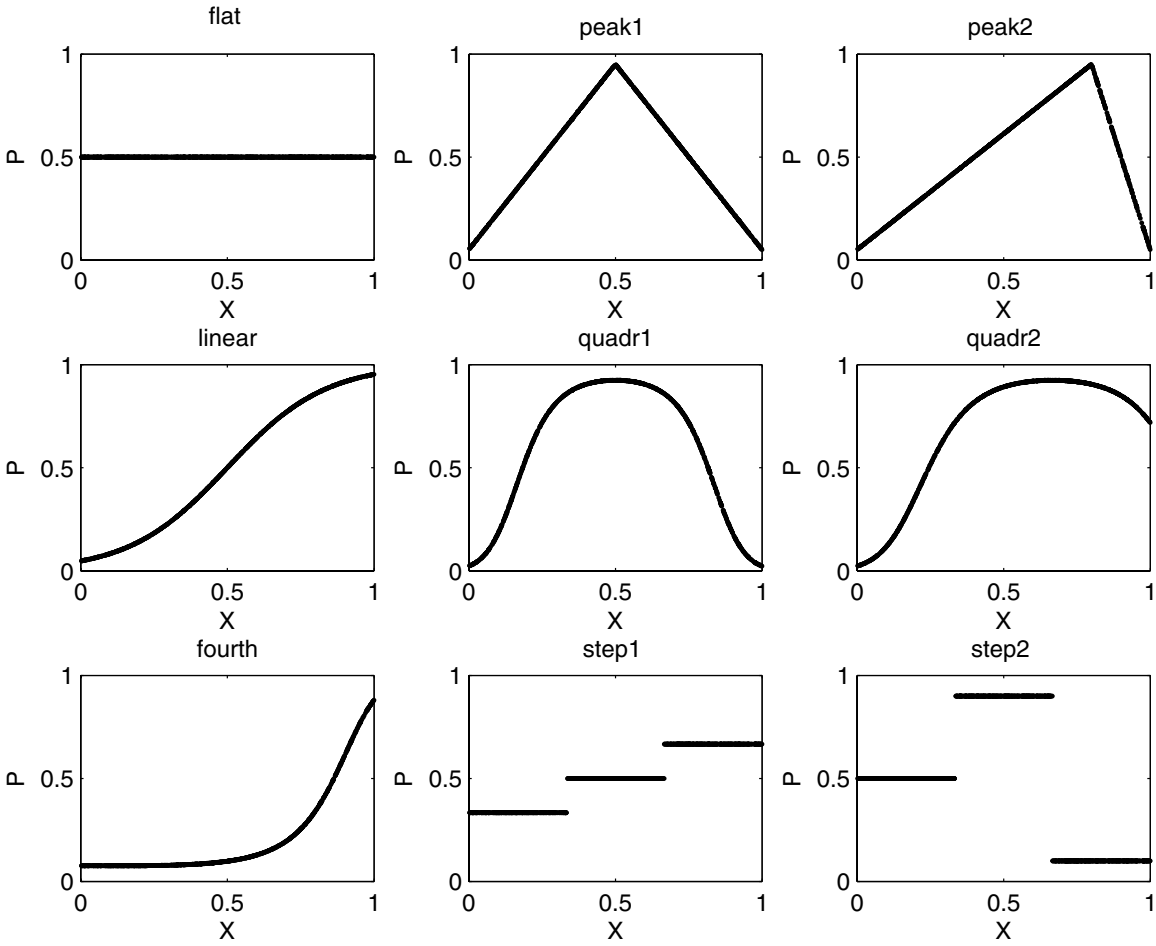

Figure 1. Univariate propensity score specifications.

Wahba (1999). Here $\beta = \{\beta_j\}_{j=1}^4$ are $[2.54, -1.67, -2.47, 1.96]$ and $\sigma^2 = 0.25$.

Estimation and Evaluation. For each data set, we estimate the propensity scores using a SLE. Following Hirano, Imbens, and Ridder (2003), we estimate the propensity scores as

$$\widehat{P}(x) = \frac{\exp[R^K(X)'\widehat{\delta}_K]}{1 + \exp[R^K(X)'\widehat{\delta}_K]},$$

where $R^K(X)$ is a vector of length K with the jth element equal to $R_j^K(x) = X^{(j-1)}$, $j = 1, \ldots, K$, and $\widehat{\delta}$ is the vector of maximum likelihood estimates. For each true model, we use $K = 1, \ldots, 10$. Thus we estimate 10 logit models with increasing degree of polynomial in the logit function.

The variance is computed as done by Hirano, Imbens, and Ridder (2003, p. 1172); specifically,

$$\hat{V} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\psi}_i + \hat{\alpha}_i)^2, \tag{8}$$

where

$$\hat{\psi} = \frac{Y_i T_i}{\hat{P}(X_i)} - \frac{Y_i(1-T_i)}{1-\hat{P}(X_i)} - \hat{\tau}$$

and

$$\hat{\alpha}_i = -\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i T_i}{\hat{P}(X_i)^2} + \frac{Y_i(1-T_i)}{(1-\hat{P}(X_i))^2}\right)R^K(X_i)\right]' \times \left(\frac{1}{N}\sum_{i=1}^{N} R^K(X_i)R^K(X_i)'\right)^{-1} R^K(X_i)(T_i - \hat{P}(X_i)).$$

To evaluate the performance of the estimators, we report the average bias, mean squared error (MSE), estimated variance, and percent coverage of the 95% confidence interval. In addition, we compute two measures of fit of the estimated propensity scores to the *true* propensity scores, namely mean integrated squared error (MISE),

$$MISE = \frac{1}{M}\sum_{j=1}^{M}[\hat{P}(Z_j) - P(Z_j)]^2,$$

and sup-norm,

$$SUP = \sup_{z\in[0,1]} |\hat{P}(Z_j) - P(Z_j)|,$$

where $P(\cdot)$ and $\hat{P}(\cdot)$ are the true and estimated propensity scores, evaluated at a grid of M points in $[0, 1]$ interval ($M = 100$). Lower values of both measures represent a better fit. Finally, we report the semiparametric efficiency bound for each true model; see the Appendix for the derivation.

Results. The results are presented in Table 1. Note that although in principal there is no issue of common support in the simulated data, we follow the suggestion of Imbens (2004) and estimate the treatment effect using only those observations with a propensity score between 0.02 and 0.98. Although this has little impact on the results, it ensures that no observation receives more than 5% weight in the causal estimator. We refer to this as the "trimming" level.

Panel I contains the results using specification Ia for the propensity score. Here the true propensity score is flat, implying that treatment is random (independent of X). This specification provides a nice check, because we expect all models to perform equally well. The results are as expected; the biases are small and equal for both causal estimators, $\widehat{\tau}$ and $\widehat{\tau}_{norm}$, and the MSEs, as well as the variance estimator, converge to the efficiency bound. However, some interesting results are observed even in this case. First, both measures of fit of the estimated propensity scores are minimized when the correct model is used (logit with zero-degree polynomial), but MSEs and the variance are higher than in overfitted models. As a result, the coverage rate exceeds 95% in the first column and exceeds the coverage rate in the remaining columns. This result is closely related to the findings of Hirano, Imbens, and Ridder (2003), because the estimated propensity score using the zero-degree polynomial is exactly the true propensity score.

Panel II presents the results for the linear model (specification Ib), which are similar to those in panel I. Specifically, the biases and MSEs are high in the underfitted model but settle

Table 1. Monte Carlo results: weighting estimators with one covariate

		Polynomial order in propensity score estimation									
		0	1	2	3	4	5	6	7	8	9
I. Flat propensity score											
1000×	Bias	−0.62	0.88	0.88	0.87	0.87	0.89	0.86	0.87	0.84	0.91
	Bias (normalized)	−0.62	0.87	0.88	0.88	0.88	0.90	0.88	0.89	0.86	0.91
	MSE	1.97	1.32	1.32	1.32	1.32	1.32	1.32	1.32	1.33	1.32
	MSE (normalized)	1.97	1.31	1.32	1.31	1.32	1.32	1.31	1.32	1.32	1.32
	Estimated variance	1.98	1.32	1.32	1.32	1.32	1.32	1.32	1.32	1.33	1.33
Propensity	MISE	0.0002	0.0005	0.0007	0.0010	0.0012	0.0014	0.0017	0.0019	0.0021	0.0023
Score	Sup norm	0.01	0.03	0.05	0.07	0.08	0.10	0.11	0.12	0.13	0.13
Coverage	Unnormalized	0.967	0.954	0.953	0.953	0.951	0.951	0.950	0.950	0.947	0.949
Rates	Normalized	0.967	0.955	0.953	0.953	0.952	0.951	0.950	0.950	0.948	0.949
II. Linear propensity score											
1000×	Bias	−536.38	1.50	0.22	1.06	0.46	1.47	1.48	1.43	1.51	1.72
	Bias (normalized)	−536.38	−1.13	−0.70	0.49	0.70	1.57	1.77	1.91	3.19	2.91
	MSE	289.33	3.30	2.92	2.90	3.06	3.20	3.27	3.26	3.37	3.43
	MSE (normalized)	289.33	3.61	3.24	2.90	3.12	3.17	3.25	3.27	3.37	3.44
	Estimated variance	1.63	3.30	2.92	2.90	3.06	3.20	3.27	3.27	3.37	3.44
Propensity	MISE	0.1287	0.0027	0.0027	0.0029	0.0030	0.0031	0.0032	0.0034	0.0035	0.0036
Score	Sup norm	0.49	0.07	0.08	0.08	0.09	0.09	0.10	0.11	0.11	0.12
Coverage	Unnormalized	0.000	0.959	0.947	0.953	0.938	0.930	0.929	0.928	0.924	0.924
Rates	Normalized	0.000	0.952	0.935	0.944	0.934	0.936	0.928	0.925	0.923	0.918
III. Quadratic (symmetric) propensity score											
1000×	Bias	−0.49	−1.87	−1.07	−3.12	1.58	1.37	1.74	2.49	3.61	3.49
	Bias (normalized)	−0.49	−0.18	0.42	0.49	0.56	0.17	0.86	1.01	0.75	0.30
	MSE	3.32	1.45	5.93	5.81	3.53	3.57	3.45	3.48	3.50	3.66
	MSE (normalized)	3.32	1.44	3.26	3.26	3.25	3.29	3.27	3.41	3.38	3.47
	Estimated variance	3.32	1.45	5.93	5.81	3.53	3.57	3.45	3.48	3.49	3.65
Propensity	MISE	0.1042	0.1041	0.0005	0.0006	0.0007	0.0008	0.0009	0.0011	0.0012	0.0013
Score	Sup norm	0.56	0.58	0.04	0.05	0.05	0.06	0.07	0.07	0.08	0.08
Coverage	Unnormalized	0.941	0.942	0.962	0.957	0.951	0.938	0.951	0.945	0.943	0.936
Rates	Normalized	0.941	0.944	0.987	0.990	0.960	0.946	0.951	0.941	0.946	0.945
IV. Quadratic 2 (asymmetric) propensity score											
1000×	Bias	−595.14	−23.45	2.45	5.83	9.45	12.60	13.85	16.88	18.42	20.23
	Bias (normalized)	−595.14	395.21	2.83	1.66	8.53	11.66	12.91	15.36	16.71	18.35
	MSE	357.05	9.19	4.96	4.10	3.54	3.79	3.85	4.09	4.10	4.12
	MSE (normalized)	357.05	167.20	3.43	3.40	3.33	3.53	3.57	3.75	3.78	3.80
	Estimated variance	2.86	8.65	4.96	4.07	3.46	3.64	3.66	3.80	3.76	3.71
Propensity	MISE	0.0880	0.0157	0.0004	0.0005	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012
Score	Sup norm	0.66	0.24	0.05	0.06	0.06	0.07	0.08	0.09	0.09	0.10
Coverage	Unnormalized	0.000	0.984	0.953	0.955	0.948	0.944	0.941	0.936	0.939	0.943
Rates	Normalized	0.000	0.111	0.985	0.965	0.944	0.944	0.942	0.937	0.945	0.946

Table 1. (Continued)

		Polynomial order in propensity score estimation									
		0	1	2	3	4	5	6	7	8	9
V. Fourth order propensity score											
1000×	Bias	−278.61	210.82	26.01	2.07	2.15	3.24	4.33	4.83	5.62	6.09
	Bias (normalized)	−278.61	61.54	−6.97	0.41	1.87	2.66	3.77	4.88	4.97	5.59
	MSE	79.61	54.89	4.07	2.95	3.04	3.08	3.15	3.19	3.26	3.35
	MSE (normalized)	79.61	9.58	3.05	3.00	3.03	3.03	3.11	3.22	3.23	3.27
	Estimated variance	1.99	10.46	3.40	2.95	3.03	3.07	3.13	3.17	3.23	3.32
Propensity	MISE	0.0462	0.0068	0.0008	0.0005	0.0006	0.0007	0.0008	0.0010	0.0011	0.0012
Score	Sup norm	0.65	0.24	0.07	0.05	0.06	0.06	0.07	0.08	0.09	0.09
Coverage	Unnormalized	0.000	0.973	0.937	0.950	0.944	0.938	0.944	0.942	0.945	0.941
Rates	Normalized	0.000	0.992	0.958	0.942	0.944	0.940	0.942	0.938	0.940	0.937
VI. Peak (symmetric) propensity score											
1000×	Bias	1.89	1.51	51.79	51.29	21.31	21.03	17.87	17.52	11.54	11.44
	Bias (normalized)	1.89	2.38	2.68	2.61	2.33	2.48	2.61	2.86	2.61	2.34
	MSE	2.45	1.21	5.24	5.31	2.44	2.41	2.70	2.72	2.42	2.44
	MSE (normalized)	2.45	1.21	2.09	2.15	1.90	1.89	2.24	2.27	2.21	2.25
	Estimated variance	2.45	1.21	2.56	2.68	1.99	1.97	2.38	2.41	2.29	2.31
Propensity	MISE	0.0664	0.0667	0.0036	0.0038	0.0027	0.0029	0.0019	0.0020	0.0019	0.0021
Score	Sup norm	0.46	0.47	0.16	0.16	0.13	0.13	0.10	0.10	0.10	0.10
Coverage	Unnormalized	0.956	0.959	0.859	0.869	0.925	0.917	0.938	0.939	0.936	0.931
Rates	Normalized	0.956	0.959	0.971	0.971	0.954	0.950	0.950	0.947	0.949	0.946
VII. Peak 2 (asymmetric) propensity score											
1000×	Bias	−269.69	12.96	56.30	−20.55	11.24	11.58	4.18	7.83	8.70	5.07
	Bias (normalized)	−269.69	81.38	−16.50	−23.09	−11.78	−12.66	−9.22	−2.01	−0.88	−1.04
	MSE	75.06	1.77	6.82	2.30	2.32	2.31	1.97	2.17	2.23	2.16
	MSE (normalized)	75.06	8.19	2.63	2.28	2.19	2.28	2.09	2.06	2.10	2.11
	Estimated variance	2.33	1.61	3.65	1.88	2.19	2.18	1.95	2.11	2.15	2.13
Propensity	MISE	0.0665	0.0404	0.0134	0.0052	0.0020	0.0021	0.0020	0.0019	0.0020	0.0021
Score	Sup norm	0.46	0.66	0.31	0.19	0.13	0.13	0.12	0.11	0.12	0.11
Coverage	Unnormalized	0.000	0.933	0.989	0.932	0.961	0.951	0.950	0.945	0.940	0.947
Rates	Normalized	0.000	0.471	0.976	0.931	0.953	0.944	0.949	0.945	0.952	0.952
VIII. Step (monotone) propensity score											
1000×	Bias	−222.99	1.21	1.34	−0.77	−0.74	−0.36	−0.34	0.14	0.08	−0.36
	Bias (normalized)	−222.99	1.40	1.68	−0.86	−0.82	−0.33	−0.37	0.37	0.27	−0.26
	MSE	51.68	1.37	1.37	1.36	1.37	1.37	1.37	1.39	1.39	1.38
	MSE (normalized)	51.68	1.40	1.37	1.36	1.36	1.37	1.37	1.39	1.39	1.38
	Estimated variance	1.95	1.37	1.37	1.36	1.37	1.37	1.37	1.39	1.39	1.38
Propensity	MISE	0.0188	0.0024	0.0026	0.0025	0.0027	0.0028	0.0030	0.0028	0.0030	0.0028
Score	Sup norm	0.18	0.180	0.11	0.12	0.12	0.13	0.13	0.14	0.14	0.14
Coverage	Unnormalized	0.001	0.952	0.953	0.951	0.953	0.954	0.956	0.957	0.956	0.955
Rates	Normalized	0.001	0.953	0.954	0.951	0.9570	0.952	0.954	0.957	0.957	0.956
IX. Step 2 (nonmonotone) propensity score											
1000×	Bias	265.03	−158.92	11.93	16.21	35.54	110.71	−32.99	−18.60	−85.39	−99.49
	Bias (normalized)	265.03	−73.83	−80.50	−91.67	−38.53	−50.02	−15.55	−16.51	−10.01	−4.18
	MSE	72.41	27.31	6.61	5.73	5.40	19.91	5.38	5.95	13.11	16.16
	MSE (normalized)	72.41	6.88	9.82	11.94	4.37	6.76	3.25	3.76	3.62	3.55
	Estimated variance	2.17	2.06	6.47	5.48	4.14	7.66	4.29	5.61	5.83	6.27
Propensity	MISE	0.1069	0.0832	0.0361	0.0329	0.0194	0.0166	0.0123	0.0105	0.0102	0.0099
Score	Sup norm	0.41	0.49	0.43	0.47	0.42	0.45	0.40	0.40	0.40	0.40
Coverage	Unnormalized	0.000	0.058	0.953	0.956	0.938	0.878	0.938	0.962	0.957	0.954
Rates	Normalized	0.000	0.717	0.861	0.768	0.937	0.942	0.976	0.985	0.991	0.996

NOTE: Estimated variance uses formula from Hirano, Imbens, and Ridder (2003). The asymptotic variance bound is 1.32 in panel I, 2.49 in panel II, 2.85 in panel III, 2.85 in panel IV, 2.85 in panel V, 1.96 in panel VI, 1.96 in panel VII, 1.40 in panel VIII, and 2.50 in panel IX. A total of 1000 simulations were used for each column, with 1000 observations per simulation. Propensity scores were trimmed at 0.02 and 0.98. See the text for further details.

down quickly, with little penalty in terms of MSE and coverage rates for overfitting. In addition, there is little difference between the unnormalized and normalized estimators.

Panel III presents the results for the symmetric quadratic model (specification Ic). These results are very surprising, indicating equally good performance in underfitted, correct, and overfitted models, with the normalized estimator fairing slightly better in the overfitted models in terms of biases, MSEs, and coverage rates. This occurs because the true propensity score is symmetric. To see this, suppose that the marginal distribution of X is U[0, 1] and the propensity score is symmetric about E[X], which is 0.5. Furthermore, note that when the estimated propensity score is constant across observations (i.e., we use a zero-degree polynomial), (4) is equivalent to the simple difference in means estimator, given by

$$\hat{\tau}_s = \frac{1}{N_1}\sum_{i=1}^{N} Y_i T_i - \frac{1}{N_0}\sum_{i=1}^{N} Y_i(1 - T_i), \tag{9}$$

where N_1 (N_0) is the size of the treatment (control) group. The first component in (9) is the mean of the $T = 1$ observations and will have mean (and converge in probability to) E[$Y|T = 1$]. Now E[$Y|T = 1$] = E[E[$Y|T = 1, X]|T = 1$] = $\int$ E[$Y|T = 1, X$]$p(X|T = 1)\,dX$, where $p(X|T = 1)$ is the conditional density of X given $T = 1$, which, by Bayes's rule, is $p(X|T = 1) = p(T = 1|X)p(X)/p(T = 1)$. Because E[$Y|T = 1$, X] = $(\beta_1 + \beta_2) + (\beta_3 + \beta_4)X$, it follows that

$$\begin{aligned}
\mathrm{E}[Y|T = 1] &= \int [(\beta_1 + \beta_2) + (\beta_3 + \beta_4)X]p(X|T = 1)\,dX \\
&= (\beta_1 + \beta_2)\int p(X|T = 1)\,dX \\
&\quad + (\beta_3 + \beta_4)\int Xp(X|T = 1)\,dX \\
&= (\beta_1 + \beta_2) + (\beta_3 + \beta_4)\int Xp(X|T = 1)\,dX \\
&= (\beta_1 + \beta_2) + (\beta_3 + \beta_4)\mathrm{E}[X|T = 1].
\end{aligned}$$

Furthermore, because $p(T = 1|X)$ is symmetric about 0.5, it follows that $p(X|T = 1)$ is symmetric about 0.5. Thus E[$X|T = 1$] = 1/2 = E[X] and E[$Y|T = 1$] = $(\beta_1 + \beta_2) + (\beta_3 + \beta_4)$E[$X$] = E[$Y(1)$]. A similar argument follows for the case where $T = 0$. As a result, the simple estimator in (9) is unbiased and consistent in this case. (We are indebted to Keisuke Hirano for this argument.)

The results are markedly different for the asymmetric quadratic model (specification Id), as shown in panel IV. The biases and MSEs are very high for underfitted models [specifications (0) and (1)], with low penalty, in terms of MSE, for overfitted models. In addition, the estimator using normalized weights performs better in the overfitted models in terms of biases, MSEs, and coverage rates, as was the case in the symmetric quadratic model.

The results for the final logit specification, polynomial of fourth degree (specification Ie), given in panel V, are not surprising. First, relatively extreme underfitting [specifications (0) and (1)] leads to very poor performance by both estimators. Second, there is little penalty for overfitting. Finally, there is little difference between the two estimators in overfitted models.

The remaining results pertain to data simulated from nonlogit true models. As a result, there is no correct estimated model. Results for the symmetric peak model (specification IIa) are given in panel VI. As in the previous symmetric specifications (Ia and Ic), we see little substantive differences across models. However, the biases, MSEs, and coverage probabilities improve when using the normalized estimator, because of the numerous estimated propensity scores close to zero or unity. In contrast, the results for the asymmetric peak model (specification IIb), given in panel VII, demonstrate obvious benefits to overfitting the propensity score, as well as little substantive difference between the unnormalized and normalized estimators.

The final set of results uses data simulated using noncontinuous propensity scores. Results for the monotonic step function (specification IIIa), given in panel VIII, indicate that all models, as well as both estimators, perform equally well, with the exception of the zero-degree polynomial logit model. The results for the nonmonotonic (asymmetric) step function (specification IIIb), displayed in panel IX, are more interesting. First, there is little penalty in terms of bias and MSE for overfitting, particularly with the normalized estimator. Second, the benefits of using the normalized estimator over the unnormalized estimator are extremely pronounced; biases and MSEs are lower in most models. This is due to the fact that one-third of the true propensity scores are close to zero, and normalized weights attenuate the effect of the small values of the estimated propensity scores in the denominator.

In sum, the results of the Monte Carlo study yield three salient implications. First, there is almost no penalty for overfitting the propensity score model and often a large penalty for underfitting. Because in practice researchers rarely know the true form of the propensity score, it appears prudent to overfit. Second, when a relatively large number of estimated propensity scores lie close to zero or unity, the normalized estimator performs better, performing relatively poorly only when propensity scores are underfitted. Thus overfitting in combination with the normalized estimator seems wise. Finally, whether the true propensity score is continuous or not does not appear to affect the performance of the different estimators and different models used to estimate the propensity score.

Based on these results, practitioners should use the normalized estimator and provide a series of estimates using increasingly sophisticated specifications of the propensity score model. One potential caveat to this conclusion pertains to sample size. In our Monte Carlo experiments, the degrees of freedom remained quite large even in the overfitted models. Overfitting may be less advisable in practice when dealing with small samples. We return to this point later in our application. We also explored whether our findings hold in other contexts, as discussed next.

3.2 Additional Experiments

To assess the applicability of the preceding results to additional situations, we undertook three additional experiments. First, we assessed the effect of overfitting on the performance of PSM estimators. Second, we extended our study to incorporate issues related to covariate selection. Finally, we also assessed

the impact of overfitting when the parameter of interest is not the population ATE, but rather the ATE for some subpopulation, given by $\tau(X) = \mathrm{E}[\tau|X] = \mathrm{E}[Y(1) - Y(0)|X]$. Specifically, we assessed the impact of overfitting in our experiments with a single covariate when the parameter of interest is $\mathrm{E}[\tau|X > 0.5]$. Here we provide details of the first two additional experiments. For the third set of experiments, we omit the results for the sake of brevity, and simply note that our conclusions with respect to overfitting remain robust.

Propensity Score Matching Estimator. To examine the more common PSM estimator, we used the same simulated data from the baseline experiments. We used kernel matching with the Gaussian kernel and several bandwidths (0.02, 0.10, 0.25, and 0.50) to assess sensitivity to this choice. The results, displayed in Table 2, were obtained using *PSMATCH2* in Stata, with the common support imposed by including only those observations within the region of overlap between the treatment and control groups.

Table 2. Monte Carlo results: matching estimators

		Polynomial order in propensity score estimation									
		0	1	2	3	4	5	6	7	8	9
I. Flat propensity score											
1000×	Bias (BW = 0.02)	−0.62	0.61	0.85	1.69	1.45	1.31	0.54	1.08	1.08	0.42
	Bias (BW = 0.10)	−0.62	−0.55	−0.35	−0.05	−0.11	−0.15	−0.22	0.17	0.26	0.11
	Bias (BW = 0.25)	−0.62	−0.67	−0.61	−0.38	−0.46	−0.59	−0.60	−0.30	−0.35	−0.44
	Bias (BW = 0.50)	−0.62	−0.69	−0.65	−0.43	−0.52	−0.67	−0.67	−0.40	−0.48	−0.55
	MSE (BW = 0.02)	1.97	1.50	1.66	1.62	1.62	1.58	1.54	1.51	1.48	1.49
	MSE (BW = 0.10)	1.97	1.94	1.96	1.95	1.95	1.96	1.93	1.89	1.89	1.89
	MSE (BW = 0.25)	1.97	1.99	1.99	1.97	1.97	1.99	1.97	1.95	1.96	1.96
	MSE (BW = 0.50)	1.97	2.00	1.99	1.98	1.98	2.00	1.98	1.97	1.98	1.99
II. Linear propensity score											
1000×	Bias (BW = 0.02)	−536.38	−2.93	−2.96	−2.68	−2.68	−2.55	−3.39	−3.04	−3.26	−2.52
	Bias (BW = 0.10)	−536.38	−58.77	−58.73	−58.32	−58.13	−58.26	−58.27	−56.78	−56.30	−55.15
	Bias (BW = 0.25)	−536.38	−216.65	−216.34	−216.50	−216.37	−216.03	−215.89	−215.33	−214.72	−213.50
	Bias (BW = 0.50)	−536.38	−410.38	−410.39	−410.46	−410.57	−410.13	−409.38	−408.25	−406.65	−404.65
	MSE (BW = 0.02)	289.33	3.05	3.05	3.07	3.08	3.07	3.09	3.11	3.08	3.05
	MSE (BW = 0.10)	289.33	6.00	5.97	5.91	5.97	5.93	5.96	5.85	5.77	5.62
	MSE (BW = 0.25)	289.33	49.00	48.86	48.95	48.94	48.77	48.75	48.49	48.22	47.70
	MSE (BW = 0.50)	289.33	170.54	170.58	170.62	170.73	170.33	169.76	168.81	167.48	165.89
III. Quadratic (symmetric) propensity score											
1000×	Bias (BW = 0.02)	−0.49	0.50	−0.28	−0.13	−0.16	0.19	−0.06	0.56	0.31	0.83
	Bias (BW = 0.10)	−0.49	1.02	−0.83	−0.61	−0.65	−0.39	−0.69	0.09	−0.06	0.41
	Bias (BW = 0.25)	−0.49	1.09	−1.15	−1.32	−1.35	−0.83	−1.04	−0.43	−0.59	0.17
	Bias (BW = 0.50)	−0.49	1.10	−0.26	−0.95	−1.04	−0.33	−0.68	−0.16	−0.43	0.60
	MSE (BW = 0.02)	3.32	6.38	3.10	3.40	3.35	3.43	3.39	3.52	3.42	3.43
	MSE (BW = 0.10)	3.32	9.60	2.69	3.07	3.03	3.03	3.02	3.61	3.54	3.47
	MSE (BW = 0.25)	3.32	9.85	2.27	3.28	3.28	3.32	3.27	3.80	3.65	3.71
	MSE (BW = 0.50)	3.32	9.89	2.46	4.96	4.92	5.96	5.85	6.42	6.02	6.05
IV. Quadratic 2 (asymmetric) propensity score											
1000×	Bias (BW = 0.02)	−595.14	35.64	22.89	23.36	24.02	24.36	26.02	25.06	25.02	26.48
	Bias (BW = 0.10)	−595.14	31.85	3.61	3.03	2.09	2.54	2.67	1.83	1.00	0.85
	Bias (BW = 0.25)	−595.14	−212.64	−136.05	−136.30	−137.05	−137.68	−137.69	−136.43	−136.04	−136.15
	Bias (BW = 0.50)	−595.14	−444.92	−374.59	−374.60	−374.72	−374.92	−374.72	−373.49	−372.66	−371.57
	MSE (BW = 0.02)	357.05	4.56	4.05	4.01	4.03	4.10	4.15	4.07	4.05	4.16
	MSE (BW = 0.10)	357.05	4.18	3.39	3.43	3.51	3.64	3.75	3.65	3.62	3.59
	MSE (BW = 0.25)	357.05	48.26	22.07	22.11	22.38	22.55	22.54	22.32	22.23	22.26
	MSE (BW = 0.50)	357.05	202.30	144.74	144.81	144.87	144.98	144.74	143.86	143.22	142.38
V. Fourth order propensity score											
1000×	Bias (BW = 0.02)	−278.61	6.27	−8.00	0.72	2.40	3.71	4.99	6.25	6.60	7.55
	Bias (BW = 0.10)	−278.61	−32.06	−42.04	−39.62	−38.67	−36.94	−35.55	−33.77	−32.96	−32.14
	Bias (BW = 0.25)	−278.61	−163.34	−149.18	−140.12	−138.83	−137.22	−135.96	−134.46	−133.74	−132.63
	Bias (BW = 0.50)	−278.61	−233.86	−236.82	−226.98	−225.48	−223.77	−222.39	−220.80	−219.84	−218.56
	MSE (BW = 0.02)	79.61	3.15	3.06	3.14	3.23	3.20	3.27	3.28	3.26	3.35
	MSE (BW = 0.10)	79.61	3.92	4.49	4.42	4.48	4.32	4.24	4.16	4.09	4.11
	MSE (BW = 0.25)	79.61	28.98	24.42	22.06	21.79	21.35	21.06	20.69	20.50	20.29
	MSE (BW = 0.50)	79.61	57.01	58.11	53.84	53.23	52.44	51.87	51.19	50.79	50.33

Table 2. (Continued)

		Polynomial order in propensity score estimation									
		0	1	2	3	4	5	6	7	8	9
VI. Peak (symmetric) propensity score											
1000×	Bias (BW = 0.02)	1.89	3.62	2.68	2.39	2.34	2.22	2.00	2.44	2.11	2.30
	Bias (BW = 0.10)	1.89	3.63	3.07	2.84	2.59	2.43	2.29	2.79	2.59	2.57
	Bias (BW = 0.25)	1.89	3.63	3.25	2.61	2.61	2.56	2.49	3.13	2.91	2.73
	Bias (BW = 0.50)	1.89	3.64	3.26	2.27	2.23	2.21	2.11	2.85	2.56	2.30
	MSE (BW = 0.02)	2.45	2.94	2.18	2.29	2.52	2.59	2.27	2.34	2.44	2.49
	MSE (BW = 0.10)	2.45	3.93	1.96	2.16	2.35	2.40	2.12	2.29	2.37	2.39
	MSE (BW = 0.25)	2.45	4.01	1.80	2.46	2.64	2.59	2.20	2.48	2.59	2.61
	MSE (BW = 0.50)	2.45	4.02	1.97	3.20	3.45	3.54	2.96	3.29	3.48	3.52
VII. Peak 2 (asymmetric) propensity score											
1000×	Bias (BW = 0.02)	−269.69	5.20	−42.43	−19.58	−3.49	−5.28	−5.54	2.40	2.85	0.70
	Bias (BW = 0.10)	−269.69	−49.92	−46.15	−47.11	−30.30	−33.02	−33.94	−26.73	−26.51	−28.67
	Bias (BW = 0.25)	−269.69	−206.26	−96.56	−123.01	−131.65	−135.41	−136.56	−128.43	−127.90	−130.24
	Bias (BW = 0.50)	−269.69	−254.55	−185.38	−218.05	−221.72	−226.49	−224.59	−211.23	−210.52	−213.12
	MSE (BW = 0.02)	75.06	2.02	3.67	2.48	2.12	2.22	2.28	2.24	2.17	2.26
	MSE (BW = 0.10)	75.06	4.42	3.89	4.23	2.86	3.16	3.26	2.79	2.72	2.91
	MSE (BW = 0.25)	75.06	45.06	11.32	17.27	19.29	20.57	20.93	18.70	18.51	19.20
	MSE (BW = 0.50)	75.06	67.98	37.02	50.18	51.63	54.16	53.39	47.49	47.12	48.36
VIII. Step (monotone) propensity score											
1000×	Bias (BW = 0.02)	−222.99	−3.73	−3.23	−2.32	−3.37	−2.89	−2.98	−0.71	−0.80	−1.28
	Bias (BW = 0.10)	−222.99	−3.73	−3.23	−2.32	−3.37	−2.89	−2.98	−0.71	−0.80	−1.28
	Bias (BW = 0.25)	−222.97	−181.28	−180.54	−179.17	−178.57	−177.79	−176.77	−175.69	−174.96	−173.56
	Bias (BW = 0.50)	−222.93	−209.80	−209.58	−209.60	−209.31	−208.98	−208.16	−207.85	−207.39	−206.84
	MSE (BW = 0.02)	51.68	1.40	1.41	1.42	1.42	1.41	1.43	1.43	1.42	1.44
	MSE (BW = 0.10)	51.68	1.40	1.41	1.42	1.42	1.41	1.43	1.43	1.42	1.44
	MSE (BW = 0.25)	51.67	34.42	34.20	33.70	33.44	33.20	32.85	32.46	32.20	31.71
	MSE (BW = 0.50)	51.65	45.87	45.80	45.81	45.63	45.53	45.20	45.07	44.87	44.64
IX. Step 2 (nonmonotone) propensity score											
1000×	Bias (BW = 0.02)	265.03	−3.97	−52.18	−25.45	−38.56	−12.46	−25.26	−22.44	−29.21	−27.71
	Bias (BW = 0.10)	265.03	56.82	−32.39	−26.94	−38.73	−18.32	−31.32	−30.74	−36.85	−38.93
	Bias (BW = 0.25)	265.03	183.55	59.42	53.62	30.45	26.06	9.36	3.57	1.63	−2.40
	Bias (BW = 0.50)	265.03	232.70	175.07	169.52	162.00	158.02	150.77	147.63	146.61	144.73
	MSE (BW = 0.02)	72.41	2.49	5.20	3.28	4.45	3.05	3.58	3.58	3.92	3.80
	MSE (BW = 0.10)	72.41	4.80	3.21	2.92	4.06	2.80	3.69	3.77	4.16	4.33
	MSE (BW = 0.25)	72.41	35.49	5.47	4.91	2.91	2.68	2.15	2.17	2.17	2.22
	MSE (BW = 0.50)	72.41	56.44	32.88	31.06	28.46	27.19	25.05	24.15	23.88	23.42

NOTE: Kernel matching using the Gaussian kernel. BW = bandwidth. Common support imposed. See Table 1 for further details.

Panel I presents the results from using specification Ia for the propensity score, with treatment assignment independent of X. Several interesting results can be seen. First, when the correct model is used (logit with zero-degree polynomial), the bias is lower than in most of the overfitted specifications using the smallest bandwidth; however, the bias is predominantly lower in all overfitted specifications using bandwidths above 0.02, and the MSEs are either essentially unchanged or reduced regardless of the bandwidth. In contrast, whereas we also obtained lower MSEs when we overspecified the propensity score in Table 1, the biases increased with the inclusion of X. Second, whereas the bias and MSE obtained using the weighting estimators remained unchanged across specifications (1)–(9) in Table 1, the bias and MSE of the PSM estimator are quite volatile across specifications, particularly at smaller bandwidths. Finally, the bias is smallest using a bandwidth of 0.10 (not 0.02), and the MSE is smallest using a bandwidth of 0.02, in all propensity score specifications including X.

The results for the linear model (specification Ib), presented in panel II, are equally telling. Regardless of bandwidth, there is no penalty for overfitting. These results are even stronger than those given in Table 1. In addition, both bias and MSE increase with the bandwidth. Panel III presents the results for the symmetric quadratic model (specification Ic), which continue to suggest little penalty for overfitting for bandwidths of 0.25 and smaller; overfitting does tend to raise the bias and MSE when the bandwidth is 0.50. One interesting difference relative to the weighting estimators, however, is that the penalty for underfitting is very large with the PSM estimator. The results in panel IV for the asymmetric quadratic model (specification Id) are similar with respect to overfitting at lower bandwidths, but here overfitting is not problematic even when the bandwidth is 0.50. But although the bandwidth does not impact the benefit

(or cost) of overfitting, the overall performance of the estimator is affected by the bandwidth; bias and MSE are smallest for all specifications when the bandwidth is 0.10. The results for the final logit specification, polynomial of fourth degree (specification Ie), given in panel V, are fairly similar to those obtained for the weighting estimators; as in the linear propensity score case (panel II), bias and MSE both increase with the bandwidth.

The remaining results pertain to data simulated from nonlogit *true* models, where, as noted earlier, there is no correct estimated model. Results for the symmetric peak model (specification IIa) are given in panel VI. As with the normalized weighting estimator, we see little substantive differences across the various specifications. In addition, there is a marginal improvement in bias and MSE at smaller bandwidths. On the other hand, the results for the asymmetric peak model (specification IIb), given in panel VII, show significant improvement in bias and MSE at smaller bandwidths. More importantly, there continues to be no penalty for overfitting, along with large gains in terms of bias reduction at smaller bandwidths. The final set of results pertaining to noncontinuous propensity scores are given in panels VIII (for specification IIIa) and IX (for specification IIIb). In both cases, there is little penalty, and sometimes a pronounced benefit (e.g., panel IX with a bandwidth of 0.25) for overfitting.

In sum, the results using the PSM estimators offer three primary conclusions. First, there is almost no penalty for overfitting the propensity score model, but often a large penalty for underfitting. Relative to weighting estimators, the cost of underfitting appears to be relatively smaller in many instances (but this is sensitive to true data-generating process and choice of bandwidth), and there also may be a *benefit* to overfitting (as opposed to simply a lack of a penalty). Thus, as with weighting estimators, it appears prudent for applied researcher to overfit. Second, with kernel PSM estimators, the efficiency gain in our experiments from larger bandwidths is far exceeded by the increase in the bias; both bias and MSE are greater with higher bandwidths (for the bandwidths that we use). Nonetheless, conclusions regarding the relative costs and benefits of underfitting and overfitting are predominantly invariant to the choice of bandwidth.

Multiple Regressors. Although our experimental results presented thus far are informative, the focus on a single covariate is not characteristic of most applications. It also forces us to abstract from issues of variable selection except as it relates to higher orders of the single regressor. Our final experiments expanded the set of regressors to three. Although also on the small side, a set of three regressors allowed us to search for situations in which overfitting may be costly. Specifically, we assess the robustness of the previous single-regressor simulations to multiple regressors and robustness to cases in which a wholly irrelevant variable included in the analysis (i.e., the irrelevant variable is not simply a higher order of an otherwise relevant variable) or a relevant variable is excluded entirely from the analysis, where the covariates may or may not be correlated.

Data Simulation. We simulated data from eight setups, which differ in terms of the specification of the true propensity score (linear or nonsymmetric quadratic), whether the covariates are correlated or independent, and the relationship between the three covariates and treatment assignment and the outcome

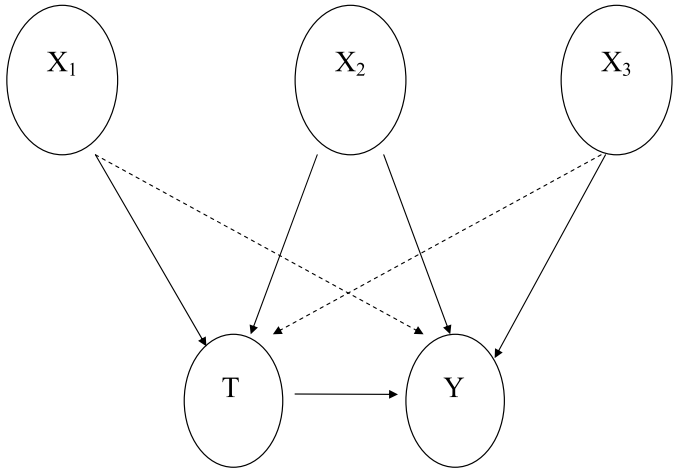

Figure 2. Experimental designs in the multiple-regressor case. The solid and dashed lines denote causal relationships in the experimental designs in which all three covariates are relevant. The solid lines denote causal relationships in the experimental designs where not all three covariates are relevant for both treatment assignment and outcomes.

(all relevant or only some relevant in the determination of outcomes). In four of the setups, all three X's are relevant for the determination of both treatment assignment and the outcome. In the remainder, two regressors are relevant for the determination of treatment assignment (X_1 and X_2), and two regressors are relevant for the determination of the outcome (X_2 and X_3). Here X_1 is irrelevant from the perspective of estimating the causal effect of the treatment, and its inclusion should lower the efficiency of the causal estimator, according to Brookhart et al. (2006); see Figure 2.

For each of these setups and for each of the *true* models, we simulated 1000 data sets, each of which contained 5 variables for each of the 1000 observations: T_i, Y_i, and the covariates, X_{1i}, X_{2i}, and X_{3i}. In four of the experiments, the three covariates were drawn independently from a U[0, 1] distribution; in the rest, the three covariates were drawn jointly from a U[0, 1] distribution (Fackler 2007), with a correlation matrix of

$$\begin{bmatrix} 1 & & \\ 0.3 & 1 & \\ 0.1 & 0.5 & 1 \end{bmatrix}.$$

To simulate treatment assignment, we again simulated the *true* propensity score, P_i, for each observation, and then drew T_i from a Bernoulli distribution with parameter P_i. We simulated the propensity scores from two logit specifications, where

$$P_i = \frac{\exp(A_i)}{1 + \exp(A_i)},$$

and A_i takes the following forms:

IV. Three relevant regressors:
   a. Linear: $A_i = 0.33[(-3+6X_{1i})+(-3+6X_{2i})+(-3+6X_{3i})]$
   b. Quadratic: $A_i = 0.33[\{2.5 - [2.5(1 - 1.5X_{1i})]^2\} + \{2.5 - [2.5(1 - 1.5X_{2i})]^2\} + \{2.5 - [2.5(1 - 1.5X_{3i})]^2\}]$.

V. Two relevant regressors:
   a. Linear: $A_i = 0.5[(-3 + 6X_{1i}) + (-3 + 6X_{2i})]$

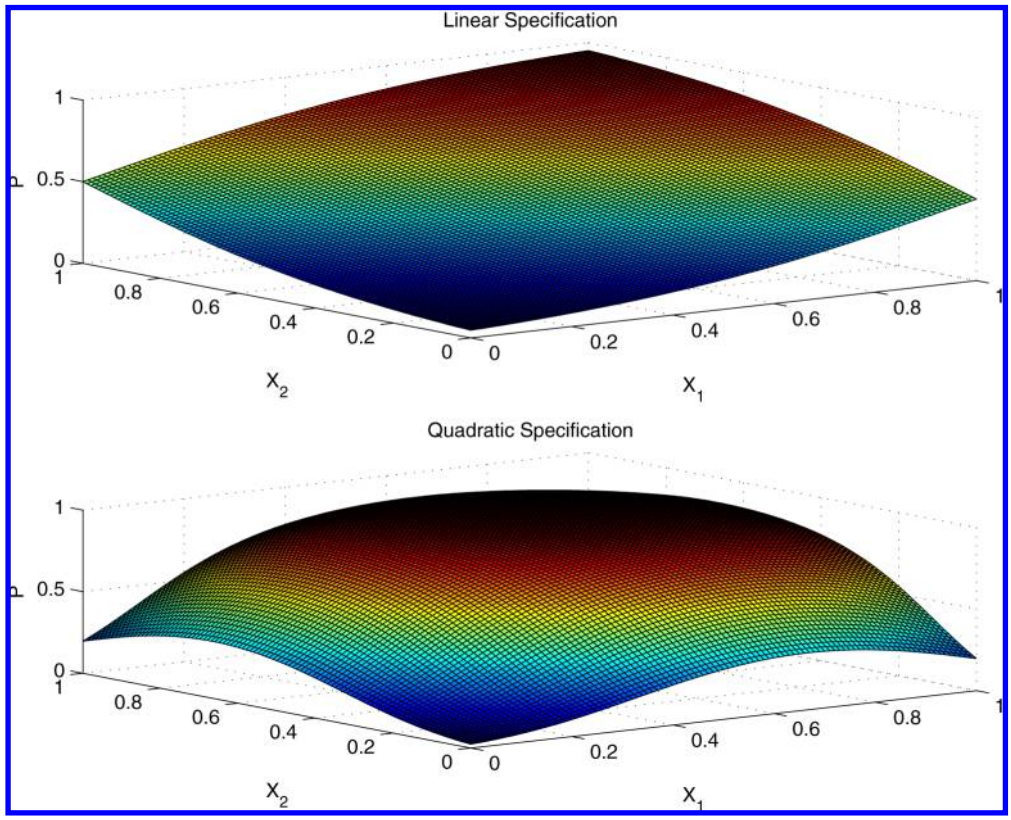

Figure 3. Multivariate propensity score specifications (two relevant regressors).

b. Quadratic: $A_i = 0.5[\{2.5-[2.5(1-1.5X_{1i})^2\}+\{2.5-[2.5(1-1.5X_{2i})^2\}]$.

The propensity scores are presented in Figure 3 for specifications Va and Vb. We did not attempt to plot the propensity scores for specifications IVa and IVb, which would require four dimensions. Note that all specifications satisfy the common support condition.

We simulated outcomes using the following specification:

$$Y_i = \beta_1 + \beta_2 T_i + \beta_3 X_i + \beta_4 T_i X_i + \varepsilon_i, \tag{10}$$

where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ and X_i is either a 3×1 or a 2×1 vector, depending on the number of relevant regressors. Parameter values are the same as in the univariate case.

For each data set, we estimated the propensity scores using a logit model with six different specifications. Specification (0) contains only a constant. Specification (1) adds linear terms for X_1, X_2, and X_3 (first-order linear approximation). Specification (2) adds interaction terms between each of the X's. Specification (3) adds quadratic terms for each X (second-order linear approximation). Specification (4) adds interaction terms between the first- and second-order terms. Specification (5) adds cubic terms for each X (third-order linear approximation). In the models excluding a relevant regressor, X_3 is excluded in specifications (1)–(5).

Results. Table 3 presents the results when all three covariates are used to estimate the propensity score and, as in the baseline experiments, the sample is trimmed to include only observations with a propensity score between 0.02 and 0.98. For brevity, we focus on the weighting estimators and report only bias and MSE. For the linear cases, specification (1) is the correct model. For the quadratic cases, there is no correct model, because specifications (1) and (2) omit relevant quadratic terms, and specifications (3)–(5) include irrelevant terms (e.g., the first-order interactions).

Panels I and II present the results for the linear and quadratic specifications where all three X's are correlated and are relevant for determining the outcome. Both panels show that there continues to be little penalty for overfitting (especially relative to underfitting), and in both cases the normalized estimator yields a lower MSE across all specifications. Panels III and IV present the results for similar specifications except with the X's independent. Again, there is little penalty for overfitting, a substantial penalty for underfitting in the quadratic case, and superior performance by the normalized estimator. Moreover, in the linear case (panel III), whereas specification (3) performs the best, recall that this is an overfitted model. Substantial overfitting [specification (5)] performs only modestly worse than the correct specification [specification (1)] in terms of bias, and performs better according to MSE.

Panels V and VI present the results for the linear and quadratic specifications where the X's are correlated and only two regressors are relevant for determining the outcome. Note that because X_3 is irrelevant in this case, there is no correct specification, because all specifications include an irrelevant regressor. Nonetheless, specifications (1) and (3), in the linear and quadratic cases, remain closest to the true specification without excluding any relevant variables.

In the linear case (shown in panel V), the MSE is lower in all overfitted models relative to the "correct" specification [specifi-

Table 3. Monte Carlo results: weighting estimators with multiple covariates

	Specification of the propensity score					
	(0)	(1)	(2)	(3)	(4)	(5)
	Trim = [0.02, 0.98]					
I. Linear propensity score: all covariates correlated & relevant						
1000 × Bias	−946.90	−5.07	−7.31	−7.10	−3.14	−1.22
Bias (normalized)	−946.90	−5.63	−6.75	−6.53	−3.33	−1.81
MSE	901.40	8.16	6.25	5.99	5.13	5.12
MSE (normalized)	901.40	6.02	5.07	4.94	4.36	4.34
II. Quadratic propensity score: all covariates correlated & relevant						
1000 × Bias	−1041.11	598.72	127.93	1.75	5.40	5.97
Bias (normalized)	−1041.11	400.13	77.81	1.24	4.29	4.85
MSE	1092.86	393.28	23.82	5.45	5.28	5.18
MSE (normalized)	1092.86	174.31	10.99	4.46	4.58	4.56
III. Liner propensity score: all covariates independent & relevant						
1000 × Bias	−626.24	1.23	0.74	−0.28	0.84	1.62
Bias (normalized)	−626.24	0.80	0.56	−0.15	0.70	1.21
MSE	395.83	4.48	3.72	3.51	3.81	3.90
MSE (normalized)	395.83	3.52	3.14	3.04	3.22	3.26
IV. Quadratic propensity score: all covariates independent & relevant						
1000 × Bias	−658.51	163.17	276.45	3.13	3.28	2.10
Bias (normalized)	−658.51	117.55	191.38	1.74	1.73	1.05
MSE	439.72	37.38	91.28	4.29	4.50	4.36
MSE (normalized)	439.72	20.49	44.59	3.46	3.62	3.63
V. Linear propensity score: all covariates correlated, not all covariates relevant						
1000 × Bias	−941.82	1.06	0.97	0.49	3.70	3.74
Bias (normalized)	−941.82	0.40	0.87	0.50	3.03	3.20
MSE	891.87	8.04	6.00	5.88	4.91	5.21
MSE (normalized)	891.87	5.87	4.85	4.79	4.16	4.38
VI. Quadratic propensity score: all covariates correlated, not all covariates relevant						
1000 × Bias	−1039.57	579.20	125.31	1.70	4.26	2.78
Bias (normalized)	−1039.57	386.50	74.90	−0.85	2.07	1.45
MSE	1089.97	371.30	22.79	5.09	4.74	4.63
MSE (normalized)	1089.97	164.24	10.36	4.11	4.09	4.09
VII. Linear propensity score: all covariates independent, not all covariates relevant						
1000 × Bias	−622.35	0.89	0.63	−0.07	0.33	0.38
Bias (normalized)	−622.35	0.38	0.41	−0.09	0.15	0.20
MSE	391.13	4.48	3.73	3.55	3.65	3.91
MSE (normalized)	391.13	3.56	3.20	3.12	3.18	3.31
VIII. Quadratic propensity score: all covariates independent, not all covariates relevant						
1000 × Bias	−661.54	165.12	268.66	5.44	5.31	4.10
Bias (normalized)	−661.54	117.90	184.37	2.93	2.88	2.12
MSE	443.70	38.03	87.17	4.14	4.25	3.95
MSE (normalized)	443.70	20.50	41.77	3.29	3.46	3.36

NOTE: Specification (0) is a zero-degree polynomial. Specification (1) adds linear terms for x_1, x_2, and x_3 to the previous specification. Specification (2) adds first-order interactions to the previous specification. Specification (3) adds quadratic terms for each regressor to the previous specification. Specification (4) adds pairwise interactions between first- and second-order terms to the previous specification. Specification (5) adds cubic terms for each regressor to the previous specification.

cation (1)], with little penalty for substantially overfitting. Similar results hold in the quadratic case. The final two panels (VII and VIII) display the results for the linear and quadratic specifications where the X's are independent and only two regressors are relevant for determining the outcome. In both cases, there is no penalty for overfitting relative to the "correct" specification [specification (1) or (3)], but a substantial penalty for underfitting. Thus panels V–VIII indicate that the efficiency loss from adding higher-order terms and interactions involving an irrelevant regressor is either not severe or nonexistent, and dwarfs the cost of underfitting. Moreover, the normalized estimator continues to perform at least as well as, if not better than, the unnormalized estimator.

The final set of results is presented in Table 4. Here we use the same data simulated for panels I–IV in Table 3 (i.e., all X's are relevant for determining treatment assignment and out-

Table 4. Monte Carlo results: weighting estimators with multiple covariates but excluding a relevant covariate

	Specification of the propensity score						Specification of the propensity score					
	(0)	(1)	(2)	(3)	(4)	(5)	(0)	(1)	(2)	(3)	(4)	(5)
	Trim = [0.02, 0.98]						Trim = [0.33, 0.67]					
I. Linear propensity score: all covariates correlated & relevant												
1000 × Bias	−946.90	−197.51	−198.79	−198.89	−194.41	−187.96	−1264.90	−181.31	−180.42	−180.33	−186.30	−190.12
Bias (normalized)	−946.90	−196.62	−197.35	−197.35	−193.64	−187.93	−1264.90	−183.16	−183.05	−183.15	−187.63	−190.21
MSE	901.40	44.61	44.33	44.10	42.00	39.50	1599.97	45.58	45.26	45.00	46.39	47.63
MSE (normalized)	901.40	43.48	43.35	43.22	41.52	39.31	1599.97	43.68	44.18	44.34	45.49	46.23
II. Quadratic propensity score: all covarates correlated & relevant												
1000 × Bias	−1041.11	206.08	24.30	−185.19	−183.78	−185.47	−1264.90	−1283.93	−1245.79	−1079.66	−1070.73	−1071.43
Bias (normalized)	−1041.11	102.23	−27.64	−186.53	−185.40	−185.60	−1264.90	−1295.72	−1270.84	−1082.33	−1073.37	−1073.47
MSE	1092.86	61.18	9.63	40.50	39.88	40.27	1599.97	1662.84	1565.42	1182.27	1164.76	1167.11
MSE (normalized)	1092.86	21.32	7.74	40.65	40.22	40.22	1599.97	1692.63	1628.90	1188.01	1170.48	1171.52
III. Linear propensity score: all covariates independent & relevant												
1000 × Bias	−626.24	−235.32	−234.99	−235.61	−235.58	−235.60	−1264.90	−234.09	−234.25	−234.11	−234.05	−233.94
Bias (normalized)	−626.24	−253.37	−235.08	−235.53	−235.51	−235.53	−1264.90	−235.08	−235.81	−234.59	−234.40	−234.32
MSE	395.83	58.65	58.35	58.52	58.44	58.45	1599.97	62.39	62.54	62.63	62.46	62.40
MSE (normalized)	395.83	58.50	58.29	58.43	58.38	58.39	1599.97	61.82	61.95	62.06	61.94	61.93
IV. Quardratic propensity score: all covariates independent & relevant												
1000 × Bias	−658.51	−143.69	−108.71	−255.48	−255.66	−256.01	−1264.90	−1032.90	−1000.39	−805.98	−802.00	−797.01
Bias (normalized)	−658.51	−167.93	−143.34	−256.11	−256.12	−256.35	−1264.90	−1030.67	−996.08	−805.09	−802.09	−797.06
MSE	439.72	27.29	19.38	70.04	70.17	70.29	1599.97	1078.58	1014.57	660.86	655.63	648.03
MSE (normalized)	439.72	33.80	26.48	70.24	70.28	70.41	1599.97	1074.36	1006.42	660.64	655.60	647.92

NOTE: Specification (0) is a zero-degree polynomial. Specification (1) adds linear terms for x_1 and x_2 to the previous specification. Specification (2) adds first-order interactions to the previous specification. Specification (3) adds quadratic terms for each regressor to the previous specification. Specification (4) adds pairwise interactions between first- and second-order terms to the previous specification. Specification (5) adds cubic terms for each regressor to the previous specification.

come), but we exclude X_3 in specifications (1)–(5) used to estimate the propensity score. Thus all specifications are misspecified, although we are still underfitting and overfitting with respect to the included covariates. We present two sets of results, based on different trimming levels. In the first set of results, we continue to trim observations with propensity scores outside the interval 0.02–0.98. In the second set, we follow the suggestion of Black and Smith (2004) and focus on observations with propensity scores closer to 0.5. As shown by Black and Smith (2004), the bias for the ATE on the treated (ATT) due to *selection on unobservables* (which is the case here when X_3 is ignored) is minimized when the propensity score is 0.5. This result follows from the assumptions of joint normality between unobservables impacting treatment assignment and unobservables affecting (potential) outcomes in the absence of treatment and additive separability between observable and unobservable determinants of potential outcomes. However, we can show that the bias for the ATE is equal to the bias for the ATT plus an additional term given by 1 minus the propensity score times the difference in the expected unobserved individual-specific gains to treatment across the treatment and control groups. Simulations reveal that the bias of the ATE due to selection on unobservables is minimized when the propensity score is above (below) 0.5 when the correlation between unobservables impacting the potential outcomes in the absence of treatment and unobservables affecting treatment assignment is positive (negative). Moreover, the distance between the "optimal" propensity score and 0.5 is increasing (in absolute value) with the correlation between unobserved individual-specific gains to treatment and unobservables affecting treatment assignment (assuming positive selection on unobserved gains). In light of this ambiguity, for comparison, we present results using trimming levels of 0.33 and 0.67, although we tried other trimming levels as well (results not shown).

The results are interesting. In the linear cases (panels I and III), there is no penalty for overfitting with either trimming level; the results are essentially unchanged across specifications (1)–(5). Moreover, the trimming level does not impact the bias of either estimator and affects the MSEs only modestly. Thus in the linear case, overfitting does not adversely affect the quality of the causal estimators when a relevant regressor has been excluded; however, overfitting does not help alleviate the bias due to selection on unobservables even when the unobservable is correlated with the observables.

Our findings differ in the quadratic cases (panel II and IV). When trimming at 0.02 and 0.98, the bias and MSE are minimized for both estimators in specification (2), which corresponds to an underfitted model; however, overfitting performs no worse than the "correct" model [specification (3)]. When trimming at 0.33 and 0.67, two changes occur. On one hand, the bias and MSE increase significantly in all specifications; on the other hand, gains result from overfitting, with both bias and MSE marginally lower in overfitted specifications relative to the "correct" specification [specification (3)]. Moreover, the benefits of underfitting in the quadratic case disappear when trimming at 0.33 and 0.67. Finally, consistent with the linear case, we see that overfitting does not help alleviate the bias due to selection on unobservables even when the unobservable is correlated with the observables.

It is not clear why underfitting should be superior when a relevant regressor is excluded from the model, and only in the quadratic case. Note, however, that when we trim at 0.33 and 0.67, we are focusing on the space where the propensity score is essentially linear. Thus it is not surprising, in light of the linear results given in panels I and III, that the benefits of underfitting disappear in the quadratic case when we trim small and large values of the propensity score. Nonetheless, we are left with a puzzle concerning the potential benefits of underfitting in cases when there is selection on unobservables and the propensity score is nonmonotonic.

In the end, however, the results from the additional experiments confirm our conclusion drawn from the baseline experiments that applied researchers should provide a series of estimates using increasingly sophisticated specifications of the propensity score model, regardless of the number of covariates or the use of matching or weighting estimators. In addition, of the weighting estimators, the normalized estimator is preferred. Finally, in terms of covariate selection, researchers should err on the side of including irrelevant regressors in the analysis, or at least conduct sensitivity analyses with new covariates as well as higher-order terms of included regressors. The only potential exception that we found to these guidelines is when the true propensity score is nonmonotonic and a relevant regressor is excluded from the estimation. But it is difficult to rationalize guidelines based on this situation, because here overfitting performed no worse than the model specified "correctly" in terms of the relevant regressors included in the estimation.

We now illustrate these guidelines with two applications involving trade-related policies.

4. APPLICATIONS

4.1 WTO and Environment

Copeland and Taylor (2004, p. 7) stated that "for the last ten years environmentalists and the trade policy community have engaged in a heated debate over the environmental consequences of liberalized trade." The debate becomes even more heated when focusing on the WTO, given the complexity of the relationship as well as the lack of empirical evidence on the environmental effects of the WTO. Specifically, there are a number of avenues by which the WTO may affect the environment. First, the WTO *may* increase the volume of trade, and the increase in trade *may* help or harm the environment. For instance, Rose (2004a, 2004b) reported surprisingly little significant association between the WTO and trade policy and only moderate evidence of expanded trade volumes. Moreover, even if the WTO does liberalize trade, the impact of trade on the environment is not clear. Antweiler, Copeland, and Taylor (2001), Frankel and Rose (2005), and others have found little evidence of a detrimental effect of trade on various measures of environmental quality (see Copeland and Taylor 2004 for a review).

Second, several provisions in the WTO *may* impede a country's ability to unilaterally enact trade-related measures designed to protect the environment (Bernasconi-Osterwalder et al. 2006). The pillars of the WTO framework—the principles of most favored nation (MFN) and national treatment—require that any advantage extended to one WTO member country be extended to "like products" from all member countries, and that imported goods be treated no less favorably by internal taxes and domestic regulations than "like" domestic products. Of relevance to environmental policy, WTO dispute settlement rulings typically fail to consider differences in environmental damage in determining likeness. Thus countries have been unable to discriminate against production methods that unduly harm the environment (Althammer and Dröge 2003). That said, there are exceptions, and these exceptions have recently been used to uphold environmental protections. For example, Article XX allows for a number of exceptions, including trade restrictions "necessary to protect human, animal, or plant life" or "relating to the conservation of exhaustible natural resources." But exceptions are limited by the chapeau of Article XX, requiring that trade measures not constitute "a means of arbitrary or unjustifiable discrimination between countries where the same conditions prevail, or a disguised restriction on international trade." In practice, before the U.S.–Shrimp/Turtle 21.5 (2001) dispute resolution, whereby a revised U.S. ban on shrimp imports from countries not adequately protecting against the accidental killing of sea turtles in the harvesting of shrimp was upheld after a previous U.S. ban had been determined to violate the chapeau, environmental policies focused on nonproduct-related PPMs were deemed inconsistent with the WTO (Bernasconi-Osterwalder et al. 2006).

Third, the aforementioned WTO provisions *may* impede the efficacy and/or viability of multilateral environmental agreements (MEAs) by preventing punishment for noncompliance or free-riding (Althammer and Dröge 2003). Finally, the WTO *may* impact the environment by providing a framework to handle such issues as a "race-to-the-bottom" or "regulatory chill" in environmental (or labor) standards (Bagwell and Staiger 2001a, 2001b). In sum, the impact of the General Agreement on Tariffs and Trade (GATT)/WTO on the environment is a priori ambiguous and merits empirical analysis.

To proceed, we used country-level data from Frankel and Rose (2005); we provide only limited details here. (We are grateful to Andrew Rose for posting the data at *http://faculty.haas.berkeley.edu/arose/*.) We analyzed five measures of environmental quality: per capita CO_2 emissions, average annual deforestation rate for 1990–1996, energy depletion, rural access to clean water, and urban access to clean water. We used three covariates in the first-stage propensity score equation, following Frankel and Rose (2005): real per capita GDP, land area per capita, and a measure of the democratic structure of the government. Finally, we supplemented the data with an indicator of whether the country is a GATT/WTO member.

In the analysis, we used observations from 1990 (before the WTO) and 1995 (after creation of the WTO). But because the treatment is defined as GATT/WTO membership, some observations were treated in both time periods and some were treated only in the second period. The sample constitutes an unbalanced panel. Table 5 provides summary statistics and descriptions of the variables, and Table 6 presents the results. For each outcome, we estimated the same six specifications as used in the multiple-regressor Monte Carlo analysis summarized in Tables 3 and 4. We also included a time dummy in specifications (1)–(5), but the results were qualitatively unchanged (data available on request). Within each specification, we display the unnormalized and normalized treatment effect estimate, as well

Table 5. Summary statistics

Variable	Mean	Standard deviation	N	Description
Per Capita CO_2	3.82	4.73	232	Carbon dioxide emissions, industrial, in metric tons per capita
Deforestation	0.68	1.28	223	Annual deforestation, average percentage change, 1990–1995
Energy depletion	3.13	7.43	223	In percent of GDP, equal to the product of unit resource rents and the physical quantities of fossil fuel energy extracted
Rural water access	51.2	27.42	137	Access to clean water, percentage of rural population, 1990–1996
Urban water access	76.28	21.76	140	Access to clean water, percentage of urban population, 1990–1996
GATT/WTO (1 = Yes)	0.78	0.41	232	Member country of GATT/WTO
Real GDP per capita	7,302.91	7,468.41	232	Real (1990) gross domestic product divided by population
Polity	3.17	6.85	232	Index, ranging from-10 (strongly autocratic) to 10 (strongly democratic)
Area per capita	51.60	89.56	232	Land area divided by population

NOTE: Source is Environmental indicators and country-level controls are from Frankel and Rose (2002, 2005); GATT/WTO membership data are from Rose (2004a, 2005). N = number of observations. Observations from 1990 and 1995. See *http://faculty.haas.berkeley.edu/arose/*.

as the standard error. We exclude observations with an estimated propensity score outside the interval [0.05, 0.95].

For per capita CO_2, the results indicate three salient findings. First, there are modest differences between the unnormalized and normalized estimates that widen as additional covariates are added to the model. But given the size of the standard errors, these differences do not appear to be statistically (or economically) meaningful. Second, although there are modest differences in the unnormalized estimates when moving from a linear specification of the propensity score to adding higher-order terms, there is little effect on the normalized estimates. Again, given the size of the standard errors, overfitting—even with the unnormalized estimator—does not qualitatively alter the results. Thus, although there is little benefit to adding additional terms in this case, there is no cost either; in fact, the standard error even improves from specification (1) to specification (4) and then returns to the original level in specification (5), which includes cubic terms. Finally, in terms of the actual point estimates, we find statistically significant evidence that GATT/WTO membership reduces per capita CO_2 emissions ($\tau = -1.28$; $\tau_{norm} = -1.15$; standard error = 0.61). The magnitude of the estimates indicate that GATT/WTO membership lowers emissions by roughly 0.25 standard deviation. This is contrary to the results of Frankel and Rose (2005); however, those authors' "treatment" is trade openness, rather than GATT/WTO membership, where openness is treated as endogenous. This result is also contrary to the argument that the GATT/WTO framework may explicitly or implicitly deter international cooperation because the damage from CO_2 emissions represent a pure global externality; however, it is consistent with the arguments of Bagwell and Staiger (2001a, 2001b), as well as with the results on energy depletion that we present later.

Table 6. Impact of GATT/WTO membership on environmental quality

	Specification of the propensity score					
	(0)	(1)	(2)	(3)	(4)	(5)
I. Per capita carbon dioxide						
τ	1.56	−1.16	−1.06	−1.18	−1.21	−1.28
τ_{norm}	1.56	−1.15	−1.02	−1.08	−1.11	−1.15
Standard error	0.58	0.61	0.50	0.41	0.41	0.61
II. Deforestation						
τ	−0.03	0.21	0.36	0.37	0.41	0.39
τ_{norm}	−0.03	0.21	0.34	0.32	0.40	0.40
Standard error	0.02	0.20	0.20	0.18	0.19	0.21
III. Energy depletion						
τ	−4.57	−3.54	−3.15	−2.73	−2.39	−2.32
τ_{norm}	−4.57	−3.47	−3.33	−3.12	−2.50	−2.34
Standard error	1.54	1.34	1.33	1.33	1.32	1.38
IV. Rural water access						
τ	8.85	3.10	2.53	2.64	3.17	0.027
τ_{norm}	8.85	3.40	2.91	3.24	1.52	0.97
Standard error	5.09	5.05	5.08	6.04	4.88	28.20
V. Urban water access						
τ	2.14	−6.20	−4.93	−2.02	0.55	−1.17
τ_{norm}	2.14	−5.02	−4.42	−1.59	−1.19	−0.61
Standard error	4.54	5.15	5.25	10.15	6.00	44.36

NOTE: See Table 3 for definition of the different propensity score specifications. Covariates are real GDP, area per capita, and a measure of polity. Propensity scores trimmed at 0.05 and 0.95. See text for further details.

For the next measure of environmental quality, deforestation, there are several instances of the interplay between the GATT/WTO and environmental quality, including two prime examples. In the dispute *Japan—Tariff on Imports of Spruce, Pine, and Fir* (*SPF*) *Dimension Lumber*, a Japanese tariff that applied to specific types of dimension lumber was found to be permissible under the GATT. More recently, Canada has disputed U.S. policies designed to protect the U.S. lumber industry from allegedly subsidized Canadian lumber in *U.S.—Softwood Lumber*. Recently, the WTO Appellate Body has upheld U.S. antidumping rates, but a WTO panel also held that the U.S. countervailing duty determination does not conform to WTO requirements.

Here we find that the choice between the normalized and unnormalized estimators has even less impact than for per capita CO_2, but the exact specification of the propensity score does have a substantial effect on inference and magnitude. Specifically, the linear propensity score equation [specification (1)] yields statistically insignificant effects of GATT/WTO membership on deforestation ($\tau = 0.21$; $\tau_{norm} = 0.21$; standard error = 0.20). The most flexible specification [specification (5)] yields statistically significant estimates that are twice as large in magnitude ($\tau = 0.39$; $\tau_{norm} = 0.40$; standard error = 0.21). Thus

we conclude that there is statistically meaningful evidence that GATT/WTO membership accentuates the annual rate of deforestation, and that this impact is economically meaningful, representing an increase of roughly 0.33 standard deviation (or nearly 60% for the mean country in the sample). Interestingly, the changes in inference across the various specifications are due solely to a change in the estimate; the standard error remains essentially constant.

The next set of results pertains to energy depletion, with a larger value indicating greater energy consumption. As with the previous outcomes, we find modest differences at best between the two estimators. Moreover, as with deforestation, we do find economically meaningful differences in the point estimates across the various specifications, although, unlike in deforestation, here statistical significance is unaffected. Specifically, we find a highly statistically and economically significant effect of GATT/WTO membership on energy conservation in the linear specification [specification (1); $\tau = -3.54$; $\tau_{norm} = -3.47$; standard error $= 1.34$]. This constitutes a decline of nearly 0.5 standard deviation (or $>100\%$ for the mean country in the sample). In the most flexible specification [specification (5)], however, the point estimates drop by roughly one-third ($\tau = -2.32$; $\tau_{norm} = -2.34$; standard error $= 1.38$). Thus adding additional terms to the model indicates a statistically significant beneficial impact of GATT/WTO membership on energy use of more reasonable magnitude; membership reduces energy depletion by 0.33 standard deviation (or roughly 75% for the mean country in the sample), consonant with Frankel and Rose's (2005) analysis of trade openness, as well as our previous results using per capita CO_2 emissions. Finally, as with the previous two environmental measures, the standard error remains essentially constant across specifications, consonant with the Monte Carlo results.

In terms of rural and urban access to clean water, two findings are noteworthy. First, across both measures, both the normalized and unnormalized estimators, and all specifications, the impact of GATT/WTO membership is never statistically significant. Second, unlike with the previous three outcomes, the standard error of the estimates does vary, sometimes quite substantially, across the various specifications; it is nearly six (eight) times larger in specification (5) compared with specification (4). This may be attributable to the smaller sample of countries for which data on clean water access are available, combined with the fact that water access is time-invariant in 1990 and 1995 (see Table 5). Thus the relative lack of cost of adding additional terms depends crucially on the sample size and the variation in outcome.

In the end, we find some evidence that GATT/WTO membership is associated with improved environmental performance measured in terms of per capita CO_2 emissions and energy depletion, but higher rates of deforestation and no statistically significant change in access to clean water. In addition, the economic and/or statistical conclusions than can be drawn are dependent on the specification of the propensity score for some outcomes (deforestation and energy depletion), but not on the choice of weighting estimator.

4.2 Euro and Trade

Amother policy-related question in the international arena that has received even more attention from economists is the impact of a common currency (i.e., currency union) on the level of international trade. Intuitively, a currency union may lead to more trade by reducing exchange rate volatility or by reducing transaction costs. Using a lengthy panel of a large number of countries, Rose (2000) found that members of currency unions trade on average more than three times more with each other compared with noncurrency union countries. This conclusion has been criticized in several dimensions, however. First, the sample pooled a very heterogeneous mix of countries; second, the empirical model may have been misspecified, due to either functional form or the omission of relevant regressors.

To address these issues, more recent studies have focused on more homogeneous samples (Micco, Stein, and Ordonez 2003), as well as applied various propensity score matching estimators (Persson 2001; Chintrakarn 2008) using different matching techniques. Specifically, Micco, Stein, and Ordonez (2003) used a smaller sample of developed countries to assess the impact of Europe's monetary union (the euro), which came into existence in 1999, on bilateral trade using a parametric framework, and found a trade-inducing effect of between 4% and 16%. Bun and Klaassen (2002, 2007) and De Nardis and Vicarelli (2003) used more complex parametric approaches, including dynamic panel estimators, and found similar expansions in trade from adoption of the euro. Finally, Chintrakarn (2008) applied matching and difference-in-difference matching estimators to the data of Micco, Stein, and Ordonez (2003), and found an effect of 9%–14%.

To build on this previous work, we used the country-level data from Micco, Stein, and Ordonez (2003) and Chintrakarn (2008); we provide only limited details here. (We are grateful to Pandej Chintrakarn for making these data available to us.) The data include bilateral trade flows across 22 developed countries, or 231 country-pairs per year, for 1999–2002, yielding a total sample size of 924. The 22 countries include Australia, Austria, Belgium-Luxembourg, Canada, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Italy, Japan, The Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, the U.K., and the U.S. The treatment group, defined as both countries having adopted the euro, includes 200 observations. We analyzed two outcomes: (log) current bilateral trade and the change in (log) bilateral trade between the current period and average bilateral trade over the period 1994–1998. The differenced outcome follows from the literature and attempts to remove unobservables that may be correlated with euro adoption and current trade levels. Following Chintrakarn (2008), we used three continuous and three discrete covariates in the first-stage propensity score equation: (log) product of real per capita GDP, (log) product of land area, (log) distance, number of landlocked countries in the country-pair, a dummy variable for sharing a common language, and a dummy variable for sharing a border. We excluded time dummies, which were found to not matter by Chintrakarn (2008). Table 7 provides summary statistics and descriptions of the variables.

Table 8 presents the results. Panel I (II) displays the results using bilateral trade in levels (changes) as the outcome. As in

Table 7. Summary statistics

Variable	Mean	Standard deviation	N	Description
Log (bilateral trade)	6.95	2.02	924	Log of bilateral trade between country pairs in 1995 U.S. dollars
Change in log (bilateral trade)	0.03	0.22	924	Difference in the log of current bilateral trade and average bilateral trade from 1994–1998
Euro (1 = Yes)	0.22	0.41	924	Both countries formal member of the European Monetary Union (EMU)
Log (distance)	7.53	1.17	924	Log of the distance in miles between capital cities
Log (product real GDPs)	53.46	2.08	924	Log of the product of the real GDP of each country in 1995 U.S. dollars
Log (product land areas)	25.05	2.26	924	Log of the product of the land area in each country
Common language (1 = Yes)	0.10	0.29	924	Countries share the same language
Share border (1 = Yes)	0.09	0.28	924	Countries share a common border
Number of landlocked countries (0, 1, 2)	0.18	0.40	924	Number of countries in the pair that are landlocked

NOTE: Source: Chintrakarn (2008). Data are from 1999–2002. N represents the number of observations. See text for further details.

the previous application, we estimate the same six specifications as used in the multiple-regressor Monte Carlo analysis given in Tables 3 and 4 for each outcome. Within each specification, we display the unnormalized and normalized treatment effect estimate, as well as the standard error. We exclude observations with an estimated propensity score outside the interval $[0.02, 0.98]$, given the sample size of close to 1000. Panel I shows that the estimated treatment effect varies considerably across specifications and estimators, although most estimates are statistically insignificant (although large in magnitude). Moreover, some evidence of a large and statistically significant *negative* impact of euro adoption is seen in specifications (1) and (2) with the unnormalized estimator; however, both the normalized and unnormalized estimates diminish in absolute value and become statistically insignificant as the specification of the propensity score increases. Panel II shows little difference across specifications [including specification (0)] or estimators in terms of inference, although the magnitude of the point estimate does increase as the specifications become more flexible. Specifically, euro adoption increases bilateral trade by roughly 10%, consonant with the results of Chintrakarn (2008) and others. Thus, sharing a common currency does have an economically meaningful impact on trade when using a sample of relatively advanced and homogeneous countries.

Table 8. Impact of euro adoption on bilateral trade

	Specification of the propensity score					
	(0)	(1)	(2)	(3)	(4)	(5)
I. Log (bilateral trade)						
τ	1.18	−1.28	−0.74	−0.29	−0.18	−0.11
τ_{norm}	1.18	0.39	0.37	0.33	0.35	0.34
Standard error	0.14	0.24	0.30	0.26	0.26	0.23
II. Change in log (bilateral trade)						
τ	0.08	0.07	0.09	0.10	0.10	0.10
τ_{norm}	0.08	0.08	0.10	0.11	0.11	0.10
Standard error	0.02	0.01	0.01	0.02	0.02	0.02

NOTE: See Table 3 for definitions of the different propensity score specifications. Covariates are the log product of real GDPs, the log product of land areas, a dummy variable for sharing a common language, a dummy variable for sharing a border, and the number of landlocked countries in the pair. Higher-order terms and interactions are included only for the first three continuous variables. Propensity scores are trimmed at 0.02 and 0.98. See the test for further details.

Given that the estimates in panel I likely are biased due to the exclusion of relevant time-invariant variables, we reestimated the treatment effects in panel I, excluding observations with an estimated propensity score outside the interval $[0.33, 0.67]$. The results (not shown) indicate that the unnormalized estimates continue to be quite volatile across specifications and predominantly negative; however, the normalized estimator yields an estimate of 0.11 in specification (5), which is quite close to the estimate obtained in panel II. Nonetheless, we caution against drawing too strong of a conclusion from these results, because the Monte Carlo experiments demonstrate no difference in the performance of the two estimators in the presence of an entirely excluded relevant variable, and the normalized estimators also are fairly volatile across specifications when trimming at $[0.33, 0.67]$, although less so than the unnormalized estimator.

5. CONCLUSION

Although the use of propensity score methods in estimating treatment effects in nonexperimental settings is proliferating, little practical guidance exists for researchers on specifying the propensity score model. The perception appears to be that although underspecifying propensity scores will yield inconsistent estimates, overspecifying propensity scores is inefficient at best and inconsistent at worst. But this perception is not necessarily correct. Using a Monte Carlo study and two weighting estimators, we found little penalty for overfitting propensity scores, and in fact found numerous cases in which overspecifying the model proved beneficial. A secondary result of our analysis is that the normalized version of the Horvitz–Thompson estimator performs as well as, and sometimes better than, the original Horvitz–Thompson estimator. As a result, we recommend that researchers report a number of estimates corresponding to different levels of polynomials used to estimate propensity scores.

The intuition underlying our findings is as follows. Overspecifying the propensity score model has two effects: reducing the bias of the estimated propensity score and increasing the variance of the estimated propensity score. If the reduction in

bias dominates the increase in variance, then there should be no penalty for overfitting. This intuition is supported by findings of Ichimura and Linton (2005) that undersmoothing (equivalent to overfitting in our case) reduces the bias without increasing the variance when the propensity score is estimated using kernel smoothing methods (see their thm. 3.1). In addition, our results are consistent with the intuition of Hirano, Imbens, and Ridder (2003), who showed, using a simple example, that even if the true propensity score is constant, including covariates in the propensity score specification results in a more efficient estimator.

To illustrate our new recommended approach, we have addressed two important questions in the international arena: (a) Does the WTO hinder environmental improvement? and (b) do currency unions promote international trade? Using data from the 1990s, we found that the WTO is beneficial in terms of the environmental measures that are global in nature and less directly tied to the sale of products that are protected by the WTO (e.g., deforestation). Moreover, these results are sensitive to the specification of the propensity score model in some cases, although not to the choice of estimator, confirming the benefits of comparing the results across numerous propensity score specifications. Using data on bilateral trade flows across developed countries for 1994–2002, we found a robust, statistically significant impact of euro adoption on trade once time-invariant unobservables are removed. We found little impact of overfitting or choice of estimator.

APPENDIX: ASYMPTOTIC VARIANCE BOUND

From theorem 1 of Hirano, Imbens, and Ridder (2003),

$$V = \mathrm{E}\left[(\tau(X) - \tau)^2 + \frac{\sigma_1^2(X)}{p(X)} + \frac{\sigma_0^2(X)}{1 - p(X)}\right],$$

where $p(X)$ is the true propensity score evaluated at X, $\tau(X) = \mathrm{E}[Y(1) - Y(0)|X]$, and

$$\sigma_j^2(X) = V[Y(j)|X], \qquad j = 0, 1.$$

Our data-generating process is

$$Y_i = \beta_1 + \beta_2 T_i + \beta_3 X_i + \beta_4 T_i X_i + e_i,$$

where $X \sim \mathrm{U}[0, 1]$ and $e_i \sim \mathrm{N}(0, \sigma^2)$. Thus

$$Y_i(0) = \beta_1 + \beta_3 X_i + e_i$$

and

$$Y_i(1) = Y_i(0) + \beta_2 + \beta_4 X_i.$$

Therefore,

$$\sigma_0^2(x) = \sigma_1^2(x) = \sigma^2,$$
$$\tau(X) = \mathrm{E}[Y(1) - Y(0)|X] = \beta_2 + \beta_4 X,$$

and

$$\tau = \mathrm{E}[Y(1) - Y(0)] = \beta_2 + \beta_4 \mathrm{E}[X] = \beta_2 + 0.5\beta_4.$$

From the foregoing expressions, we have

$$\mathrm{E}(\tau - \tau(X))^2 = \mathrm{E}(\beta_4(X - 0.5))^2$$
$$= \beta_4^2 \mathrm{E}(X - 0.5)^2 = \beta_4^2 V(X) = \beta_4^2/12.$$

Thus,

$$V = \frac{\beta_4^2}{12} + \sigma^2 \mathrm{E}\left[\frac{1}{p(X)}\right] + \sigma^2 \mathrm{E}\left[\frac{1}{1 - p(X)}\right].$$

ACKNOWLEDGMENTS

The authors thank the editor, an anonymous associate editor, and two anonymous referees for detailed suggestions that greatly improved the article. They also thank Keisuke Hirano, Alberto Abadie, Juan Carlos Escanciano, David Jacho-Chavez, Pravin Trivedi, Konstantin Tyurin, and participants at the Midwest Econometrics Group meetings in October 2006 for helpful comments and discussions.

[Received August 2006. Revised November 2007.]

REFERENCES

Abadie, A. (2005), "Semiparametric Difference-in-Differences Estimators," *The Review of Economic Studies*, 72, 1–19.

Althammer, W., and Dröge, S. (2003), "International Trade and the Environment: The Real Conflicts," in *Environmental Policy in an International Perspective*, eds. L. Marsiliani, M. Rauscher, and C. Withagen, Dordrecht, The Netherlands: Kluwer Academic Publishers.

Antweiler, W., Copeland, B. R., and Taylor, M. S. (2001), "Is Free Trade Good for the Environment," *The American Economic Review*, 91, 877–908.

Bagwell, K., and Staiger, R. W. (2001a), "Domestic Policies, National Sovereignty and International Economic Institutions," *The Quarterly Journal of Economics*, 116, 519–562.

——— (2001b), "The WTO as a Mechanism for Securing Market Access Property Rights: Implications for Global Labor and Environmental Issues," *The Journal of Economic Perspectives*, 15, 69–88.

Bernasconi-Osterwalder, N., Magraw, D., Oliva, M. J., Orellana, M., and Tuerk, E. (2006), *Environment and Trade: A Guide to WTO Jurisprudence*, London: Earthscan.

Black, D. A., and Smith, J. A. (2004), "How Robust Is the Evidence on the Effects of College Quality? Evidence From Matching," *Journal of Econometrics*, 121, 99–124.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006), "Variable Selection for Propensity Score Models," *American Journal of Epidemiology*, 163, 1149–1156.

Bryson, A., Dorsett, R., and Purdon, S. (2002), "The Use of Propensity Score Matching in the Evaluation of Labour Market Policies," Working Paper 4, Dept. for Work and Pensions, London, England.

Bun, M. J. G., and Klaassen, F. J. G. M. (2002), "Has the Euro Increased Trade?" Discussion Paper 02-108/2, Tinbergen Institute.

——— (2007), "The Euro Effect on Trade Is Not as Large as Commonly Thought," *Oxford Bulletin of Economics and Statistics*, 69, 473–496.

Chintrakarn, P. (2008), "Estimating the Euro Effects on Trade With Propensity Score Matching," *Review of International Economics*, 16, 186–198.

Copeland, B. R., and Taylor, M. S. (2004), "Trade, Growth, and the Environment," *Journal of Economic Literature*, 42, 7–71.

D'Agostino, R. B., Jr. (1998), "Tutorial in Biostatistics: Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group," *Statistics in Medicine*, 17, 2265–2281.

De Nardis, S., and Vicarelli, C. (2003), "The Impact of the Euro on Trade: The (Early) Effect Is Not So Large," Working Paper 017, European Network of Economic Policy Research Institutes.

Dehejia, R. H., and Wahba, S. (1999), "Casual Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.

Fackler, P. L. (2007), "Generating Correlated Multidimensional Variates," unpublished manuscript, North Carolina State University.

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.

Frankel, J. A., and Romer, D. (1999), "Does Trade Cause Growth," *The American Economic Review*, 89, 379–399.

Frankel, J. A., and Rose, A. K. (2002), "Is Trade Good or Bad for the Environment? Sorting Out the Causality," Working Paper 9201, NBER.

——— (2005), "Is Trade Good or Bad for the Environment? Sorting Out the Causality," *The Review of Economics and Statistics*, 87, 85–91.

Geman, S., and Hwang, C. (1982), "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," *The Annals of Statistics*, 10, 401–414.

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

Heckman, J. J., and Robb, R. (1985), "Alternative Methods for Evaluating the Impact of Intervientions," in *Longitudinal Analysis of Labor Market Data*, eds. J. J. Heckman and B. Singer, Cambridge, England: Cambridge University Press.

Hirano, K., and Imbens, G. W. (2001), "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization," *Health Services and Outcomes Research Methodology*, 2, 259–278.

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

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685.

Ichimura, H., and Linton, O. (2005), "Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators," in *Identification and Inference for Econometric Models: Essays in Honour of Thomas Rothenberg*, eds. D. W. K. Andrews and J. Stock, New York: Cambridge University Press.

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

Imbens, G. W., Newey, W., and Ridder, G. (2005), "Mean-Square-Error Calculations for Average Treatment Effects," IEPR Working Paper 05-34, University of Southern California.

LaLonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs With Experimental Data," *The American Economic Review*, 76, 604–620.

Micco, A., Stein, E., and Ordonez, G. (2003), "The Currency Union Effect on Trade: Early Evidence From EMU," *Economic Policy*, 18, 315–356.

Newey, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," translated in *Statistical Science* (with discussion), 5 (1990), 465–480.

Newey, W. (1994), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Persson, T. (2001), "Currency Union and Trade: How Large Is the Treatment Effect," *Economic Policy*, 33, 433–462.

Robins, J., and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129.

Rose, A. K. (2000), "One Money, One Market: Estimating the Effect of Common Currencies on Trade," *Economic Policy*, 30, 7–46.

——— (2004a), "Do We Really Know That the WTO Increases Trade," *The American Economic Review*, 94, 98–114.

——— (2004b), "Do WTO Members Have More Liberal Trade Policy," *Journal of International Economics*, 63, 209–235.

——— (2005), "Which International Institutions Promote International Trade?" *Review of International Economics*, 13, 682–698.

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

Roy, A. D. (1951), "Some Thoughts on the Distribution of Income," *Oxford Economic Papers*, 3, 135–146.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Rubin, D., and Thomas, N. (1996), "Matching Using Estimated Propensity Scores: Relating Theory to Practice," *Biometrics*, 52, 249–264.

Shaikh, A. M., Simonsen, M., Vytlacil, E. J., and Yildiz, N. (2005), "On the Identification of Misspecified Propensity Scores," unpublished manuscript, Columbia University, Dept. of Economics.

Smith, J. A., and Todd, P. E. (2005), "Does Matching Overcome LaLonde's Critique," *Journal of Econometrics*, 125, 305–353.

Todd, P. E. (1996), "Three Essays on Empirical Methods for Evaluating the Impact of Policy Interventions in Education and Training," unpublished Ph.D. dissertation, University of Chicago, Dept. of Economics.

Zhao, Z. (2008), "Sensitivity of Propensity Score Methods to the Specifications," *Economics Letters*, 98, 309–319.