CAEPR Working Paper
#2006-013

On the Specification of Propensity Scores: with an Application to the WTO-Environment Debate

Daniel Millimet
Southern Methodist University

Rusty Tchernis
Indiana University Bloomington

September 29, 2006

On the Specification of Propensity Scores: with an Application to the WTO-Environment Debate

Daniel L. Millimet*
Southern Methodist University

Rusty Tchernis
Indiana University

August 2006

Abstract

The use of propensity score methods for program evaluation with non-experimental data typically requires the propensity score be estimated, often with a model whose specification is unknown. While theoretical results suggest that estimators utilizing more flexible propensity score specifications perform better, this has not filtered into applied research. Here, we provide Monte Carlo evidence indicating the benefits of *over-specifying* the propensity score when using weighting estimators, as well as using *normalized* weights. We illustrate these results with an application assessing the environmental effects of GATT/WTO membership. We find that membership has a mixed impact, and that under-fitting the propensity score yields misleading inference in several cases.

JEL Classifications: C21, C52, F18
Key words: Treatment Effects, Propensity score, Specification, WTO, Environment

*The authors are indebted to Keisuke Hirano for helpful comments and discussions. Corresponding author: Daniel Millimet, Department of Economics, Southern Methodist University, Box 0496, Dallas, TX 75275-0496, USA; Email: millimet@mail.smu.edu; Tel. (214) 768-3269; Fax: (214) 768-1821.

1 Introduction

Estimation methods utilizing the propensity score (i.e., the probability of an observation receiving a particular treatment conditional on covariates) are widely used in economics, as well as other disciplines, in the evaluation of programs or interventions. In the majority of applications, the *true* propensity score is unknown, and therefore must be estimated. Even in the case when the true propensity score is known, Hahn (1998) and Hirano et al. (2003) show that using the true propensity score is inefficient. However, there are few guidelines for applied researchers on how to proceed in the estimation of the propensity score. In particular, there are two inter-related specification issues that arise when estimating the propensity score in practice. First, one must decide which variables are to be included in the propensity score model. Second, one must decide if higher order terms (and/or interaction terms) are to be included as well. A third specification issue, the choice of the actual model itself (e.g., probit, logit, nonparametric, etc.), is not the focus of the present paper.

As noted in Smith and Todd (2005) and the references therein, using too crude of a propensity score specification is likely to yield biased estimates of the causal effect of the treatment. However, the inclusion of irrelevant variables in the propensity score model may have a similar impact. Bryson et al. (2002) argue against the inclusion of irrelevant variables on efficiency grounds; the variance of the treatment effect estimator is likely to increase. Brookhart et al. (2006) suggest 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 as their inclusion results in higher mean squared error of the treatment effect estimate. Zhao (2005, p.15) argues that while including irrelevant variables will not bias estimates of the treatment effects, over-fitting the propensity score model (i.e., including irrelevant higher order terms) may be "counterproductive," although "in practice, this point may not be important." Rubin and Thomas (1996) argue in favor of including variables in the propensity score model unless there is consensus that they do not belong. Lastly, the intuition behind the result in Hirano et al. (2003) that using the true propensity score is inefficient even when it is known suggests that over-fitting the propensity score model may have little negative consequence on the properties of treatment effect estimates in practice.

To date, 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 means to implement matching or stratification estimators. If a variable is found not to be balanced, the usual approach is re-specify the propensity score model adding (additional) higher order terms and/or interactions (e.g. Dehejia and Wahba 1999). However, there is no single, universally accepted balancing

test (Smith and Todd 2005). Recently, Shaikh et al. (2005) develop an alternative test for misspecification of the propensity score, although investigating the consequences of mis-specifying the propensity score due to over-fitting is not a goal of the paper.

In light of this background, the aim of the present paper is twofold. First, we assess whether the practical effects of mis-specifying the propensity score model through the *exclusion* of relevant higher order terms and the *inclusion* of irrelevant higher order terms via a Monte Carlo study. Our study focuses 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). The results indicate that in many cases over-fitting the propensity score model results in a more efficient estimator, and in the remaining cases does no worse than the correctly specified model.[1] 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. Since the penalty for over-fitting is minimal, this suggests that practitioners report a set of causal estimates corresponding to various models of propensity scores.

The second goal of this study is to illustrate these ideas with a timely application related to the impact of membership in the World Trade Organization (WTO) on environmental quality. Using cross-country data from the 1990s, we find a detrimental impact of WTO membership on per capita carbon dioxide emissions, and a beneficial impact of membership on energy consumption and access to clean water. The unfavorable effect of the WTO on per capita carbon dioxide emissions is consonant with previous research suggesting that the WTO may impede international cooperation on global environmental issues, while the positive effect on local measures of environmental quality is consistent with previous research on the beneficial effects of trade openness. In addition, the application illustrates the relative costlessness of over-fitting the propensity score model.

The remainder of the paper is as follows. Section 2 presents a brief review of the treatment effects setup as well as the estimators analyzed. Section 3 describes the Monte Carlo study and results. Section 4 contains the WTO-environment application. Section 5 concludes.

2 Setup

Consider a random sample of N individuals from a large population indexed by $i = 1, ..., N$. As in the now commonplace 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, $t \in \mathcal{T}$. Here, we consider only the

[1] Similar results are found in Ichimura and Linton (2005), who use a kernel estimator to estimate propensity scores.

case of binary treatments: $\mathcal{T} = \{0, 1\}$. The causal effect of one treatment, say $t = 1$, relative to the other, $t = 0$, is defined as the difference between the corresponding potential outcomes. Formally,

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

and the population average treatment effect 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 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 makes clear that only one potential outcome is observed for any individual. As such, estimating τ is not trivial as there is an inherent missing data problem, and some assumptions are required to proceed.

One such assumption is *unconfoundedness* or selection on observables (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) propose using the propensity score, $P(x) = \Pr(T_i = 1 | X_i = x)$, 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) or Smith and Todd (2005)), the average treatment effect can be estimated in a number of ways (D'Agostino (1998) and Imbens (2004) offer summaries). Here, we focus on the inverse probability weighted estimator of Horvitz and Thompson (1952), given by

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

We refer to (4) as the unnormalized estimator. A sample estimate of τ can be computed using estimated propensity scores as follows:

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

Alternatively, Hirano and Imbens (2001) propose 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

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

The advantage of estimator in (6), which we refer to as the normalized estimator, is that the weights sum to unity within each group (Imbens et al. 2005). In addition, Imbens et al. (2005) show that the normalized estimator is asymptotically equivalent to the unnormalized estimator and therefore is asymptotically efficient.

This exposition shows that researchers face two issues in practice. First, since the propensity score is typically unknown, this 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 et al. 2003). Second, practitioners must decide between the unnormalized Horvitz and Thompson (1952) estimator and the normalized Hirano and Imbens (2001) estimator. Hirano et al. (2003) show 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 propose a feasible variance estimator.

To help inform applied researchers, we assess the practical performance of the two estimators, (5) and (6), as well as various specifications of the propensity score. With respect to the latter, we focus in particular on whether it pays to over-specify the propensity score equation. We now turn to our Monte Carlo study.

3 Monte Carlo Study

To compare the two estimators, as well as assess the benefit to over-specifying the propensity score equation, we simulate the treatment assignment based on a number of different, and increasingly difficult for estimation, *true* models, and then compare the performance of the estimators in (5) and (6). For each of the *true* models, we simulate 1,000 data sets. Each data set contains three variables for each of the 1,000 observations: T_i, a binary indicator of the treatment received, Y_i, the outcome, and a single covariate, X_i. X is randomly drawn from a $U[0,1]$ distribution. Simulation of the treatment assignment and outcome are discussed below.

3.1 Treatment Assignment

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. The propensity scores are simulated

from nine distinct settings, ranging from continuous, smooth and monotone to non-continuous and non-monotone. Thus, we compare the performance of the causal estimators using models for which SLE would be expected to perform well (e.g. when propensity scores are simulated using a logit model), as well as more difficult models (e.g. non-continuous propensity scores). The nine settings are:

1. Logit specification:

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

   where A_i is

   (a) Flat: $A_i = 0$

   (b) Linear: $A_i = -3 + 6X_i$

   (c) Quadratic & Symmetric: $A_i = 2.5 - (2.5(1 - 2X_i))^2$

   (d) Quadratic & Non-Symmetric: $A_i = 2.5 - (2.5(1 - 1.5X_i))^2$

   (e) Fourth degree: $A_i = -2.5 + 4.5X_i^4$

2. 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) Non-Symmetric:

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

3. 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) Non-Monotonic:

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

While the relationships between the propensity scores and covariate are continuous and smooth in all the logit specifications, Specifications 1a, 1b, and 1e represent monotonic relationships and Specifications 1c and 1d are non-monotonic. In addition, Specification 1e is also asymmetric. Specifications 2a and 2b, on the other hand, represent non-smooth, but continuous, relationships, while Specifications 3a and 3b represent non-continuous relationships. The propensity scores are presented in Figure 1. Note that all specifications satisfy the common support condition.

3.2 Outcome Equation

Outcomes are simulated using the following specification:

$$Y_i = \beta_1 + \beta_2 T_i + \beta_3 X_i + \beta_4 T_i X_i + \varepsilon_i, \tag{7}$$

where $\varepsilon_i \sim N(0, \sigma^2)$. To decide on the values of the parameters, we use the data from Lalonde (1986). Specifically, we estimate (7) using this data, where X represents the first principal component of the set of the covariates used in Deheija and Wahba (1999). The values of $\beta = \{\beta_j\}_{j=1}^4$ are [2.54, -1.67, -2.47, 1.96], and $\sigma^2 = 0.25$.

3.3 Estimation and Evaluation

3.3.1 Estimation

For each data set, we estimate the propensity scores using a SLE. Following Hirano et al. (2003), we estimate the propensity scores as

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

where $R^K(X)$ is a vector of length K with the j^{th} 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 $K = 1, \ldots, 10$. Hence, we estimate ten logit models with increasing degree of polynomial in the logit function.

The variance is computed as in Hirano et al. (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

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

$$\times \left(\frac{1}{N} \sum_{i=1}^{N} R^K(X_i) R^K(X_i)' \right)^{-1} R^K(X_i)(T_i - \hat{P}(X_i)).$$

3.3.2 Evaluation

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)

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

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, respectively, evaluated at a grid of M points in $[0, 1]$ interval (M equals 100 in the study). Lower values of both measures represent a better fit. Finally, we report the semiparametric efficiency bound for each true model; see the appendix for derivation.

3.4 Results

The results are presented in Table 1. Panel I contains the results using Specification 1a for the propensity score. Here, the true propensity score is flat, implying treatment is random (independent of X). This specification provides a nice check as 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 variance are higher than in over-fitted models. As a result, the coverage rate is above 95% in the first column, and above the rest of the estimated models. This result is closely related to the findings in Hirano et al. (2003), since the estimated propensity score using zero degree polynomial is exactly the true propensity score.

The results for the linear model (Specification 1b) are presented in Panel II, and are similar to Panel I. Specifically, biases and MSEs are high in the under-fitted model, but settle down quickly, with little penalty in terms of MSE and coverage rates for over-fitting. In addition, there is little difference between the unnormalized and normalized estimators.

Panel III contains the results for the symmetric quadratic model (Specification 1c). Here, the results are very surprising, indicating equally good performance in under-fitted, correct, and over-fitted models. This

arises 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 $\mathrm{E}[X]$, which is $1/2$. Further, note that when the estimated propensity score is constant across observations (i.e. we utilize 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) $\mathrm{E}[Y|T=1]$. Now, $\mathrm{E}[Y|T=1] = \mathrm{E}[\mathrm{E}[Y|T=1,X]|T=1] = \int \mathrm{E}[Y|T=1,X=x]p(x|T=1)dx$, where $p(x|T=1)$ is the conditional density of X given $T=1$, which by Bayes' rule is $p(x|T=1) = p(T=1|x)p(T=1)/p(x)$. Since $\mathrm{E}[Y|T=1,X=x] = (\beta_1+\beta_2)+(\beta_3+\beta_4)x$, it follows that

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

Furthermore, since $p(T=1|x)$ is symmetric about $1/2$, it follows that $p(x|T=1)$ is symmetric about $1/2$. Thus, $\mathrm{E}[X|T=1] = 1/2 = \mathrm{E}[X]$ and $\mathrm{E}[Y|T=1] = (\beta_1+\beta_2)+(\beta_3+\beta_4)\,\mathrm{E}[X] = \mathrm{E}[Y(1)]$. A similar argument follows for the $T=0$ case. As a result, the simple estimator in (9) is unbiased and consistent in this case.[2]

The results are markedly different for the asymmetric quadratic model (Specification 1d); results are shown in Panel IV. The biases and MSEs are very high for under-fitted models (columns 1 and 2), with low penalty, in terms of MSE, for over-fitted models. In addition, the estimator using normalized weights performs better in the over-fitted models in terms of biases, MSEs, and coverage rates.

The results for the final logit specification, polynomial of forth degree (Specification 1e), are not surprising; results are given in Panel V. First, relatively extreme under-fitting (columns 1 and 2) leads to very poor performance by both estimators. Second, there is little penalty for over-fitting the propensity score. Finally, there is little difference between the unnormalized and normalized estimators in over-fitted models.

The remaining results pertain to data simulated from non-logit *true* models. As a result, there is no correct model. Results for the symmetric peak model (Specification 2a) are given in Panel VI. As in the previous symmetric specifications (1a and 1c), we find little substantive differences across models.

[2]The authors are indebted to Keisuke Hirano for this argument.

However, the biases, MSEs, and coverage probabilities improve using with the normalized estimator due to a large number of estimated propensity score close to zero or unity. On the other hand, in the asymmetric peak model (Specification 2b), we find obvious benefits to over-fitting the propensity score, as well as little substantive difference between the unnormalized and normalized estimators (Panel VII).

The final set of results uses data simulated using non-continuous propensity scores. Results for the monotonic step function (Specification 3a) are shown in Panel VIII, and indicate that all models, as well as both estimators, perform equally well with the exception of the zero degree polynomial logit model. More interesting results arise with the non-monotonic (asymmetric) step function (Specification 3b); results are displayed in Panel IX. First, there is little penalty in terms of bias and MSE for over-fitting, particularly with the normalized estimator. Second, the benefits of using the normalized – relative to the unnormalized – estimator are extremely pronounced: biases and MSEs are lower in most models. This occurs 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 over-fitting the propensity score model and often a large penalty for under-fitting. Since in practice researchers rarely know the true form of the propensity score, it appears prudent to over-fit. Specifically, in cases where the true propensity score is symmetric (and not flat), there is little benefit (and little cost) to over-fitting. Conversely, in all asymmetric cases, with the exception of an asymmetric, non-continuous, and monotonic, propensity score (Specification 3a), over-fitting is beneficial. 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 under-fitted. Thus, over-fitting in combination with the normalized estimator seems wise. Finally, whether or not 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 ought to use the normalized estimator and provide a series of estimates using increasingly sophisticated specifications of the propensity score model. We illustrate this approach with a timely application assessing the environmental effects of the GATT/WTO.

4 Application

4.1 Motivation

Copeland and Taylor (2004, p. 7) state 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."

Given the stakes involved, Taylor (2004, p. 1) argues that this constitutes "one of the most important debates in trade policy." The debates become even more heated when one focuses on the interplay between the WTO and the environment, as evidenced by the political demonstrations at the WTO Ministerial Meetings in Seattle in 1999. As noted on CNN's website[3]: "Opponents of the WTO claim it puts too much control into the hands of government and big business. They believe international trade agreements do not often take into account the impact on the health and safety of workers, the damage to the environment and the welfare of poorer nations." On the other hand, Mark Vaile, Australia's trade minister told CNN in the same article: "On both the labor standards and the environment there are very clear and demonstrable benefits that can flow from improved trade and trade liberalization across the world to those sectors – particularly in the developing world." Weinstein and Charnovitz (2001, p. 147) state: "[T]he WTO has started to develop an environmental conscience. With only a few tweaks, it can turn greener still... The organization is in fact developing constructive principles for accommodating both trade and environmental concerns."

In large part, the disagreement between WTO proponents and environmentalists stems from the complex relationship between trade and the environment in general, and the WTO and the environment in particular. Specifically, there are a number of avenues via which the WTO may affect the environment. First, the WTO *may* increase the volume of trade through the relaxation of barriers to trade, and the increase in trade *may* help or harm the environment. With respect to the impact of the WTO (and its predecessor, the GATT) on trade, Rose (2004a, 2004b, 2005) finds surprisingly little significant association between the WTO and liberal trade policy and only moderate evidence on expanded trade volumes. Moreover, even if the WTO does liberalize trade, the impact of trade on the environment is also not clear. Recent evidence in Antweiler et al. (2001), Frankel and Rose (2005), Chintrakarn and Millimet (2006), and others find little, if any, evidence of a detrimental effect of trade on various measures of environmental quality.[4]

Second, several provisions in the WTO *may* impede a country's ability to unilaterally enact trade-related measures designed to protect the environment.[5] The pillars of the WTO framework – the principles of Most Favored Nation (MFN) and national treatment – require that any advantage extended 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. Relevant for environmental policy, WTO dispute settlement rulings typically fail to consider differences in process

[3]See http://www.cnn.com/SPECIALS/2000/globaljustice/seattle.html.

[4]See Copeland and Taylor (2004) for a review of the literature.

[5]See Bernasconi-Osterwalder et al. (2006) for an excellent summary of the relevant jurisprudence.

and production methods (PPMs) in determining likeness. Thus, countries have been unable to discriminate against goods on the basis of non-product related PPMs (Atlhammer and Dröge 2003). This is particularly relevant for environmental policy since many trade-related environmental measures discriminate on the basis of characteristics of the producer which fail to alter the final product. That said, 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". However, exceptions are limited by the chapeau of Article XX, requiring trade measures to 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, prior to the *US-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 non-product related PPMs had been deemed inconsistent with the WTO; now the door is open (Bernasconi-Osterwalder et al. 2006). DeSombre and Barkin (2002, p. 17) state: "The recent ruling on the shrimp/turtle issue represents the first time that a DSM [Dispute Settlement Mechanism] ruling has clearly supported a breach of international trade rules for the purpose of environmental protection." Thus, while complex, there is room in the WTO provisions for unilateral trade restrictions on the basis of non-product related, environmentally hazardous PPMs.

Third, the WTO provisions discussed above *may* impede the efficacy and/or viability of multilateral environmental agreements (MEAs) in two ways (Atlhammer and Dröge 2003). First, MEAs typically rely on the voluntary cooperation of countries to abide by a treaty's obligations. In the event of non-compliance, trade restrictions are a useful enforcement mechanism. However, if the offending party is a WTO member, such restrictions may be disallowed. Second, trade restrictions on countries opting not to sign or ratify a particular MEA may limit free-riding, but again may not be allowed under the WTO system. Thus, WTO provisions may make it much more difficult to get an MEA adopted, or to enforce an existing MEA. Eckersley (2004, p. 27) investigates these claims, arguing that the threat of challenges under the WTO has had a "chilling effect" on MEA negotiations and imposition of trade restrictions under existing MEAs.

Finally, the WTO *may* provide a framework, as argued in Bagwell and Staiger (2001a, 2001b), to handle issues such as a 'race-to-the-bottom' or 'regulatory chill' in environmental (or labor) standards. Specifically, countries can potentially compensate domestic import-competing firms for higher (lower) domestic (foreign) standards by raising tariffs so that market access remains constant. In the end, then, the impact of the GATT/WTO on the environment is *a priori* ambiguous, and requires empirical analysis.

4.2 Data

The data are at the country-level and come from Frankel and Rose (2005); thus, we provide only limited details.[6] We analyze five measures of environmental quality: per capita carbon dioxide emissions, the average annual deforestation rate from 1990 – 1996, energy depletion, rural access to clean water, and urban access to clean water. We employ 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 supplement the data from Frankel and Rose (2005) with an indicator of whether the country is a GATT/WTO member. In the analysis, we use observations from 1990 (prior to the WTO) and 1995 (after the creation of the WTO). Table 2 provides summary statistics and descriptions of the variables.

4.3 Results

The results are presented in Table 3. For each outcome, we estimate 19 specifications to assess the impact of over-fitting the propensity score equation. Specification (1) includes only a constant in the propensity score equation. Each successive specification then adds the variable listed in the second column. Thus, specification (4) includes a constant and a linear term for each of the three covariates. Specification (7) ((10)) includes a constant as well as linear and squared (and cubic) terms for each of the three covariates. Finally, specification (19) corresponds to a third order linear approximation. Within each specification, we display the unnormalized and normalized treatment effect estimate, as well as the standard error.[7]

For per capita carbon dioxide, the various results indicate three salient findings. First, there are large differences between the unnormalized and normalized estimates in all specifications except for (1); the unnormalized estimates are typically twice as large and are therefore much more likely to be statistically significant. Second, there are large differences in both the unnormalized and normalized estimates when moving from a linear specification of the propensity score to adding higher order terms. However, over-fitting – even relative to specification (5) that includes a quadratic term for only one of the three covariates – does not qualitatively alter the results. Thus, while there is little benefit in this case to over-fitting, there is no cost either. Finally, in terms of the actual point estimates, we find moderately statistically significant evidence that GATT/WTO membership raises per capita carbon dioxide emissions ($\tau = 1.69$; $\tau_{norm} = 0.91$; standard error = 0.52). This is consistent with the results in Frankel and Rose (2005) assessing the impact of trade openness on per capita carbon dioxide, as well as the argument that the GATT/WTO framework may explicitly or implicitly deter MEAs, since international cooperation is required to limit carbon dioxide

[6]The authors are grateful to Andrew Rose for making the data available (http://faculty.haas.berkeley.edu/arose/).

[7]For all estimated models there was a sufficient overlap in the distribution of propensity scores for treated and controls.

emissions give that it is purely a global externality.

In terms of deforestation, there are two primary examples of the interplay between the GATT/WTO and environmental quality. 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 *US - Softwood Lumber*. Recently, the WTO Appellate Body has upheld U.S. anti-dumping rates, but a WTO panel also held that the U.S. countervailing duty determination does not conform to WTO requirements.

Turning to the results, we find that the use of the choice of normalized or unnormalized estimator has substantially less impact than for per capita carbon dioxide. On the other hand, the exact specification of the propensity score does have a substantial effect on inference. Specifically, entering each covariate linearly in the propensity score equation (specification (4)) yields statistically insignificant effects of GATT/WTO membership on deforestation ($\tau = 0.18$; $\tau_{norm} = 0.11$; standard error = 0.18). Adding quadratic and cubic terms for each covariate, but excluding any interactions (specification (10)) also yields insignificant estimates ($\tau = 0.22$; $\tau_{norm} = 0.20$; standard error = 0.15). However, adding interactions between the linear terms yields statistically significant effects ($\tau = 0.26$; $\tau_{norm} = 0.30$; standard error = 0.14), whereas moving to the third order linear approximation (specification (19)) yields statistically insignificant estimates ($\tau = 0.24$; $\tau_{norm} = 0.22$; standard error = 0.15). Thus, we conclude that there is no evidence that GATT/WTO membership accentuates the annual rate of deforestation. Interestingly, the changes in inference across the various specifications is solely due to a change in the estimate; the standard errors are essentially constant. This confirms the results from the Monte Carlo study: there is no efficiency loss from over-fitting the propensity score.

The next set of results pertain to energy depletion, where a larger value indicates greater energy consumption. As with the previous outcomes, we find that there are important differences across the unnormalized and normalized estimates, as well as across the various propensity score specifications. Moreover, the efficiency of the estimates are not hindered by over-fitting the propensity score. Specifically, we find a moderately statistically significant effect of GATT/WTO membership on energy conservation when fitting a third order linear approximation (specification (19)) and using the normalized estimator ($\tau = -1.65$; $\tau_{norm} = -1.85$; standard error = 1.04), as well as when we omit all interactions, but include third order polynomials for each covariate (specification (10)) ($\tau = -2.10$; $\tau_{norm} = -2.27$; standard error = 1.06). However, only including some linear interactions makes the normalized estimate statistically insignificant; for example, see specification (13) ($\tau = -1.80$; $\tau_{norm} = -1.43$; standard error = 1.02). In the end, using the normalized weights and over-fitting the propensity score indicates a statistically significant, beneficial

impact of GATT/WTO membership on energy use, consonant with Frankel and Rose's (2005) analysis of trade openness.

In terms of rural and urban access to clean water, we initially note two salient findings. First, the unnormalized estimates are much larger than the corresponding normalized estimates. For rural (urban) access, the unnormalized estimates are roughly 1.5 to 2.5 (five to six) times as large. Second, unlike for the previous three outcomes, the variance of the estimates does vary, sometimes quite substantially, across the various specifications. For instance, the standard error increase by roughly 270% (225%) from specification (17) to (19) for rural (urban) water access. 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 2). Thus, the relative costlessness of over-fitting depends crucially on the sample size and the variation in the outcome. That said, we find a statistically significant effect of GATT/WTO membership on rural and urban access to clean water using the unnormalized estimator, and moderately statistically significant effects on rural access to clean water using the normalized estimator, across the majority of specifications. For example, in the cubic specification (specification (10)), we find statistically significant effects (rural: $\tau = 16.94$; $\tau_{norm} = 9.48$; standard error = 4.92; urban: $\tau = 16.59$; $\tau_{norm} = 2.74$; standard error = 5.95).

5 Conclusion

While the use of propensity score methods in the estimation of treatment effects in non-experimental settings is proliferating, little practical guidance exists for researchers in terms of how one ought to specify the propensity score model. The perception appears to be that, although under-specifying propensity scores will yield inconsistent estimates, over-specifying propensity scores is inefficient at best, and inconsistent at worst. However, this perception need not be correct. Using a Monte Carlo study and two weighting estimators, we find little penalty to over-fitting propensity scores, and in fact find numerous cases where over-specifying the model proves 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.

Illustrating this approach, we address an important question in the international arena: Does the WTO hinder environmental improvement? Using data from the 1990s, we find that the WTO is beneficial in terms of the environmental measures analyzed that are local in nature, and detrimental in terms of those that are global. Moreover, these results are sensitive to the specification of the propensity score model,

thus confirming the benefits of comparing the results across numerous propensity score specifications.

Open questions still remain, however. Additional research is called for to further assess the effects of over-specifying the propensity score model through the inclusion of irrelevant variables. In particular, there are two types of irrelevant variables: irrelevant higher order terms of relevant variables and variables that are irrelevant at all orders. Aside from the flat propensity score specification, our Monte Carlo study focuses only on the former. As such, future work should aim to offer guidance to applied researchers with respect to the latter.

A Asymptotic Variance Bound

From Theorem 1 in Hirano et al. (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, $\tau(x) = \mathrm{E}[Y(1) - Y(0)|X = x]$, and $\sigma_j^2(x) = V[Y(j)|X = x]$, $j = 0, 1$.

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 $X \sim U[0, 1]$ and $e_i \sim N(0, \sigma_2)$.

Thus

$$\begin{aligned} Y_i(0) &= \beta_1 + \beta_3 X_i + e_i \\ Y_i(1) &= Y_i(0) + \beta_2 + \beta_4 X_i. \end{aligned}$$

Therefore

$$\sigma_0^2(x) = \sigma_1^2(x) = \sigma^2$$

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

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

From the expressions above

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

Hence

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

References

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

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

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

[4] Bagwell, K. and R.W. Staiger (2001b), "The WTO as a Mechanism for Securing Market Access Property Rights: Implications for Global Labor and Environmental Issues," *Journal of Economic Perspectives*, 15, 69-88.

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

[6] Brookhart, M.A., S. Schneeweiss, K.J. Rothman, R.J Glynn, J. Avorn, and T. Stürmer (2006), "Variable selection for propensity score models," *American Journal Of Epidemiology*, 163, 1149-1156.

[7] Bryson, A., R. Dorsett, and S. Purdon (2002), "The Use of Propensity Score Matching in the Evaluation of Labour Market Policies," Working Paper No. 4, Department for Work and Pensions.

[8] Chintrakarn, P. and D.L. Millimet (2006), "The Environmental Consequences of Trade: Evidence from Subnational Trade Flows," *Journal of Environmental Economics and Management*, forthcoming.

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

[10] 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.

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

[12] DeSombre, E.R. and J.S. Barkin (2002), "Turtles and Trade: The WTO's Acceptance of Environmental Trade Restrictions," *Global Environmental Politics*, 2, 12-18.

[13] Eckersley, R. (2004), "The Big Chill: The WTO and Multilateral Environmental Agreements," *Global Environmental Politics*, 4, 24-50.

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

[15] Frankel, J.A. and D. Romer (1999), "Does Trade Cause Growth?" *American Economic Review*, 89, 379-399.

[16] Frankel, J.A. and A.K. Rose (2005), "Is Trade Good or Bad for the Environment? Sorting Out the Causality," *Review of Economics and Statistics*, 87, 85-91.

[17] Geman. S. and C. Hwang (1982), "Nonparametric Maximum Likelihood Estimation by the Method of Sieves," *Amlals of Statistics*, 10, 401-414.

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

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

[20] 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.

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

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

[23] Ichimura H. and O. Linton (2005), "Asymptotic expansions for some semiparametric program evaluation estimators,"in *Volume in Honour of Tom Rothenberg*, Eds. D.W.K. Andrews and J. Stock, Cambridge University Press.

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

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

[26] LaLonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76,604-620

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

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

[29] Rose, A.K. (2004a), "Do We Really Know that the WTO Increases Trade?" *American Economic Review*, 94, 98-114.

[30] Rose, A.K. (2004b), "Do WTO Members have More Liberal Trade Policy?" *Journal of International Economics*, 63, 209-235.

[31] Rose, A.K. (2005), "Which International Institutions Promote International Trade?" *Review of International Economics*, 13, 682-698.

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

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

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

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

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

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

[38] Weinstein, M.M. and S. Charnovitz (2001), "The Greening of the WTO," *Foreign Affairs*, 80, 147-156.

[39] Zhao, Z. (2005), "Sensitivity of Propensity Score Methods to the Specifications," IZA Discussion Paper No 1873.

Figure 1: Propensity Scores Specifications

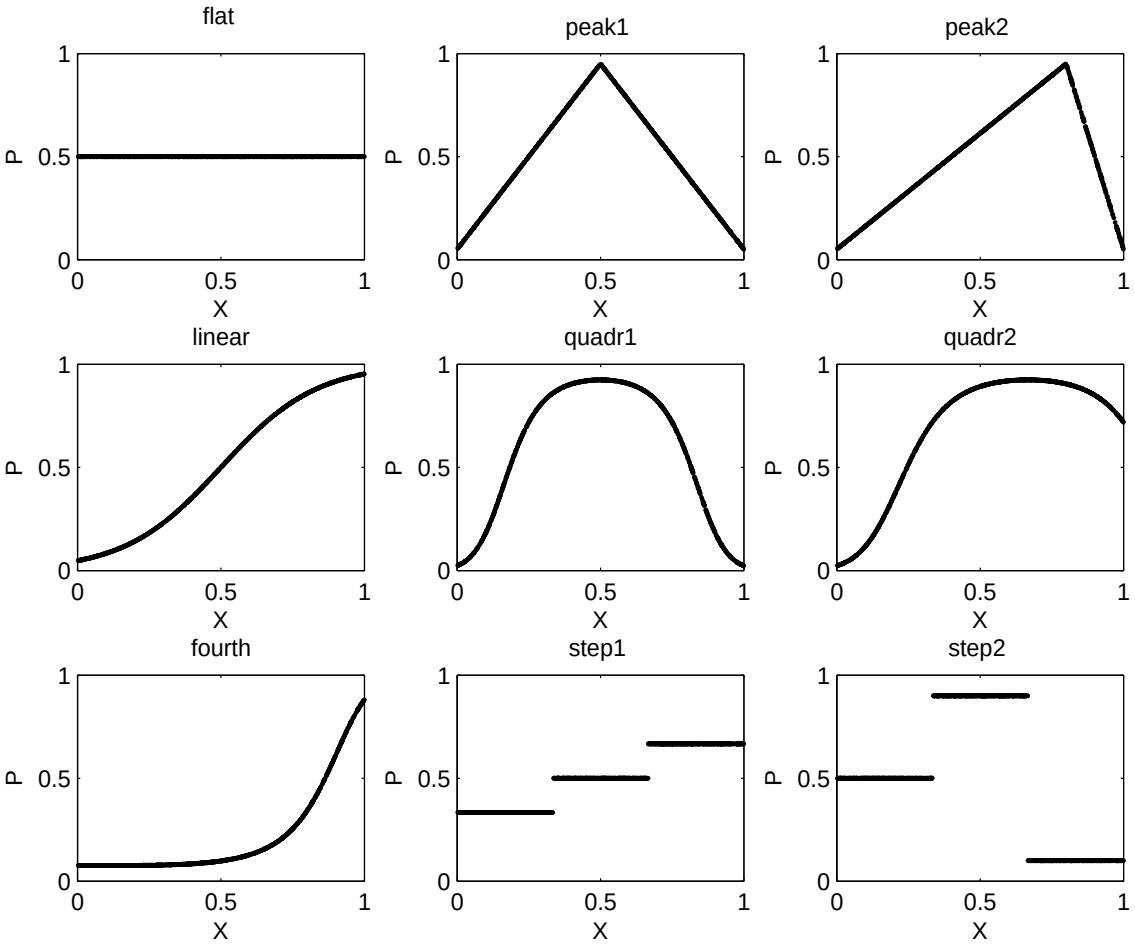

Table 1. Monte Carlo Results

		Polynomial Order in Propensity Score Estimation									
		0	1	2	3	4	5	6	7	8	9
I. Flat Propensity Score											
1000 x	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	2.06	1.32	1.32	1.32	1.32	1.32	1.32	1.32	1.32	1.32
Propensity	MISE	0.0002	0.0005	0.0007	0.0010	0.0012	0.0014	0.0017	0.0019	0.0021	0.0023
Score	SUP	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 x	Bias	-536.38	1.50	0.10	0.60	0.32	0.48	0.47	-0.84	-1.21	-2.98
	Bias (normalized)	-536.38	-1.13	-0.71	0.66	-0.07	0.03	-1.43	-2.53	-4.05	-6.27
	MSE	289.33	3.30	2.90	2.81	2.88	2.94	2.95	2.98	3.02	3.01
	MSE (normalized)	289.33	3.61	3.23	2.88	2.98	3.19	3.04	3.16	3.16	3.18
	Estimated Variance	1.65	3.45	2.92	2.92	3.67	6.02	4.14	14.68	8.07	3.40
Propensity	MISE	0.1287	0.0027	0.0027	0.0029	0.0030	0.0031	0.0032	0.0034	0.0035	0.0036
Score	SUP	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	0.952	0.947	0.947	0.942	0.938	0.933	0.934
Rates	Normalized	0.000	0.952	0.935	0.946	0.944	0.955	0.943	0.937	0.931	0.927
III. Quadratic (Symmetric) Propensity Score											
1000 x	Bias	-0.49	-1.87	-1.54	-4.00	-0.41	-1.06	-3.82	-6.65	-8.25	-9.61
	Bias (normalized)	-0.49	-0.18	0.02	0.37	0.55	0.50	0.41	0.02	-0.02	-0.08
	MSE	3.32	1.45	5.89	5.96	3.46	3.54	3.43	3.32	3.30	3.34
	MSE (normalized)	3.32	1.44	3.23	3.28	3.26	3.13	3.11	3.00	2.97	2.97
	Estimated Variance	3.15	1.38	8.07	8.38	7.15	27.66	13.01	7.19	6.15	4.11
Propensity	MISE	0.1042	0.1041	0.0005	0.0006	0.0007	0.0008	0.0009	0.0011	0.0012	0.0013
Score	SUP	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	0.966	0.960	0.962	0.951	0.954	0.952	0.946	0.939
Rates	Normalized	0.941	0.944	0.989	0.989	0.965	0.955	0.959	0.962	0.967	0.965
IV. Quadratic 2 (Asymmetric) Propensity Score											
1000 x	Bias	-595.14	-24.72	-1.78	-2.05	-2.52	-3.66	-4.52	-6.19	-7.78	-9.02
	Bias (normalized)	-595.14	429.98	-1.44	-1.91	-2.01	-2.12	-2.25	-2.79	-3.15	-3.34
	MSE	357.05	10.43	5.00	4.17	3.29	3.36	3.48	3.44	3.54	3.53
	MSE (normalized)	357.05	192.02	3.36	3.35	3.12	3.10	3.16	3.19	3.18	3.16
	Estimated Variance	2.63	19.72	6.73	6.12	9.99	8.78	23.74	5.59	4.43	4.46
Propensity	MISE	0.0880	0.0157	0.0004	0.0005	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012
Score	SUP	0.66	0.24	0.05	0.06	0.06	0.07	0.08	0.09	0.09	0.10
Coverage	Unnormalized	0.000	0.986	0.953	0.952	0.953	0.946	0.946	0.942	0.938	0.933
Rates	Normalized	0.000	0.073	0.985	0.967	0.956	0.958	0.951	0.951	0.953	0.949

Notes: MSE = mean squared error; MISE = mean integrated squared error; SUP = sup norm. Estimated variance uses formula from Hirano et al. (2003). Asymptotic variance bound is 1.32 (Panel I); 2.49 (Panel II); 2.85 (Panel III); 2.85 (Panel IV). 1,000 simulations used for each column. See text for further details.

Table 1 (cont.). Monte Carlo Results

		Polynomial Order in Propensity Score Estimation									
		0	1	2	3	4	5	6	7	8	9
V. Fourth Order Propensity Score											
1000 x	Bias	-278.61	279.32	26.01	1.69	1.27	1.34	1.12	0.33	-0.32	-1.12
	Bias (normalized)	-278.61	40.49	-6.97	0.08	1.15	1.15	1.38	1.48	0.80	0.18
	MSE	79.61	95.34	4.07	2.89	2.93	2.90	2.93	2.97	2.97	2.97
	MSE (normalized)	79.61	6.79	3.05	2.97	2.96	2.88	2.92	2.97	3.01	2.97
	Estimated Variance	2.09	111.54	3.66	2.99	3.07	3.12	3.29	9.11	3.22	3.34
Propensity	MISE	0.0462	0.0068	0.0008	0.0005	0.0006	0.0007	0.0008	0.0010	0.0011	0.0012
Score	SUP	0.65	0.24	0.07	0.05	0.06	0.06	0.07	0.08	0.09	0.09
Coverage	Unnormalized	0.000	0.996	0.937	0.951	0.946	0.944	0.948	0.947	0.953	0.952
Rates	Normalized	0.000	1.000	0.958	0.942	0.945	0.947	0.947	0.944	0.949	0.945
VI. Peak (Symmetric) Propensity Score											
1000 x	Bias	1.89	1.51	51.79	51.37	21.31	21.01	18.26	17.71	9.66	8.57
	Bias (normalized)	1.89	2.38	2.69	2.64	2.33	2.48	2.71	2.75	2.43	2.03
	MSE	2.45	1.21	5.24	5.29	2.44	2.41	2.87	2.98	2.35	2.22
	MSE (normalized)	2.45	1.21	2.09	2.13	1.90	1.88	2.23	2.31	2.09	2.07
	Estimated Variance	2.50	1.26	2.86	3.08	1.90	1.89	12.97	25.70	16.97	2.17
Propensity	MISE	0.0664	0.0667	0.0036	0.0038	0.0027	0.0029	0.0019	0.0020	0.0019	0.0021
Score	SUP	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	0.870	0.925	0.918	0.945	0.948	0.941	0.937
Rates	Normalized	0.956	0.959	0.971	0.971	0.954	0.951	0.960	0.957	0.952	0.953
VII. Peak 2 (Asymmetric) Propensity Score											
1000 x	Bias	-269.69	12.96	56.75	-20.55	10.38	10.87	2.96	2.94	2.57	-0.79
	Bias (normalized)	-269.69	81.38	-17.88	-23.09	-12.53	-13.74	-9.91	-6.02	-5.76	-4.66
	MSE	75.06	1.77	7.11	2.30	2.26	2.27	1.92	2.05	2.24	2.00
	MSE (normalized)	75.06	8.19	2.67	2.28	2.17	2.27	2.04	2.03	2.05	2.00
	Estimated Variance	2.23	1.58	10.42	1.91	2.93	3.86	2.15	4.82	40.96	2.36
Propensity	MISE	0.0665	0.0404	0.0134	0.0052	0.0020	0.0021	0.0020	0.0019	0.0020	0.0021
Score	SUP	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	0.990	0.932	0.963	0.956	0.953	0.953	0.946	0.946
Rates	Normalized	0.000	0.471	0.978	0.931	0.954	0.950	0.952	0.949	0.946	0.950

Notes: MSE = mean squared error; MISE = mean integrated squared error; SUP = sup norm. Estimated variance uses formula from Hirano et al. (2003). Asymptotic variance bound is 2.85 (Panel V); 1.96 (Panel VI); 1.96 (Panel VII). 1,000 simulations used for each column. See text for further details.

Table 1 (cont.). Monte Carlo Results

		Polynomial Order in Propensity Score Estimation									
		0	1	2	3	4	5	6	7	8	9
VIII. Step (Monotone) Propensity Score											
1000 x	Bias	-222.99	1.21	1.34	-0.77	-0.74	-0.36	-0.34	0.14	0.08	-0.35
	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	1.98	1.40	1.40	1.39	1.39	1.40	1.40	1.40	1.40	1.40
Propensity	MISE	0.0188	0.0024	0.0026	0.0025	0.0027	0.0028	0.0030	0.0028	0.0030	0.0028
Score	SUP	0.18	0.10	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	0.950	0.952	0.954	0.957	0.957	0.956
IX. Step 2 (Non-Monotone) Propensity Score											
1000 x	Bias	265.03	-158.92	88.60	598.09	35.58	145.28	-32.42	-7.27	-82.26	-99.81
	Bias (normalized)	265.03	-73.83	-79.98	-124.15	-38.46	-42.91	-14.86	-15.17	-7.67	-2.10
	MSE	72.41	27.31	17.01	2138.43	5.40	28.99	5.36	7.81	12.79	16.62
	MSE (normalized)	72.41	6.88	11.31	33.53	4.36	5.99	3.25	4.38	3.70	3.57
	Estimated Variance	2.30	2.36	79.34	4.03E+08	4.34	27.79	4.92	8129.52	15.49	16.69
Propensity	MISE	0.1069	0.0832	0.0361	0.0329	0.0194	0.0166	0.0123	0.0105	0.0102	0.0099
Score	SUP	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	0.995	0.998	0.938	0.892	0.939	0.974	0.963	0.960
Rates	Normalized	0.000	0.717	0.916	0.971	0.937	0.971	0.975	0.985	0.992	0.997

Notes: MSE = mean squared error; MISE = mean integrated squared error; SUP = sup norm. Estimated variance uses formula from Hirano et al. (2003). Asymptotic variance bound is 1.40 (Panel VIII); 2.50 (Panel IX). 1,000 simulations used for each column. See text for further details.

Table 2. 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	7302.91	7468.41	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

Source: 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/.

Table 3. Impact of GATT/WTO Membership on Environmental Quality

		Per Capita Carbon Dioxide			Deforestation			Energy Depletion		
	Additional	Estimator		Standard	Estimator		HIR	Estimator		Standard
Specification	Variable	τ	τ_{norm}	Error	τ	τ_{norm}	Std Error	τ	τ_{norm}	Error
(1)	constant	1.56	1.56	0.58	-0.03	-0.03	0.20	-4.57	-4.57	1.54
(2)	x_1	0.94	0.42	0.64	0.18	0.12	0.18	-3.49	-4.16	1.33
(3)	x_2	0.82	0.27	0.71	0.19	0.13	0.18	-3.50	-4.16	1.33
(4)	x_3	0.87	0.33	0.69	0.18	0.11	0.18	-3.09	-3.69	1.20
(5)	x_1^2	1.72	1.03	0.53	0.21	0.27	0.15	-2.10	-1.60	1.14
(6)	x_2^2	1.77	1.06	0.50	0.21	0.26	0.15	-2.19	-1.76	1.14
(7)	x_3^2	1.78	1.05	0.49	0.21	0.21	0.15	-2.22	-2.18	1.07
(8)	x_1^3	1.77	1.04	0.50	0.21	0.23	0.15	-2.14	-1.94	1.06
(9)	x_2^3	1.74	1.02	0.51	0.22	0.21	0.15	-2.09	-2.20	1.07
(10)	x_3^3	1.75	1.02	0.50	0.22	0.20	0.15	-2.10	-2.27	1.06
(11)	x_1*x_2	1.75	1.01	0.49	0.24	0.29	0.14	-1.99	-1.52	1.06
(12)	x_1*x_3	1.75	1.01	0.50	0.26	0.33	0.14	-1.71	-1.12	0.96
(13)	x_2*x_3	1.77	1.04	0.48	0.26	0.30	0.14	-1.80	-1.43	1.02
(14)	$x_1^2*x_2$	1.80	1.04	0.47	0.25	0.22	0.14	-1.76	-2.08	1.10
(15)	$x_1^2*x_3$	1.79	1.03	0.46	0.25	0.20	0.14	-1.72	-2.22	1.10
(16)	$x_2^2*x_1$	1.79	1.03	0.47	0.25	0.20	0.14	-1.72	-2.22	1.09
(17)	$x_2^2*x_3$	1.78	1.00	0.48	0.25	0.20	0.14	-1.72	-2.22	1.09
(18)	$x_3^2*x_1$	1.69	0.86	0.52	0.25	0.21	0.14	-1.68	-1.97	1.06
(19)	$x_3^2*x_2$	1.69	0.91	0.52	0.24	0.22	0.15	-1.65	-1.85	1.04

Notes: The propensity score equation in specification (1) contains only a constant. Each specification after that includes the variables from the previous specification plus the additional variable. x_1 is real GDP; x_2 is area per capita; and, x_3 is a measure of polity. See text for further details.

Table 3 (cont.). Impact of GATT/WTO Membership on Environmental Quality

Specification	Additional Variable	Rural Water Access			Urban Water Access		
		Estimator		Standard Error	Estimator		Standard Error
		τ	τ_{norm}		τ	τ_{norm}	
(1)	constant	8.85	8.85	5.09	2.14	2.14	4.54
(2)	x_1	10.08	6.24	5.64	5.72	-1.64	6.38
(3)	x_2	12.29	7.72	5.24	7.94	-0.95	6.37
(4)	x_3	12.35	8.01	5.29	6.93	-1.59	6.20
(5)	x_1^2	17.12	10.24	4.67	16.55	3.18	5.27
(6)	x_2^2	17.11	8.79	4.63	17.91	2.49	5.27
(7)	x_3^2	17.31	9.58	4.62	17.49	2.86	5.53
(8)	x_1^3	17.27	9.55	4.64	17.51	2.86	5.60
(9)	x_2^3	16.88	9.54	4.88	16.55	2.84	5.96
(10)	x_3^3	16.94	9.48	4.92	16.59	2.74	5.95
(11)	x_1*x_2	16.90	9.08	4.93	17.71	2.90	5.43
(12)	x_1*x_3	16.71	8.73	4.86	17.74	2.66	5.33
(13)	x_2*x_3	15.74	8.93	8.73	15.95	2.69	8.18
(14)	$x_1^2*x_2$	15.69	8.06	6.79	18.12	3.34	6.17
(15)	$x_1^2*x_3$	17.04	7.25	4.91	21.11	3.12	5.46
(16)	$x_2^2*x_1$	17.14	7.34	4.88	21.34	3.07	5.63
(17)	$x_2^2*x_3$	18.35	8.09	4.74	22.52	3.25	5.66
(18)	$x_3^2*x_1$	18.87	7.48	11.24	24.19	3.87	9.79
(19)	$x_3^2*x_2$	18.51	7.64	12.77	23.01	3.72	12.68

Notes: The propensity score equation in specification (1) contains only a constant. Each specification after that includes the variables from the previous specification plus the additional variable. x_1 is real GDP; x_2 is area per capita; and, x_3 is a measure of polity. See text for further details.