# Misspecification and the Propensity Score: The Possibility of Overadjustment*

Kevin A. Clarke†
Brenton Kenkel‡
Miguel R. Rueda§

June 11, 2015

**Abstract**

The popularity of propensity score matching has given rise to a robust, sometimes informal, debate concerning the number of pre-treatment variables that should be included in the propensity score. The standard practice when estimating a treatment effect is to include all available pre-treatment variables, and we demonstrate that this approach is not always optimal when the goal is bias reduction. We characterize the conditions under which including an additional relevant variable in the propensity score increases the bias on the effect of interest across a variety of different implementations of the propensity score methodology. Moreover, we find that balance tests and sensitivity analysis provide limited protection against overadjustment.

Keywords: Matching; Omitted variable bias; Conditioning

*We thank Jake Bowers, John Jackson, and Michael Peress for helpful comments and discussion. A previous version of this paper was given at the 27th Annual Summer Meeting of the Society for Political Methodology. We thank the participants for their comments. Errors remain our own.

†Corresponding author. Associate Professor, Department of Political Science, University of Rochester, Rochester, NY 14627-0146. Email: kevin.clarke@rochester.edu.

‡Assistant Professor, Department of Political Science, Vanderbilt University, Nashville, TN 37203. Email: brenton.kenkel@vanderbilt.edu.

§Assistant Professor, Department of Political Science, Emory University, Atlanta, GA 30233. Email: miguel.rueda@emory.edu

# Introduction

The increasing popularity of propensity scores raises questions regarding our statistical intuitions. Political scientists have a strong and natural tendency to transfer lessons learned from years of regression training to the newer technique. Our regression-based intuitions, however, do not hold necessarily when it comes to the use of propensity scores. Covariance adjustment is a prime example. The current wisdom regarding the specification of propensity score models is to include all available covariates. This intuition is mistaken in a regression context (see, for example, Clarke, 2005, 2009; Luca, Magnus and Peracchi, 2015), but it remains an open question whether the intuition is mistaken in a propensity score context.

Consider a field experiment where the treatment is a get-out-the-vote effort aimed at swing voters. Suppose that data on partisanship is available to the analyst, but data on personal income is not. Should the analyst adjust for partisanship in the propensity score? (We revisit this example in Section 4.)

Such questions are of interest to political scientists and statisticians alike. Political scientist and statistician Andrew Gelman has engaged a number of prominent scholars in the debate without a convincing resolution.[1] Similar debates have taken place in the formal statistical literature (see Pearl, 2009). The discussion takes place within the potential outcomes framework (see the appendix for a brief primer), and the question is whether conditioning on a new covariate can only decrease confounding bias when estimating a treatment effect. Much of the debate lacks the hard evidence that is likely to convince researchers one way or another.

To that end, this paper accomplishes four goals. First, assuming that political scientists do not have access to the full list of covariates comprising the data-generating process, we characterize the conditions under which including an additional covariate in the propen-

[1] http://stat.columbia.edu/∼gelman/blog/

sity score increases the bias on a treatment estimator. We find that these circumstances are not unusual. Second, we characterize these conditions across a variety of different implementations of the propensity score methodology— nearest neighbor matching, matching within calipers, stratification by blocks, weighting, and regression—to assess which, if any, of these implementations helps mitigate the problem. Our results show that differences across techniques are minor. Third, we assess the ability of post-estimation tools such as balance tests and sensitivity analysis to alert us to the problem of overconditioning. The results are not encouraging. Finally, on a more positive note, political scientists have a lot of theoretical knowledge, and we demonstrate how to leverage that knowledge to determine which pre-treatment confounders to exclude in the presence of known omitted variables.

Our findings lend little credence to the claim that a researcher should condition on all available pretreatment covariates. Which variables should be included in a data analysis depends on factors that vary from situation to situation. We can tackle the problem using theory, judgement, and common sense, and we end with a discussion of how our results can be helpful to researchers.

# 1 The debate

In responding to a paper by Rubin (2007) on the design of observational studies, Shrier (2008) raises the issue of selection bias caused by controlling for a covariate that is the common effect of two independent variables. He is interested in "M-structures," where a treatment $X$ causes an outcome $Y$, an unmeasured covariate, $U_1$, causes both $X$ and a measured covariate $Z$, and a second unmeasured covariate, $U_2$, causes both the measured covariate $Z$ and $Y$ (see Figure 1). In this situation, if a researcher controls for the measured covariate $Z$, a spurious dependence between $X$ and $Y$ is created that would bias the estimate of the causal effect of $X$ on $Y$ (because the value of $Z$ predicts the value of $[X, Y]$).

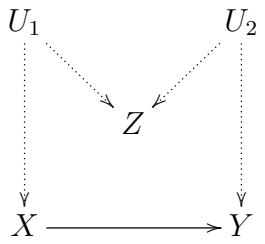



Figure 1: A causal directed acyclic graph of an M-structure. The dashed lines denote the effects of unmeasured covariates.

Pearl (2009, 1415) expands on Shrier's (2008) point and argues that the use of propensity score techniques increases the bias on the estimated ATE whenever

> ...treatment is strongly ignorable to begin with and becomes non-ignorable at some levels of $e_i$ [probability of assignment to treatment]. In other words, although treated and untreated units are balanced in each stratum of $e_i$, the balance only holds relative to the covariates measured; unobserved confounders may be highly unbalanced in each stratum of $e_i$, capable of producing significant bias.

That is, the propensity score balances treated and untreated observations relative only to observed covariates. A new association is introduced between treatment and outcome by conditioning on a variable that is not causally related to either, but is an indicator of unobserved factors that are not balanced. This new association may increase or decrease the bias on the estimate of the ATE. Pearl (2009, 1416) concludes that the effectiveness of propensity score techniques depends "critically on the choice of covariates, and that choice cannot be left to guesswork."

Rubin's (2009) response (and to a lesser extent Gelman's blog response[2]) to these claims is that not controlling for an observed covariate is bad practical advice "in all but the most unusual circumstances." Furthermore, Rubin argues that even if one were to condition on $Z$ in Figure 1, the result would be inefficient, but not biased. In the end, he argues that not

[2]http://www.stat.columbia.edu/c̃ook/movabletype/archives/2009/07/disputes_about.html

conditioning on an observed covariate because of fears of increasing bias "is neither Bayesian nor scientifically sound but rather it is distinctly frequentist and nonscientific ad hocery" (Rubin, 2009, 1421).

Our interest lies in assessing the risk of increasing bias through conditioning, and whether the choice of propensity score techniques makes a difference. Numerous articles have been published on misspecification of the propensity score, and the evidence is mixed. Drake (1993) finds that there is little difference between propensity score methods and prognostic (regression) models with regard to omitted confounders. Augurzky and Schmidt (2001) recommend including only highly significant variables in the propensity score equation. Brookhard et al. (2006) find that in small studies situations exist where it would be better, in terms of MSE, to exclude a true confounder from the propensity score model.

# 2 Motivating Example

Our analysis begins with a simple example that illustrates how conditioning on additional information may increase the bias of a treatment effect estimate. In the example, treatment assignment $T_i$ and the observed outcome $Y_i$ are functions of two covariates, $W_i$ and $Z_i$, of which only $W_i$ is observed by the analyst. We show that the naive difference of means estimator of the average treatment effect may be less biased than one that conditions on all available information, namely $W_i$.

The data-generating process of the example is as follows. The covariates $W_i$ and $Z_i$ are independent Bernoulli(0.5) random variables. The value of each potential outcome is increasing in $W_i$ and decreasing in $Z_i$,

$$Y_{i0} = 1 + W_i - 2Z_i,$$

$$Y_{i1} = 2 + W_i - 2Z_i.$$

Given these equations, the true value of the average treatment effect, our estimand of interest, is $\tau = E[Y_{i1} - Y_{i0}] = 1$. The probability of receiving treatment is increasing in both covariates,

$$\Pr(T_i = 1 \,|\, W_i, Z_i) = \frac{1}{4} + \frac{1}{4} W_i + \frac{1}{4} Z_i.$$

We can summarize the setup with a pair of $2 \times 2 \times 2$ tables:

| | $T_i = 0$ | | $T_i = 1$ | |
|---|---|---|---|---|
| | $Z_i = 0$ | $Z_i = 1$ | $Z_i = 0$ | $Z_i = 1$ |
| $W_i = 0$ | 1 | $-1$ | 2 | 0 |
| $W_i = 1$ | 2 | 0 | 3 | 1 |

(a) The observed value $Y_i$ as a function of the covariates and treatment assignment.

| | $T_i = 0$ | | $T_i = 1$ | |
|---|---|---|---|---|
| | $Z_i = 0$ | $Z_i = 1$ | $Z_i = 0$ | $Z_i = 1$ |
| $W_i = 0$ | $\frac{3}{16}$ | $\frac{2}{16}$ | $\frac{1}{16}$ | $\frac{2}{16}$ |
| $W_i = 1$ | $\frac{2}{16}$ | $\frac{1}{16}$ | $\frac{2}{16}$ | $\frac{3}{16}$ |

(b) The joint distribution of the covariates and treatment assignment, $\Pr(W_i, Z_i, T_i)$.

Table 1: Summary of the data-generating process for the motivating example.

Here, treatment assignment is strongly ignorable conditional on $W_i$ and $Z_i$. This assumptions allows us to write

$$\tau = E_{(W_i, Z_i)} \left[ E[Y_{i1} \,|\, W_i, Z_i, T_i = 1] - E[Y_{i0} \,|\, W_i, Z_i, T_i = 0] \right], \tag{1}$$

where $E_{(W_i, Z_i)}$ denotes expectation with respect to the distribution of the covariates. If both covariates were observed, the average treatment effect could be estimated consistently via subclassification (Rosenbaum and Rubin, 1983, Corollary 4.2):

$$\hat{\tau} = \frac{1}{N} \sum_{w=0}^{1} \sum_{z=0}^{1} \sum_{i=1}^{N} \delta_i(w, z)[Y_i T_i - Y_i(1 - T_i)],$$

where $\delta_i(w, z)$ is an indicator for $(W_i, Z_i) = (w, z)$. However, this first-best solution is not available because $Z_i$ is unobserved. If one followed the standard advice of conditioning on all covariates, and subclassified just on $W_i$, the resulting estimator would have asymptotic mean

$$\begin{aligned}
&\sum_{w=0}^{1} \Pr(W_i = w) \left[ E[Y_{i1} \,|\, T_i = 1, W_i = w] - E[Y_{i0} \,|\, T_i = 0, W_i = w] \right] \\
&\qquad = \frac{1}{2}\left(\frac{2}{3} - \frac{1}{5}\right) + \frac{1}{2}\left(\frac{9}{5} - \frac{4}{3}\right) \\
&\qquad = \frac{7}{15}.
\end{aligned}$$

Because the true treatment effect is $\tau = 1$, the asymptotic bias of this estimator is $-\frac{8}{15}$. Now consider estimating the average treatment effect via a naive difference of means in $Y_i$ between the observed treated and untreated groups. The mean of this estimator is

$$E[Y_{i1} \,|\, T_i = 1] - E[Y_{i0} \,|\, T_i = 0] = \frac{11}{8} - \frac{5}{8} = \frac{3}{4},$$

giving it a bias of $-\frac{1}{4}$. Although both estimators are biased, the magnitude of the bias is lower under the one that does not condition on all available covariates.

The reason it is not optimal to condition on $W_i$ in this scenario is that the unobserved variable $Z_i$ has countervailing effects. Recall that $W_i$ affects both treatment assignment and the response positively. Therefore, failure to adjust for $W_i$ causes the estimated average treatment effect to be biased upward. However, $Z_i$ has a positive effect on treatment assignment and a negative effect on the response, so its exclusion causes the estimate to be biased downward. Because $Z_i$ has a stronger effect on the response, the entirely unadjusted estimate of the treatment effect is biased downward. Balancing on $W_i$ lowers the estimated treatment effect, which increases the magnitude of the bias if $Z_i$ cannot also be balanced on. As such, balancing on a covariate that has countervailing effects with an unobserved

confounder may actually worsen the quality of a treatment effect estimate. In the Monte Carlo simulations described in the following sections, we find that this is often the case.

# 3 Design of the Experiments

We now turn to a series of Monte Carlo experiments illustrating that conditioning on all observed variables may be suboptimal if some confounding variables are unobserved. The results of these simulations show that the result of our motivating example—that conditioning on all available pre-treatment variables may be suboptimal in the presence of unobserved confounders—also holds in a less stylized setting. In each simulation, we compare the performance of various ATE estimators (described below) under two specifications: when conditioning on all available covariates, and when conditioning on all except a single variable $W_i$. Our goal is to find sets of parameters, if any, under which the magnitude of the bias is lower when $W_i$ is excluded from the specification.

## Experiment 1: baseline specification

In our baseline experiment, there are three covariates affecting treatment assignment and the outcome of interest. Two of these are observed, $X_i \sim N(1,1)$ and $W_i \sim N(1,1)$, while $Z_i \sim N(2,1)$ is unobserved to the analyst. In this set of simulations, the equations describing the data-generating process for the potential outcomes and the log odds of treatment assignment are linear in all covariates. The equations describing the data-generating process are as follows:

Outcome in treated state

$$Y_{i1} = \beta_{10} + \beta_{11}X_i + \beta_{12}Z_i + \beta_{13}W_i + \epsilon_{i1}$$

Outcome in untreated state

$$Y_{i0} = \beta_{00} + \beta_{01}X_i + \beta_{02}Z_i + \beta_{13}W_i + \epsilon_{i0}$$

Latent index function

$$T_i^* = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \gamma_3 W_i + u_i$$

Treatment indicator

$T_i = I(T_i^* > 0)$, where $I(\cdot)$ is the indicator function

There are three moving parts in this set of simulations. The first is the coefficient on $W_i$ in the propensity score equation, which varies between $-0.5$ and 0.5, $\gamma_3 \in \{-0.5, -0.25, 0, 0.25, 0.5\}$. The second is the coefficient on $W_i$ in the outcome equations, which varies between $-2$ and 2, $\beta_{13} \in \{-2, 0, 2\}$. The third is the canonical correlation between $X_i$ and $\mathbf{Z}_i = (W_i, Z_i)$, $cc_{XZ} \in \{0.2, 0.5, 0.8\}$. The sample size is $N = 1{,}000$ in each iteration. The error terms, $\epsilon_0$ and $\epsilon_1$, are normally distributed with mean 0 and variance 1 and are correlated at 0.25.[3] The error term $u$ in the propensity score equation has a logistic distribution. The intercept $\gamma_0$ is chosen so that 25% of the observations are in the treatment group. The remaining coefficients are set at reasonable values.[4] We performed 500 replications of the experiment per combination of the three moving parameters, $\gamma_3$, $\beta_{13}$, and $cc_{XZ}$. For each combination, we confirmed that the estimators we examine recover the treatment effect without bias when all three covariates are included in the propensity score specification. The goal of the experiment is to compare the bias of each estimator when only $X_i$ is included to the bias when both $X_i$ and $W_i$ are included.

[3] We also ran the experiment with $\epsilon_0$ and $\epsilon_1$ uncorrelated; the results were essentially identical.

[4] $\gamma_1 = 0.5$, $\gamma_2 = 1$, $\beta_{00} = 1$, $\beta_{01} = 1$, $\beta_{02} = 1$, $\beta_{10} = 2.5$, $\beta_{11} = 1.5$, and $\beta_{12} = 0.5$. Extensive robustness checks were performed; see the Results section.

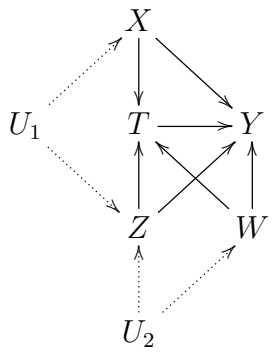



Figure 2: DGP 1. $W$ is related to both treatment and outcome.

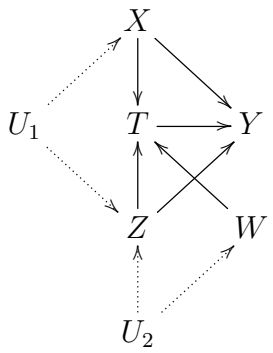



Figure 3: DGP 2. $W$ is related only to the treatment.

An easy way to understand the various data-generating processes (DGPs) used in the experiment is to look at the directed acyclic graphs (DAGs) in Figures 2–3. In each graph, an arrow indicates variables that affect one another; dashed arrows indicate unobserved variables. In Figure 2, $W_i$ affects both the treatment $T_i$ and the outcome $Y_i$. In Figure 3, $W_i$ is related only to the treatment, which corresponds to the case where $\beta_{13} = 0$. A graph representing the case where $\gamma_3 = 0$ would not have a solid arrow from $W_i$ to $T_i$.

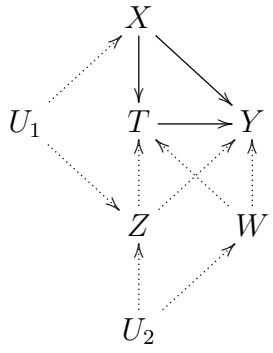



Figure 4: Misspecified model 1. $Z$ and $W$ are unobserved.

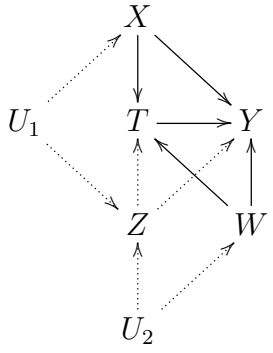



Figure 5: Misspecified model 2. $Z$ is unobserved, and $W$ is related to treatment and outcome.

The key to this experiment is the difference in the estimated ATE between the two misspecified models depicted in Figures 4 and 5. In the first misspecified model (Figure 4),

both $Z_i$ and $W_i$ are unobserved. In the second misspecified model (Figure 5), $W_i$ is included in the propensity score while $Z_i$ remains unobserved. The question is whether the bias on the ATE in the second misspecified model is ever greater than the bias on the ATE in the first. If so, then there are cases where controlling for all observed covariates is incorrect advice.

## Experiment 2: covariates from real data

In the previous experiment, all of the variables are drawn from a normal distribution—an ideal case for common matching methods (Rubin and Thomas, 1992), but one that hardly resembles a typical data set. To assess the relative performance of the propensity score estimators in a more realistic setting, we ran a second experiment using data from an experimental study. The data come from the well-known study of the impact of the National Support Work Demonstration (NSW) labor training program on post-intervention income LaLonde's (1986). [5] As in the prior experiments, the treatment variable $T_i$ and the potential outcomes $(Y_{i0}, Y_{i1})$ are simulated, as is the "new variable" $W_i$. However, all of the other observed and unobserved covariates are taken from LaLonde's data. The variables we use are *Age* (in years), *Education* (in years), ln(1+*Earnings 1974*) (in logged USD), and ln(1+*Earnings 1975*) (in logged USD). The last of these is assumed to be unobserved to the analyst, and thus always omitted from the propensity score specification, while the other three are always included. In each trial, we generate a variable $W_i$ (mean 1, variance 1) that is correlated with the omitted variable, ln(1+*Earnings 1975*), at the level $\rho \in \{-0.75, -0.25, 0, 0.25, 0.75\}$. As in the previous experiments, we compare the bias in the estimated ATE when all observed covariates are included in the propensity score specification to the bias when all but $W_i$ are included. The data-generating process is as follows:

[5] Following Dehejia and Wahba (1999), we use the subset of male participants for which 1974 earnings are available and the original experimental control units, giving us 445 observations.

Outcome in treated state

$$Y_{i1} = 2 + 0.1\, Age_i + 0.15\, Educ_i + 0.4\, \ln(1 + Earn74)_i + 0.4\, \ln(1 + Earn75)_i + \beta W_i + \epsilon_{i1}$$

Outcome in untreated state

$$Y_{i0} = 1 + 0.05\, Age_i + 0.1\, Educ_i + 0.2\, \ln(1 + Earn74)_i + 0.2\, \ln(1 + Earn75)_i + \beta W_i + \epsilon_{i0}$$

Latent index function

$$T_i^* = \gamma_0 + 0.03\, Age_i + 0.14\, Educ_i + 0.12\, \ln(1 + Earn74)_i + 0.12\, \ln(1 + Earn75)_i + \gamma_1 W_i + u_i$$

Treatment indicator

$T_i = I(T_i^* > 0)$, where $I(\cdot)$ is the indicator function

The other moving parts are once again the coefficient of $W_i$ in the true propensity score, $\gamma_1 \in \{-0.3, -0.15, 0, 0.15, 0.3\}$, and its coefficient in the outcome equation for treated units, $\beta \in \{-0.5, -0.25, 0, 0.25, 0.5\}$. The intercept in the propensity score equation, $\gamma_0$, is chosen so that 25% of observations are treated. The error terms in the outcome equations, $\epsilon_{i0}$ and $\epsilon_{i1}$, have a standard deviation of 3 and are correlated at 0.25. As in the prior experiment, our interest focuses on whether including $W_i$ in the propensity score specification decreases the bias and mean squared error of various estimators of the average treatment effect.

## Techniques

Because there are continuous covariates in each experiment, we focus on propensity score estimators of the average treatment effect. Each of the five estimators that we consider involves obtaining a first-stage estimate of the propensity score, $e(\mathbf{X}_i) = \Pr(T_i = 1 \mid \mathbf{X}_i)$, where $\mathbf{X}_i$ denotes the full set of confounding variables. In each Monte Carlo iteration, we obtain two distinct propensity score estimates, both using the fitted values from a logistic regression of $T_i$ on observed covariates. The first estimate, $\hat{e}_{XWi}$, is obtained from a specification that

includes all of the observed covariates. For the first experiment, this is

$$\hat{e}_{XWi} = \Lambda(\hat{\gamma}_0^{XW} + \hat{\gamma}_1^{XW} X_i + \hat{\gamma}_2^{XW} W_i),$$

where $\hat{\gamma}^{XW}$ represents the logistic regression coefficients. The second estimate, $\hat{e}_{Xi}$, comes from a specification that includes all of the observed variables except $W_i$. For the first experiment, this is

$$\hat{e}_{Xi} = \Lambda(\hat{\gamma}_0^{X} + \hat{\gamma}_1^{X} X_i).$$

After obtaining the two sets of propensity score estimates, we plug them into each of the following ATE estimators:[6]

- Nearest-neighbor matching (with replacement): In this method, all the units are ordered randomly, and the first treated unit is matched with the control unit having the nearest propensity score. The first treated unit is then removed from the data set while its matched control unit is kept to be used in future matches. The process is repeated for the second treated unit and so on. The causal effect is estimated by averaging the outcome differences between the matched treatment and control groups. Following Rosenbaum and Rubin (1985, 36), we use the linear predictor in place of the estimated probability to avoid compression of the propensity scores near 0 and 1.

- Caliper matching (with replacement): In this method, all the units are ordered randomly, and for the first treated unit, the control units with propensity scores (again the linear predictor) within a specified distance of the treated unit are gathered (0.25 standard deviations of the linear predictor), and the treated unit is matched with the closest control unit within the group in terms of Mahalanobis distance. The first treated unit is then removed from the data set while its matched control unit is kept to

[6] We implement the first two estimators using Sekhon's (2011) R package `Matching`.

be used in future matches. The process is repeated for the second treated unit and so on. Again, the causal effect is estimated by averaging the outcome differences between the matched treatment and control groups.

- Blocked matching: In this method, all the observations are divided into blocks, or strata, based on the value of the propensity score, which should be approximately constant within the strata. (We use deciles.) The difference of means between the treated and the control is estimated within each block, and the estimated causal effect is the weighted mean of these differences.

- Weighting: Wooldridge (2002, 616) provides a consistent estimator of the ATE based on simple weighting that is identical to the Horvitz-Thompson estimator (Horvitz and Thompson, 1952),

$$\begin{aligned} \hat{\tau} &= N^{-1} \sum_{i=1}^{N} \left\{ \frac{[T_i - \hat{e}_i] Y_i}{\hat{e}_i (1 - \hat{e}_i)] } \right\} \\ &= N^{-1} \sum_{i=1}^{N} \left\{ \frac{T_i Y_i}{\hat{e}_i} - \frac{(1 - T_i) Y_i}{1 - \hat{e}_i} \right\} \end{aligned}$$

  where $\hat{e}_i$ is the estimated propensity score.

- Covariance adjustment: In this method, the ATE is estimated from a regression of the response variable on a constant, a variable denoting treatment assignment, the estimated propensity score, and a multiplicative term comprising the treatment variable and deviations about the sample mean of the estimated propensity score. The regression equation is

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 \hat{e}_i + \beta_3 T_i \left[ \hat{e}_i - \hat{\mu}_e \right] + \epsilon_i,$$

  where $\hat{\mu}_e$ is the sample average of the estimated propensity score, $\hat{e}_i$. Rosenbaum and

Rubin (1983, 46) demonstrate, assuming that $E[Y_{i0} \mid e(\mathbf{X}_i)]$ and $E[Y_{i1} \mid e(\mathbf{X}_i)]$ are linear in $e(\mathbf{X}_i)$, that $\hat{\beta}_3$ is a consistent estimator of the ATE:

In each Monte Carlo iteration, we yield ten estimates of the ATE: one for each method using the propensity score specification with all observed covariates, and one for each method using the propensity score specification from which $W_i$ is excluded. We write $\hat{\tau}_{XW}$ to denote an estimator that uses the propensity score specification with $W_i$ included and $\hat{\tau}_X$ for one that uses the specification where $W_i$ is excluded. Our interest is in whether there are parameter combinations under which $\hat{\tau}_{XW}$ is more biased than $\hat{\tau}_X$ (for some or all of the five particular procedures), meaning it is better not to condition on all observed covariates.

# 4 Results

## Experiment 1: baseline specification

Figure 6 summarizes the main findings.[7] Each point in a subplot represents the difference in absolute bias on the estimated ATE between the model that included $W$ in the propensity score and the model that did not, as a function of $\gamma_3$ ($W$'s coefficient in the true propensity score equation). Each subplot represents one combination of the other varying parameters: $cc_{XZ}$, the canonical correlation between $\mathbf{X}$ and $\mathbf{Z}$, and $\beta_{13}$, the coefficient on $W$ in the outcome equation. Each line within a subplot represents the quantity of interest estimated with a specific method. Positive values indicate that the absolute value of the bias is greater when $W$ is included in the estimated propensity score equation than when it is not. For example, the subplot in the top left corner presents the difference in absolute bias when $\beta_{13} = -2$ and $cc_{XZ} = 0.2$. In this example, intermediate values of $\gamma_3$ result in the bias increasing when $W$ is included in the estimated propensity score, regardless of which method

[7] A replication file is available upon request.

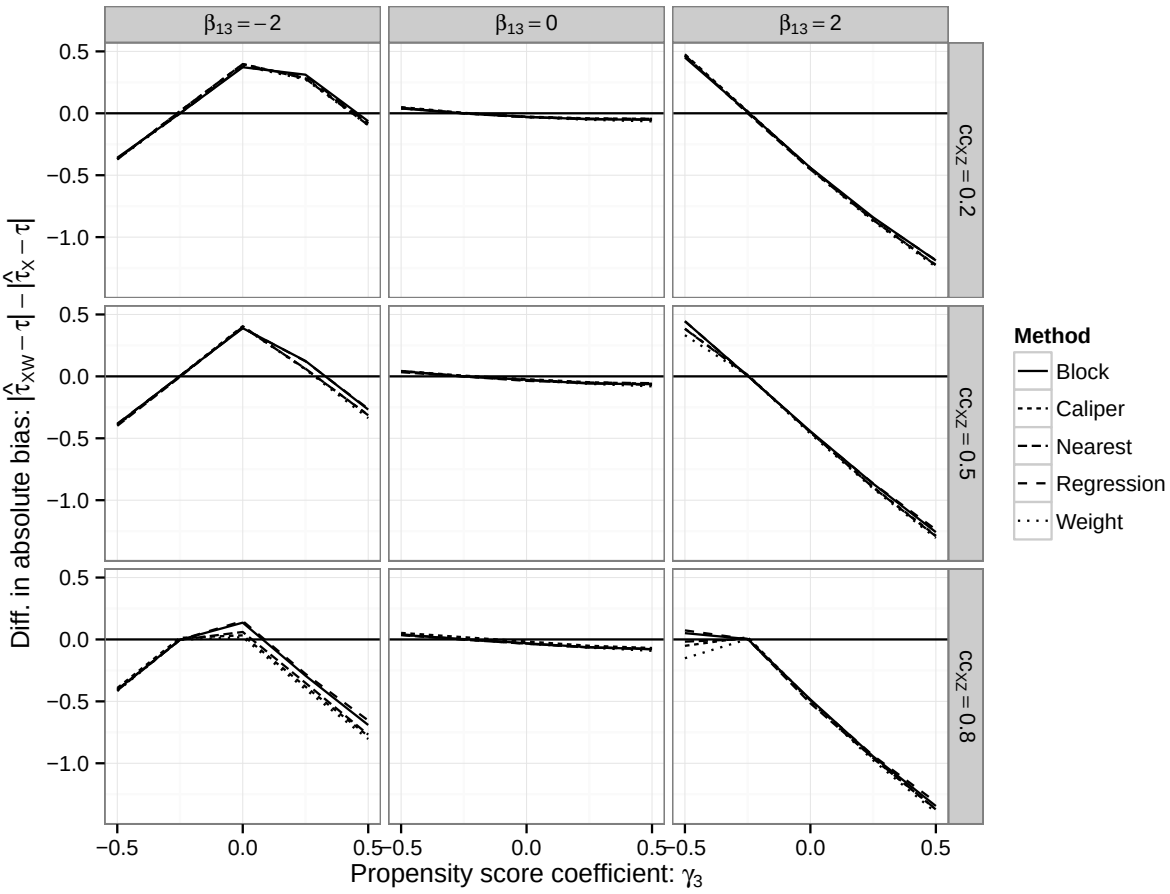
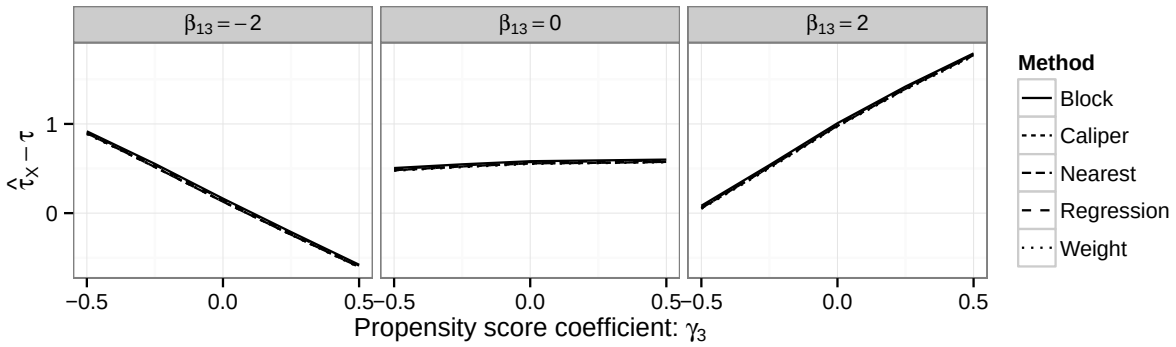


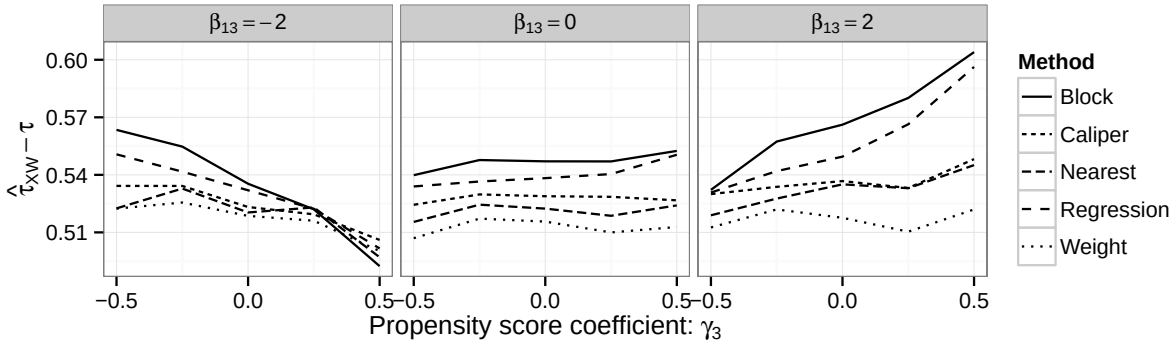

Figure 6: Difference in absolute bias with linear outcome equations.

is used to estimate the ATE. We ran a similar set of experiments in which the estimand was the average treatment effect on the treated (ATT), and obtained substantively identical results.

Two observations are apparent. First, the absolute bias on the estimated ATE increases when $W$ is included in the estimated propensity score equation for several parameter combinations. That is, it is *not* always optimal to condition on all available pre-treatment variables. Specifically, we find that when $\beta_{13}$ is positive but $\gamma_3$ is equal to $-0.5$, it is often worse to include $W$ in the propensity score. The same is true if $W$'s effect on the outcome is negative but its effect on the probability of treatment is weakly positive ($\gamma_3 = 0.25$). This




(a) Bias when $W$ is excluded.




(b) Bias when $W$ is included.

Figure 7: Biases in the experiment with linear outcome equations (the canonical correlation is held at 0.2, $cc_{XZ} = 0.2$).

pattern is stronger when the canonical correlation between $X$ and $Z$ is low. Notice also that for all values of the canonical correlation, whenever there is no effect of $W$ on the outcome, we see that there is not much difference between the bias when the propensity score model includes the variable and when it does not. If anything, when $\gamma_3$ is negative and $\beta_{13} = 0$, it is slightly worse to include the variable.

The second observation is that the five estimation procedures generally agree regarding whether including $W$ in the propensity score equation worsens the bias on the ATE. Figure 7 shows the bias when $W$ is excluded from the estimated propensity score and when it is not for a low canonical correlation between $X$ and $Z$. When $W$ is not included, all five methods

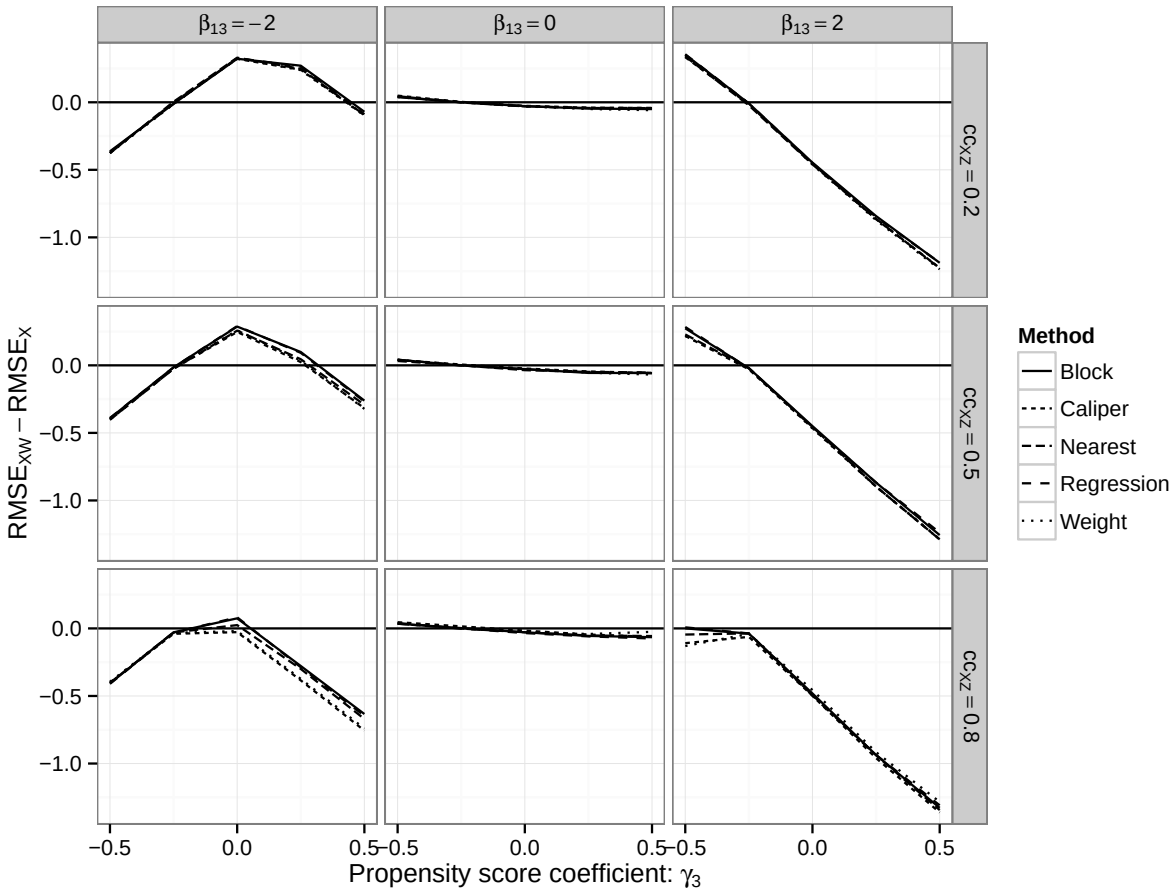



Figure 8: Difference in root mean squared error with linear outcome equations.

return nearly identical estimates but the variation between methods is greater when $W$ is included.[8] It appears, that weighting and nearest neighbor matching perform slightly better at reducing the bias when $W$ is included; however, as we will see later, this finding is not robust to changes implemented in other experiments.

We can also see how the inclusion vs. exclusion of $W$ from the propensity score equation affects the root mean squared error of the ATE estimates.[9] Figure 8 presents plots for the

[8]This is true accounting for the fact that the figure has different $y$ scales for both specifications of the propensity score.

[9]The estimate of the root mean squared error $RMSE_k$ with $k = X, XW$ is $\sqrt{\frac{\sum_{j=1}^{500}(\widehat{\tau}_{k,j}-\tau)^2}{500}}$, where $\widehat{\tau}_{k,j}$ is the estimate of the average treatment effect for a given parameter combination in one replication of the experiment.

difference in the root mean squared error (RMSE) of these experiments. The plots keep the same basic structure of the ones in Figure 6. We find similar results. When the coefficient of $W$ in the treatment and outcome equations have opposite signs, the treatment effect is weak, and the canonical correlation is low, it is worse to include $W$. In addition, if the outcome effect is null and the treatment effect is strong and negative, including $W$ slightly increases the RMSE.

If the magnitude of the increment in bias or RMSE were negligible for the conditions that we have characterized, we could be confident that adding all available pre-treatment variables would not seriously affect the conclusions derived by the results of our empirical analysis. Unfortunately, this is not always the case. If we consider the parameter combination of $\beta_{13} = 2$, $\gamma_3 = -0.5$ and $cc_{XZ} = 0.2$, the bias when $W$ is included is approximately 0.53, compared to 0.05 when it is excluded. Including $W$ increases by more than 10 times the bias for this parameter combination. Moreover, the increment in the bias represents 16% of the total ATE (which is 3 for this parameter combination).

To confirm that the relationships between the sign and magnitude of the coefficients in the propensity and outcome equations are indeed driving the results, we ran a number of additional Monte Carlo experiments. We ran the baseline model with each of the following changes (separately):

- Coefficient on $Z$ in the propensity score equation reversed to $-1$;
- Coefficients on $Z$ in the outcome equations, $Y_0$ and $Y_1$, reversed to $-1$ and $-0.5$ respectively;
- Correlation between $W$ and $Z$ reversed to $-0.25$;
- Inclusion of a quadratic term of $Z$ and an interaction of $Z$ and $X$ in outcome equations.

In the first two cases, the results were nearly a mirror image of those shown in Figure 6:

adding the new variable $W$ increased the bias mainly in cases where it had the same effect on the treatment assignment and outcome.[10] In the third case, with the correlation between $Z$ and $W$ reversed, the results were substantively the same as in the original experiment. Finally, when we added an interaction and quadratic terms to the outcome equations, we found the same basic pattern (including $W$ is worse in terms of bias when the signs of $\beta_{13}$ and $\gamma_3$ are different). This time, however, the result holds regardless of the canonical correlation.

Our findings are summarized in Table 2, which describes the conditions or patterns of signs of effects under which adding a variable $W$ to the propensity score equation increases the bias in the ATE when another variable $Z$ is unavailable.

| Pattern | $\gamma_2 \cdot \beta_{12}$ | $\gamma_3 \cdot \beta_{13}$ | Including W increases bias? |
|---|---|---|---|
| 1 | + | + | no |
| 2 | + | - | yes |
| 3 | - | + | yes |
| 4 | - | - | no |

Table 2: When does including a new variable $W$ in the propensity score equation increase the bias of the estimated average treatment effect? ($\gamma_2$ and $\beta_{12}$ are the coefficients on $Z$ in the propensity and outcome equations, respectively. $\gamma_3$ and $\beta_{13}$ are the coefficients on $W$ in the propensity and outcome equations, respectively.)

The cases where including $W$ in the estimated propensity score equation increases the bias are those where balancing on $W$ and balancing on $Z$ have countervailing effects (patterns 2 and 3 in Table 2), and the effect of $Z$'s confounding is stronger than that of $W$'s. In this experiment, $Z$ has a positive effect on the probability of treatment and a positive direct effect on the outcome, meaning the ATE would be overstated if only $Z$ were omitted. If $W$ has a negative effect on treatment assignment but a positive direct effect on the outcome, then $W$'s omission causes the ATE to be understated. In this situation, balancing on $Z$ causes the estimated ATE to decrease, and balancing on $W$ causes an increase. Suppose the

[10] In particular, these were $(\gamma_3, \beta_{13}) = (-0.5, -2)$ and $(0.5, 2)$, whereas in the original model the cases where including $W$ increased bias were $(0.5, -2)$ and $(-0.5, 2)$.

confounding effect of $Z$ is larger than that of $W$, so the estimated ATE is too high when both are omitted. Including $W$ in the propensity score specification would further increase the estimated ATE, exacerbating the bias due to the omission of $Z$.[11] This is precisely what we observe in Figure 6 when $\beta_{13} = 2$ and $\gamma_3 = -0.5$.

The patterns that we have consistently found in which adding a relevant pre-treatment variable increases the bias of the ATE could easily occur in empirical applications. Imagine a get-out-the-vote effort ($T$) targeted at swing voters (e.g., those registered as independents or who have not donated to campaigns). A high level of partisanship ($W$) decreases one's probability of receiving the treatment, but strong partisans are going to be more likely to vote ($Y$). At the same time, higher-income ($Z$) individuals are both more likely to receive the treatment and to vote. This case corresponds to pattern 2 in Table 2: Partisanship has opposite effects on treatment and outcome, while income has the same effect on both, so controlling for partisanship when income data are unavailable will likely increase the bias.

## Experiment 2: real data

Results from the second experiment, in which most of the covariates are taken from LaLonde's data rather than being simulated, are summarized in Figure 9. These results are similar to those of the first two experiments, though the pattern identified in Table 2 is somewhat weaker. For example, take the fourth column of Figure 9, where $\beta = 0.25$ (i.e., the new variable has a positive effect on the outcome). When the correlation between $W$ and the omitted variable ($\ln(1+Earn75)$) is 0.25, including $W$ in the estimated propensity score equation increases the bias when $W$'s true effect on receiving the treatment is negative. Since the omitted variable has a positive effect on both assignment and outcome, this corresponds to the second row of Table 2. There are some anomalies when the magnitude of the correlation

[11] The only exception is if $W$ and $Z$ are so strongly positively correlated that balancing on $W$ substantially improves balance on $Z$.

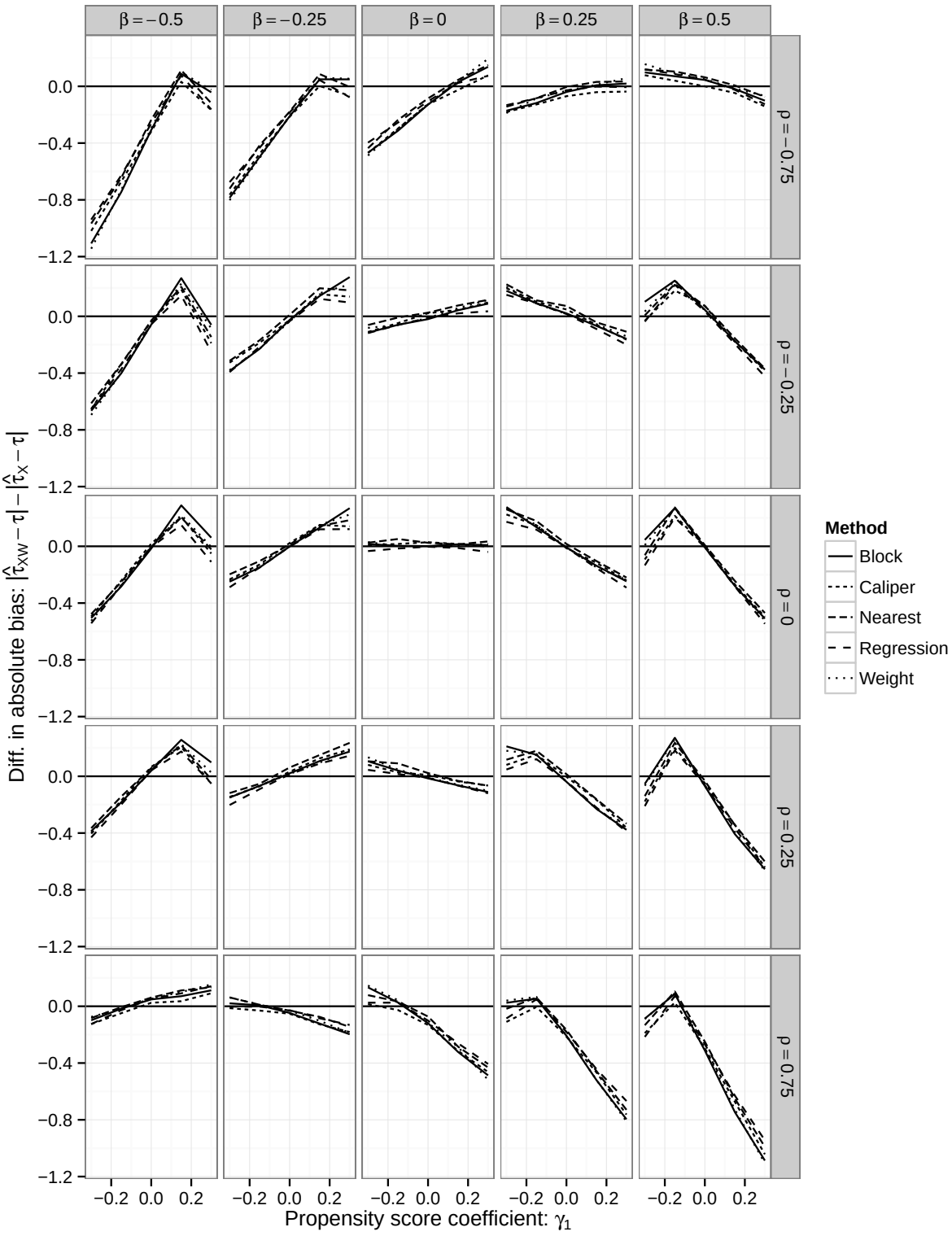



Figure 9: Difference in absolute bias with regressors taken from the LaLonde (1986) data.

between $W$ and $Z$ is high. For example, when $\beta = -0.25$, inclusion of $W$ in the propensity score specification when $\gamma_1$ is positive exacerbates bias for all $\rho$ except 0.75. In this case, conditioning on $W$ increases balance on $Z$ enough to offset any potential exacerbation of the bias.

The differences between estimation methods are still relatively small in this experiment, but they are more noticeable than in the first two experiments. We find that including $W$ in the propensity score equation is less likely to increase bias under caliper matching than under other methods. Covariance adjustment (regression on the propensity score) also seems to perform well with the LaLonde data, which was not the case in the fully simulated experiments. When using covariance adjustment to estimate the ATE, including $W$ in the estimated propensity score equation increases the magnitude of bias in 40 cases, out of 125 total combinations of the parameters ($\beta$, $\gamma$, and $\rho$). That figure increases to 58 cases under nearest-neighbor matching; in between are caliper matching (46), weighting (53), and blocking (55).

## Balance tests

The previous experiments identified situations where conditioning on all available pre-treatment variables could lead to an increase in bias on the estimated ATE (or ATT). The results show that when considering whether to include a pre-treatment variable in the propensity score equation a researcher should take into account the potential effect of unobservables on treatment and outcomes, as well as their relation with the variable in question. At this point it is natural to ask whether a post-matching balance test on treated and untreated units could be thought of as an alternative way to identify whether a covariate should be included in the propensity score equation. If balance on matched units improves once the variable is included in the propensity score, the inclusion might be justified. Improving balance on observed variables does not necessarily mean reducing bias when there are unobserved con-

founders. Adopting this practice could lead to worse estimation results. Using the LaLonde data, we show that it is not uncommon to find situations where a post-matching test indicates that a candidate variable should be included, when in fact its inclusion worsens the bias. What this result suggests is that balance tests are not substitutes for careful consideration of the potential effects of unobservables when choosing a specification.

For the following exercises, we use LaLonde's subset of treated units and the CPS (Current Population Survey) control individuals (LaLonde, 1986).[12] This data set has been used to evaluate the performance of treatment effect estimators on observational data by comparing them with an experimental benchmark. We are interested in finding pairs of variables from this data set that would play the role of $Z$ and $W$ in our previous experiments and that simultaneously satisfy three conditions: 1) They have one of the countervailing effects identified in Table 2, 2) balance seems to improve with $W$'s inclusion in the propensity score according to a post-matching balance test, and 3) the estimated ATT is more biased when W is included. Identifying such cases gives evidence of the risks of relying exclusively on balance tests, and the potential benefit of using the identified countervailing patterns in justifying the inclusion of additional variables on the propensity score. Given that we are using observational data, we do not know the true effects of $W$ and $Z$ on the outcome and treatment. We generate estimates of these effects by relying on the propensity score specification that in Dehejia and Wahba (1999) gave the authors the closest ATT estimate to the experimental benchmark when using the CPS control individuals.[13]

To further clarify the exercise, consider the pair of variables $re75$ (real earnings in 1975) and *nodegree* (indicator variable of not possessing a degree) and suppose that the first one

[12]The data set is included in the R package of Random Recursive Partitioning (Iacus, 2007). It has a total of 16177 observations with no missingness on the variables from the original study. For a description of the data set see Dehejia and Wahba (1999, 1054)

[13]The specification includes the following variables: *age, age*$^2$*, education, education*$^2$*, no degree, married, black, hispanic, real earnings in 1974 and 1975 (re74 and re75), indicators of zero earnings in 1974 and 1975 (u74 and u75), education* $\times$ $re74$ *and age*$^3$.

takes the role of $Z$ (the unobserved variable) and the second one is the candidate variable to be included $W$. This would represent a situation where a researcher interested in finding the effect of the training program on income is considering to include the indicator of no degree in the propensity score equation when past earnings are not available. Following Dehejia and Wahba's specification that gives the best estimate of the ATT with this data, it is found that the effect of $re75$ and *nodegree* on income are both positive and that the effect of $re75$ on treatment assignment is negative while this same effect for *nodegree* is positive. This example follows pattern 3 in Table 2. For this specific case, and using as a benchmark the experimental ATT, we calculate that the bias after including *nodegree* (while leaving out $re75$) increases by 150.71, which is approximately 8.5% of the experimental ATT.[14]

We also compared the results of a balance test after matching when *nodegree* was not included in the treatment equation with the results of the same test when it was. We found that after including it, the p-value of a t-test between the treated and untreated matched units increased for 6 variables (out of a total of 11 covariates) compared to the p-value of the same test when it was left out of the propensity score. This result suggests that in more than half of the rest of the covariates, balance seems to have improved after matching when *nodegree* was included. In this example, a researcher expecting a positive relation between past earnings and income, could infer with the help of Table 2 that the inclusion of this variable could bring an increase in bias if there is a positive relation between *nodegree* and income once other variables are controled for in the outcome equation.

The case of $re75$ and *nodegree* is not the only one. Table 3 shows 14 other cases where a countervailing effect is present and including $W$ improves balance for a given number of covariates, but increases bias. These results are not definitive as we have relied on a particular specification from Dehejia and Wahba to find estimates of the true effects on income and

[14]The ATT was calculated using caliper matching using the same estimation setup as in the Monte Carlo experiments.

participation on the program. However, given that the bias increased in the situations where we expected it to happen, based on our Monte Carlo findings, we can be more confident that the results are not unique to this data set or this particular specification.

| Countervailing pattern | Z | W | $\Delta$ Bias | $\frac{\text{Vars. p-val increased}}{\text{Total}}$ |
|---|---|---|---|---|
| 2 | *nodegree* | $age^2$ | 27.13 | 7/12 |
| 2 | *education* | *age* | 171.64 | 4/10 |
| 3 | *black* | *nodegree* | 112.75 | 8/12 |
| 3 | *black* | *u*74 | 690.72 | 7/12 |
| 3 | *u*75 | *nodegree* | 150.70 | 6/11 |
| 3 | *age* | *education* | 104.99 | 8/10 |
| 3 | *age* | *u*74 | 94.37 | 8/10 |
| 3 | $age^2$ | *education* | 104.99 | 8/10 |
| 3 | $age^2$ | *u*74 | 94.37 | 8/10 |
| 3 | *black* | *education* | 11.85 | 8/12 |
| 3 | *married* | *nodegree* | 107.37 | 9/12 |
| 3 | *married* | *u*74 | 24.87 | 9/12 |
| 3 | *re*75 | *nodegree* | 150.70 | 6/11 |
| 3 | $education^3$ | *education* | 104.99 | 8/10 |
| 3 | $education^3$ | *u*74 | 94.37 | 8/10 |

Notes: The column 'Countervailing patterns' refers to the patterns defined in Table 2. $\frac{\text{Vars. p-val increased}}{\text{Total}}$ is the number of control variables that had an increase in the p-value of the t-test after matching when $W$ was included over the number of common controls between the specifications with and without $W$.

Table 3: Balance test and propensity score selection of variables

## Sensitivity analysis

We ran additional simulations to determine whether researchers can use sensitivity analysis to inform their decisions on whether a covariate should be included in the propensity score. The logic of using sensitivity analysis in this situation is as follows: If after including a variable in the propensity score the results of the ATE estimation are less sensitive to unobserved factors than when the variable is left out, this could be used to justify its inclusion.

We use Rosenbaum's (2002) method for calculating bounds on the ATE estimate when the

log odds of treatment assignment differ by up to $\Gamma$ due to unobserved confounding.[15] Using the same parameters as in the simulations described above (with both the fully simulated data and the LaLonde data), we calculated the average lowest level of $\Gamma$ at which the bounds on the treatment effect contained 0, under both inclusion and exclusion of $W$ from the propensity score equation. If the level of $\Gamma$ was larger when including $W$ than when it was left out, this would suggest that the results are less sensitive to unobservables, and that therefore $W$ should be added to the propensity score. However, what we found is that in all cases, this level is almost entirely determined by the magnitude of the estimated ATE and that there are no significant differences between including and excluding $W$ on the calculated $\Gamma$. Therefore, this kind of sensitivity analysis is not useful for determining whether it is worse to include an observed pre-treatment variable because of its interaction with an unobserved confounder.

# 5 Discussion

We investigate claims made by both sides in recent debates regarding conditioning and matching using propensity scores. The results of our experiments indicate that conditioning on all available pre-treatment variables is not always optimal. In every case, the researcher must consider the effects of unobserved pre-treatment variables and their relationships with observed pre-treatment variables. Whether conditioning on an additional observed pre-treatment variable increases or decreases the bias on the ATE depends on these relationships. Specifically, in the linear case, we show that when the newly included covariate has a positive effect on the outcome and a negative effect on the propensity (and when there is an unobserved covariate whose effects on the outcome and treatment have the same sign), it is often worse to include the covariate. This basic pattern also holds in nonlinear specifications

[15] To calculate the bounds, we used the R package `rbounds` (Keele, 2011).

and in simulations using real data.

We have yet to address how researchers can best make use of our findings. Our results suggest that researchers cannot rely on advice such as condition on all pre-treatment covariates or on balance and sensitivity tests. Some progress can be made if we consider the two kinds of unobserved covariates that plague empirical analyses. To paraphrase Donald Rumsfeld (Morris, 2010), there are known unknowns and unknown unknowns. That is, there are covariates, perhaps suggested by theory, that cannot be measured or perhaps measurement is infeasible. These are the known unknown covariates. A researcher can hypothesize about the relationships of such a covariate with previously included variables and any variables that are candidates for inclusion. Our results provide some guidance in such a situation. If the candidate covariate and the unobserved covariate have countervailing effects, a case can be made for leaving the candidate covariate unadjusted.

On the other hand, there exist, in Rumsfeldian terms, unknown unknown covariates. These are variables that have not been suggested by theory and have not crossed the mind of the researcher in question (or anyone else). In such a case, no theorizing can take place, and our results demonstrate that including a new covariate in a propensity score equation may increase or decrease the bias on the estimated ATE. Sensitivity analysis that explicitly takes unobserved covariates into account, e.g. Rosenbaum (2002), seems to be of little use. The only surefire response a researcher has to the problem discussed in this paper is to be modest in the claims she makes based on her results. Scientific progress is rarely the result of a single study, and empirical generalizations are accepted only after many repeated demonstrations across varying spatial and temporal domains.

# Appendix

The debate takes place within the context of the potential outcomes framework. Let $Y_{i1}$ be the value of the response variable when unit $i$ receives the treatment ($T_i = 1$), and let $Y_{i0}$ be the value of the response variable when unit $i$ does not receive the treatment ($T_i = 0$). $Y_{i1}$ and $Y_{i0}$ are potential outcomes as they cannot be observed simultaneously for unit $i$. The observed outcome is $Y_i = T_i Y_{i1} + (1 - T_i)Y_{i0}$, and the effect of the treatment for unit $i$ is $\tau_i = Y_{i1} - Y_{i0}$. As the individual-level causal effects generally cannot be estimated, interest centers on the ATE:

$$E[\tau_i] = E[Y_{i1} - Y_{i0}].^{16} \tag{2}$$

Equation (1) is estimable under the the stable unit treatment value assumption (SUTVA), which states if unit $i$ receives treatment $j$, the observed value of $Y$ is $Y_{ij}$ (no interference between units and no variation in treatment) (Rubin, 1980, 591).

In observational studies, focus moves to the ATE conditional on a set of observed pre-treatment covariates $\mathbf{X}$:

$$\tau = E[Y_i|\mathbf{X}_i, T_i = 1] - E[Y_i|\mathbf{X}_i, T_i = 0]. \tag{3}$$

Equation (3) can be used to consistently estimate the ATE when treatment is said to be strongly ignorable. That is, the potential outcomes and the treatment must be independent within levels of the covariates,

$$\{Y_{i1}, Y_{i0}\} \perp\!\!\!\perp T_i|\mathbf{X}_i, \tag{4}$$

and for every value of $\mathbf{X}$ there are treated and nontreated cases, $0 < \Pr(T_i = 1|\mathbf{X}_i) < 1$, for

[16] We will also be interested in the average treatment effect on the treated (ATT): $E[Y_{i1} - Y_{i0}|T_i = 1]$.

all $\mathbf{X}_i$. If unobserved confounding variables exist that are not included in $\mathbf{X}$, as is likely in an observational study, any estimator of the ATE will be biased.

Rosenbaum and Rubin (1983) show that if the set of covariates is of high dimension, thereby creating a problem for techniques such as matching, one can condition on the propensity score, which is the probability of assignment to treatment, conditional on the set of covariates, $e(\mathbf{X}_i) = \Pr(T_i = 1|\mathbf{X}_i)$. Free of behavioral assumptions, the propensity score is generally estimated with a simple logit model:

$$\Pr(T_i = 1|\mathbf{X}_i) = \frac{e^{\boldsymbol{\beta} h(\mathbf{X}_i)}}{1 + e^{\boldsymbol{\beta} h(\mathbf{X}_i)}},$$

where $h(\mathbf{X})$ comprises linear and higher order terms of the pretreatment covariates, and $\boldsymbol{\beta}$ is the set of parameters to be estimated.

# References


Augurzky, Boris and Christoph M. Schmidt. 2001. The Propensity Score: A Means to An End. Technical Report 271 The Institute for the Study of Labor Bonn, Germany: .

Brookhard, M. Alan, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn and Til Stürmer. 2006. "Variable selection for propensity score models." *American Journal of Epidemiology* 163(12):1149–1156.

Clarke, Kevin A. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22(4):341–352.

Clarke, Kevin A. 2009. "Return of the Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 26(1):46–66.

Dehejia, R. and S. Wahba. 1999. "Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448):1053–1062.

Drake, Christiana. 1993. "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effects." *Biometrics* 49(4):1231–1236.

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(260):663–685.

Iacus, S.M. 2007. *rrp: Random Recursive Partitioning.* R package version 0.7.
**URL:** *http://cran.r-project.org/web/packages/rrp/*

Keele, Luke J. 2011. *rbounds: Perform Rosenbaum bounds sensitivity tests for matched data.* R package version 0.7.
**URL:** *http://CRAN.R-project.org/package=rbounds*

LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4):604–620.

Luca, Guiseppe De, Jan Magnus and Franco Peracchi. 2015. On the ambiguous consequences of omitting variables. Technical report Einaudi Institute for Economics and Finance Italy: .

Millimet, Daniel L. and Rusty Tchernis. 2009. "On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies." *Journal of Business and Economic Statistics* 27(3):397–415.

Morris, Errol. 2010. "The Anosognosic's Dilemma: Something's Wrong but You'll Never Know What It Is (Part 1)." *The New York Times* .

Pearl, Judea. 2009. "Remarks on the method of propensity score." *Statistics in Medicine* 28(9):1415–1416.

Rosenbaum, Paul R. 2002. *Observational studies*. New York: Springer.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effect." *Biometrika* 70(1):41–55.

Rosenbaum, Paul R. and Donald B. Rubin. 1985. "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score." *The American Statistician* 39(1):33–38.

Rubin, Donald B. 1980. "Randomization Analysis of Experimental Data: The Fisher Randomization Test: Comment." *Journal of the American Statistical Association* 75(371):591–593.

Rubin, Donald B. 2007. "The design *versus* the analysis of observational studies for causal effects: parallels with the design of randomized trials." *Statistics in Medicine* 26(1):78–97.

Rubin, Donald B. 2009. "Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups?" *Statistics in Medicine* 28(9):1420–1423.

Rubin, Donald B. and Neal Thomas. 1992. "Affinely Invariant Matching Methods with Ellipsoideal Distributions." *Annals of Statistics* 20(2):1079–1093.

Sekhon, Jasjeet S. 2011. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R." *Journal of Statistical Software* 42(7):1–52.

Shrier, Ian. 2008. "Letter to the editor." *Statistics in Medicine* 27(14):2740–2741.

Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data.* Cambridge, MA: The MIT Press.