NBER TECHNICAL WORKING PAPER SERIES

DO INSTRUMENTAL VARIABLES BELONG IN PROPENSITY SCORES?

Jay Bhattacharya
William B. Vogt

Technical Working Paper 343
http://www.nber.org/papers/t0343

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2007

We thank Paul Rosenbaum, Frank Wolak, Salvador Navarro-Lozano, Azeem Shaikh, and Ed Vytlacil for helpful comments. Bhattacharya thanks the National Science Foundation and the National Institute on Aging for partially funding his work on this paper. All errors remain our own. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Do Instrumental Variables Belong in Propensity Scores?
Jay Bhattacharya and William B. Vogt
NBER Technical Working Paper No. 343
September 2007, Revised September 2009
JEL No. C1,I1,I2

ABSTRACT

Propensity score matching is a popular way to make causal inferences about a binary treatment in observational data. The validity of these methods depends on which variables are used to predict the propensity score. We ask: "Absent strong ignorability, what would be the effect of including an instrumental variable in the predictor set of a propensity score matching estimator?" In the case of linear adjustment, using an instrumental variable as a predictor variable for the propensity score yields greater inconsistency than the naive estimator. This additional inconsistency is increasing in the predictive power of the instrument. In the case of stratification, with a strong instrument, propensity score matching yields greater inconsistency than the naive estimator. Since the propensity score matching estimator with the instrument in the predictor set is both more biased and more variable than the naive estimator, it is conceivable that the confidence intervals for the matching estimator would have greater coverage rates. In a Monte Carlo simulation, we show that this need not be the case. Our results are further illustrated with two empirical examples: one, the Tennessee STAR experiment, with a strong instrument and the other, the Connors' (1996) Swan-Ganz catheterization dataset, with a weak instrument.

Jay Bhattacharya
117 Encina Commons
Center for Primary Care
and Outcomes Research
Stanford University
Stanford, CA 94305-6019
and NBER
jay@stanford.edu

William B. Vogt
Senior Economist
RAND Corporation
4570 Fifth Ave
Pittsburgh, PA 15213
and NBER
william.b.vogt@gmail.com

1 Introduction

Propensity score matching is a popular way to make causal inferences about a binary treatment in observational data. These methods seek to create the observable covariate balance which randomization creates in a randomized controlled trial. Rosenbaum and Rubin (1983) demonstrate that if treatment assignment is strongly ignorable given observed covariates then a consistent estimate of the average treatment effect may be obtained via matching on, sub-classifying on, or covariance adjusting for the propensity score.

Key to propensity score-based methods is the decision of which variables to use in the predictor set for the propensity score. As a practical matter, predictor variable selection for propensity scores seems to be guided most often by some measure of goodness-of-fit of the propensity score to the treatment assignment (Weitzen et al., 2005, 2004; Hirano and Imbens, 2001).

Under the maintained hypothesis of strong ignorability, omitting a relevant variable from the construction of a propensity score leads to inconsistency; whereas, the inclusion of an irrelevant variable leads only to greater variance of the estimator, leading some to favor generous inclusion of variables in predictor sets.

In an article published in the Journal of the American Medical Association (JAMA), a premier medical journal, bio-statisticians D'Agostino and D'Agostino provide the following guidance on the choice of covariates:

> [T]he analysis can be more liberal with inclusion of covariates in the model than in most traditional settings. For instance, covariates with $p > 0.05$ can be included in the propensity score model. (D'Agostino and D'Agostino, 2007)

Despite the prevalence of inclusion algorithms emphasizing a variable's utility in predicting assignment, both the theoretical and Monte Carlo literatures show that the key criterion for inclusion in a predictor set is a variable's effect on the outcome of interest. (Rubin,

1997; Rubin and Thomas, 1996; Heckman and Navarro-Lozano, 2004; Robins et al., 1992; Brookhart et al., 2006)

In the absence of strong ignorability, the choice of predictor variables is even more fraught. Wooldridge (2005) shows that including a predictor variable which is affected by the treatment decision leads to inconsistency in propensity score estimation. In a paper closely related to ours, Heckman and Navarro-Lozano (2004) examine a case in which the analyst does not have access to the full set of necessary predictor variables to ensure strong ignorability–the case in which there are omitted variables. In that case, adding variables to the predictor set of a propensity score-based estimator may decrease or increase the inconsistency caused by the omitted variable. In a factor model of selection using normally distributed factors, they find that adding a variable to the propensity score predictor set that is strongly correlated with assignment to treatment but weakly correlated to outcome generally increases inconsistency.

One path out of this difficulty is through the use of instrumental variables. (Angrist et al., 1996; Imbens and Angrist, 1994) If an analyst does not have access to a set of predictor variables satisfying strong ignorability but does have access to an instrumental variable (a variable which does not directly affect the outcome but does affect the assignment to treatment) then a consistent instrumental variables estimator may be constructed.

In this paper, we focus on the case in which an analyst does not have a set of predictor variables satisfying strong ignorability but does have access to an instrumental variable. We investigate the effect of introducing an instrumental variable to the predictor set of a propensity score. We find that, at least in the case of linear adjustment, using an instrumental variable as a predictor variable in a propensity score method yields greater inconsistency than would be obtained by calculating the naive estimator (the simple difference in means between treatment and control). Furthermore, we find that the inconsistency increases in the strength of the instrument used, that is, in how well the instrument predicts assignment.

The intuition for our results is straightforward, at least in a linear model. The variation in assignment may be decomposed into "good" variation — variation that is uncorrelated with outcomes — and "bad" variation — variation that is correlated with outcomes. The naive estimator uses both sources of variation to identify the treatment effect, and is therefore inconsistent. An instrument is a variable which identifies some of the good variation, and the instrumental variables estimator uses this subset of the good variation to identify the treatment effect (leading both to its consistency and its larger standard errors). A propensity score estimator using the instrument as a predictor *controls for* and thereby removes some of the good variation, so that the treatment effect is identified by the remaining variation which now has a greater proportion of bad to good variation. Since a stronger instrument removes more good variation, the stronger the instrument, the worse it is to control for it. Thus in the case of an instrumental variable, creating balance between the treatment and control groups can be undesirable.

Since the propensity score matching estimator with the instrument in the predictor set is both more biased and more variable than is the naive estimator, it is conceivable that the confidence intervals for the matching estimator would have greater coverage rates. In a Monte Carlo simulation, we show that this need not be the case. We exhibit a simple example in which naive estimator coverage rates are consistently below matching estimator coverage rates.

Our results suggest that the typical guidance by statisticians provided to researchers conducing propensity score analyses regarding the selection of variables to include in a propensity score analysis should be modified. D'Agostino and D'Agostino summarize this advice as follows:

> In addition, in the same way that randomization in a clinical trial will create balance on all patient characteristics, both those related to outcomes to be assessed later and those unrelated to outcomes, the focus should be on including variables

> in propensity score models that are unbalanced between the treated and control groups, and not necessarily be concerned specifically whether they are related to the outcomes of interest. (D'Agostino and D'Agostino, 2007)

A variable which is unbalanced between treatment and control is, *ipso facto*, predictive of assignment, so this advice is to include variables which are predictive of assignment without regard to whether they are predictive of outcome.

Hirano and Imbens (2001) suggest the following algorithm for selecting the set of propensity score predictors:

> After estimating this logistic regression [which includes the first covariate] by maximum likelihood we compute the t-statistic for the test of the null hypothesis that the slope coefficient [on the covariate] is equal to zero. If the t-statistic is larger in absolute value than [a pre-specified cut-off value] this variable will be included in the vector of covariates used in the final specification of the propensity score. After estimating all [univariate] logistic regressions we end up with the subset of covariates whose marginal correlation with the treatment indicator is relatively high. We orthogonalize the set of selected covariates, and use these to estimate the propensity score." Hirano and Imbens (2001)

The advice in these "how to" articles recommending generous and treatment-prediction-centric inclusion criteria appear to have been widely followed in practice. Weitzen et al. (2004) reports a methodological review of medical articles published in 2001 which used propensity score modeling. Of the 47 studies reviewed, 23 reported the inclusion criteria for predictors in their propensity score models. Of these 23, 12 used the predictive value of the variable for treatment (univariate p-values, stepwise inclusion algorithms, or explicit goodness-of-fit tests). Of the remaining 11 studies, seven used "non-parsimonious" approaches to predictor set development, in one case involving over one hundred predictor

Table 1: Criteria for inclusion of propensity score predictors

	Effect on		
Literature	Outcome	Treatment	Goodness of Fit
Economics	55%	45%	15%
Medical	35%	80%	30%

variables. The remaining four used *a priori* criteria for inclusion, although the review does not report what these criteria were.

We performed a brief literature review of the recent use of propensity score techniques in the economic and medical literatures.[1] On May 13, 2009 we searched for the keyword "propensity score" in Econlit for economics articles and Medline for Medical articles. We reviewed the twenty most recent empirical articles for which we had full text access and examined them for their description of how their propensity scores were calculated.[2] We report our results in Table 1. In the medical literature, eighty percent (16/20) of the articles explicitly mention that they use prediction of treatment as a criterion for selection of propensity score predictors while only 35% mention prediction of outcome as a criterion. Furthermore, 30% mention explicitly the use of goodness-of-fit criteria in the construction of propensity score predictor sets. The economics literature, by contrast, is less likely to use goodness-of-fit, less likely to mention prediction of treatment and more likely to mention prediction of outcome in discussing propensity score predictor set choice.

The traditional advice on covariate selection in propensity score analyses, which focuses on the statistical association between the treatment indicator and covariates, will lead merely to inefficiency under the assumption of strong ignorability, but it may worsen bias in the

[1] A list of the papers reviewed and the categories to which they were assigned is available from the authors on request

[2] In some cases, we inferred these reasons. For example, some authors explain why each variable belongs in the outcome equation and then include each of these variables in their propensity score prediction equation. We classified those cases as choosing predictor variables on the basis of their effect on outcomes.

estimation of the treatment effect when strong ignorability does not hold. And this case is likely typical in economic applications and in medical studies where patients are not randomized into treatment. The upshot of our analysis is that there is no substitute for carefully considering whether each potential treatment predictor is or is not an instrument; in empirical economics applications this is typically most convincingly done in the context of a well-developed economic model.

We also present two illustrative case studies: the Tennessee STAR experiment to illustrate the case of strong instruments, and observational data on the use of Swan-Ganz catheterization to illustrate the case of a weak instrument.[3] As the theory predicts, the naive estimator is less inconsistent than is the propensity score estimator in each case, and the inconsistency is larger in the strong instrument case, in each case assuming that our assumption of instrument validity is true.

2 Model

In the Rubin (1974) causal model, let D be a binary variable taking the value 1 if the subject has received a treatment of interest. If an omnipotent experimenter were to assign the subject to receive treatment, the subject's outcome would be Y_1, and it would be Y_0 if not assigned to receive treatment. Thus, the subject's outcome is:

$$Y = DY_1 + (1 - D)Y_0 = Y_0 + D(Y_1 - Y_0) = Y_0 + D\Delta$$

The final equality serves to define Δ. The object of the inquiry is then to estimate the distribution of Δ, the treatment effect. As discussed in Heckman and Robb (1985), we can

[3]Our inclusion of the Tennessee STAR case study in this paper is not meant to suggest that analysts commonly include a random assignment indicator from a randomized trial among the set of predictors in a propensity score regression. Rather, we include it to illustrate a situation where the instrumental variable is incontrovertibly strong.

write:

$$E\{Y|D\} = E\{Y_0|D\} + E\{\Delta\}D + E\{\Delta - E\{\Delta\}|D=1\}D$$

A naive estimator of the population average treatment effect, $E\{\Delta\}$, is the regression coefficient from an ordinary least squares (OLS) regression of Y on D. Under standard regularity conditions, it converges to:

$$E\{\Delta\} + E\{Y_0|D=1\} - E\{Y_0|D=0\} + E\{\Delta - E\{\Delta\}|D=1\}$$

The first term, $E\{\Delta\}$, is the population average treatment effect. The first and fourth terms together, $E\{\Delta\} + E\{\Delta - E\{\Delta\}|D=1\}$ are the effect of treatment on the treated. The second and third terms, $E\{Y_0|D=1\} - E\{Y_0|D=0\}$ are the selection bias terms.

Heckman (1997), Heckman and Robb (1985), and Ichimura and Taber (2001) examine the case in which the treatment effect is known to be uncorrelated with the treatment variable, D, and there exists an instrument, Z, which is mean-independent of (Y_0, Y_1). Heckman (1997) shows that Z is a valid instrument if:[4]

$$E\{Y_i|Z\} = E\{Y_i\} \qquad i = 0, 1 \tag{1}$$

$$\text{Cov}(D, \Delta|Z) = 0 \tag{2}$$

$$V(E\{D|Z\}) \neq 0 \tag{3}$$

Except where otherwise noted, we assume that these assumptions are satisfied. Since Y_0 is potentially correlated with D, there is selection (on unobservables) bias that renders the naive estimator inconsistent.

Under these assumptions, the treatment effect varies neither with the instruments (as-

[4]Assumptions 1 through 3 imply Heckman's assumptions, which he shows are sufficient for instrument validity.

sumption 1) nor with treatment (assumption 2). The assumption that the decision to seek treatment is uncorrelated with effect size, conditional on observables, is standard in the propensity score literature but not in the instrumental variables literature (c.f. Imbens and Angrist, 1994). The assumption that the instruments are uncorrelated with outcomes is standard in the instrumental variables literature, as this is part of the definition of an instrument.

Since we are interested in comparing the performance of propensity score and instrumental variables estimators, we make the union of standard assumptions in the two literatures. Under these assumptions the naive, propensity score, and instrumental variables estimators are aiming to estimate the same thing, the global average treatment effect. It is only in a setting where the treatment effect does not vary with treatment or with the instruments that these three estimators are even aiming at estimating the same treatment effect.

We relax the assumption that the treatment effect is uncorrelated with treatment in some of our discussion below. Heckman (1997) also shows that, if we replace assumption 2 with the assumption that $\text{Cov}\left(D, E\{\Delta|D=1, Z\}\right) = 0$, then the IV estimator becomes consistent for $E\{\Delta|D=1\}$, the average effect of treatment on the treated.

Let $e(Z) = P(D=1|Z)$ be the propensity score calculated using the instrument, Z. Since $E\left\{Y|Z\right\} = E\left\{Y_0\right\} + E\left\{\Delta\right\}e(Z)$, the average treatment effect, $E\left\{\Delta\right\}$ may be consistently estimated by an OLS regression of Y on an intercept and $e(Z)$. Similarly, it may be estimated via instrumental variables estimation of Y on D, using $e(Z)$ as an instrument. In either case, the estimator of $E\left\{\Delta\right\}$ is the sample analogue of:

$$\frac{\text{Cov}\left(Y, e(Z)\right)}{V(e(Z))}$$

2.1 Instruments as propensity score predictors

We consider what would happen were a researcher to use conventional propensity score methods when D is correlated with (Y_1, Y_0) and when he possesses an instrumental variable Z, but does not know it.

First, observe that, if Y and D are correlated, then:

$$\begin{aligned}\frac{\text{Cov}(Y,D)}{V(D)} &= E\{Y|D=1\} - E\{Y|D=0\} \\ &= E\{\Delta\} + \frac{\text{Cov}(Y_0,D)}{V(D)} + \frac{\text{Cov}(\Delta - E\{\Delta\}, D)}{E\{D\}} \\ &= E\{\Delta|D=1\} + \frac{\text{Cov}(Y_0,D)}{V(D)}\end{aligned}$$

Under the assumption that Δ is uncorrelated with D, the naive estimator, regarded as an estimator of the average treatment effect, $E\{\Delta\}$, will have an inconsistency of $\frac{\text{Cov}(Y_0,D)}{V(D)}$. If we drop the assumption that treatment effect is uncorrelated with treatment, then the naive estimator becomes an estimator for the effect of treatment on the treated, with, again, a bias of $\frac{\text{Cov}(Y_0,D)}{V(D)}$. Regarded as an estimator of the average treatment effect, the naive estimator's bias is $\frac{\text{Cov}(Y_0,D)}{V(D)} + E\{\Delta|D=1\} - E\{\Delta\}$.

Consider a researcher who observes Y, D, and a scalar instrument Z. As we mention above, under conditions (1)-(3), $E\{\Delta\}$ may be consistently estimated by an instrumental variables (IV) regression.

Now, imagine that the researcher does not know that Z is an instrumental variable. He would, nevertheless, be able to establish that Z is predictive of D (condition (3)). Furthermore, he would be able to establish that Z is predictive of Y since assumptions (1) and (2) imply that $E\{Y|Z\} = E\{Y_0\} + E\{\Delta\}e(Z)$.

These facts would likely lead him to the conclusion that a propensity score-based method would be a good way to estimate $E\{\Delta\}$. After all, Z is both predictive of Y and unbalanced

in the treatment and control groups.

2.2 Regression propensity score adjustment

One common method of propensity score adjustment is the regression-based approach which, in this case, involves regressing Y on D and $e(Z)$, with the coefficient on D being interpreted as an estimate of $E\{\Delta\}$. This method produces consistent estimates of average treatment effects if, in addition to the assumption of strong ignorability, the conditional expectation of Y given D and $e(Z)$ is linear in $e(Z)$ and D. [5]

Since D is correlated with (Y_0, Y_1), the regression adjustment method will lead to inconsistency. By a standard result in the algebra of least squares, the estimator of the coefficient on D in the regression of Y on D and $e(Z)$ is the sample analogue of:

$$\frac{V(e(Z))\mathrm{Cov}\left(Y,D\right) - \mathrm{Cov}\left(D,e(Z)\right)\mathrm{Cov}\left(Y,e(Z)\right)}{V(D)V(e(Z)) - \left(\mathrm{Cov}\left(D,e(Z)\right)\right)^2} =$$

$$E\{\Delta\} + \frac{1}{1-R^2_{D|e(Z)}}\frac{\mathrm{Cov}\left(Y_0,D\right)}{V(D)}$$

where $R^2_{D|e(Z)}$ is the squared correlation between D and $e(Z)$. The inconsistency is composed of two multiplicative terms, $\frac{1}{1-R^2_{D|e(Z)}}$ and $\frac{\mathrm{Cov}(Y_0,D)}{V(D)}$. The second is the inconsistency of the naive estimator and the first is a factor greater than or equal to one.

If we drop assumption (2) and allow treatment effect to be correlated with treatment,

[5] There is no reason for this assumption to hold in our setting, since we are not assuming strong ignorability.

this expression becomes:

$$\frac{V(e(Z))\text{Cov}\left(Y,D\right)-\text{Cov}\left(D,e(Z)\right)\text{Cov}\left(Y,e(Z)\right)}{V(D)V(e(Z))-\left(\text{Cov}\left(D,e(Z)\right)\right)^2}=$$

$$E\left\{\Delta|D=1\right\}+\frac{1}{1-R^2_{D|e(Z)}}\frac{\text{Cov}\left(Y_0,D\right)}{V(D)}-\frac{\text{Cov}\left(e(Z),E\{\Delta|D=1,Z\}\right)}{V(D)-V(e(Z))}$$

The first two terms are similar to those in the case with no correlation between treatment effect and treatment. The second term gives the propensity-score-adjusted estimator for the effect of treatment on the treated a larger bias than the naive estimator has. It is not possible to sign the third term in general. Continuing to allow for differential treatment effects on the treated and untreated, if we make the alternative assumption discussed above that $\text{Cov}\left(D,E\{\Delta|D=1,Z\}\right)=0$, then the third term vanishes. Then we can again conclude that the propensity score estimator is more biased than the naive estimator for the effect of treatment on the treated.

To formalize the variance decomposition intuition of the introduction, consider the two different variance decompositions below, denoting by $\hat{D}$ the best linear predictor of D given Y_0:

$$\begin{aligned}Y=Y_0+\Delta D&=Y_0+\Delta\hat{D}+\Delta(D-\hat{D})\\&=Y_0+\Delta(D-e(Z))+\Delta e(Z)\end{aligned}$$

In the first line, the variance in D has been apportioned into the "good" (i.e. uncorrelated with Y_0) variance $(D-\hat{D})$ and the "bad" variance $\hat{D}$. In the third line, this variance has been alternatively apportioned into the good variance explained by $e(Z)$ and the other variance contained in $(D-e(Z))$. A regression of Y on either $e(Z)$ or on $(D-\hat{D})$ would produce consistent estimates of $E\left\{\Delta\right\}$, with the latter providing a narrower standard error owing to

the greater variance of $(D - \hat{D})$.

In case the propensity score, $e(Z)$, is uninformative about D, the naive and propensity score estimators are equally inconsistent. As the strength of the instrument rises, $R^2_{D|e(Z)}$ rises, and the propensity score method becomes progressively more relatively inconsistent than the naive method.

2.3 Propensity score stratification and inverse probability weighting

Another popular use of propensity scores is to adjust via stratification on the propensity score. In that approach, the researcher calculates the difference in Y separately for each value of the propensity score and then averages the estimates:

$$\begin{aligned} E\left\{E\left\{Y|D=1,e(Z)\right\} - E\left\{Y|D=0,e(Z)\right\}\right\} &= E\left\{\frac{\text{Cov}\left(Y_0, D|e(Z)\right)}{V(D|e(Z))}\right\} \\ &= E\left\{E\left\{Y_0|D=1,e(Z)\right\} - E\left\{Y_0|D=0,e(Z)\right\}\right\} \end{aligned}$$

In another popular method, the researcher uses the propensity score as an inverse weighting of the outcomes and estimates the treatment effect as:

$$\begin{aligned} E\left\{\frac{YD}{e(Z)} - \frac{Y(1-D)}{1-e(Z)}\right\} &= E\left\{\frac{Y(D-e(Z)}{e(Z)(1-e(z))}\right\} \\ &= E\left\{\frac{\text{Cov}\left(Y_0, D|e(Z)\right)}{V(D|e(Z))}\right\} \\ &= E\left\{E\left\{Y_0|D=1,e(Z)\right\} - E\left\{Y_0|D=0,e(Z)\right\}\right\} \end{aligned}$$

In this setting, these two methods yield an inconsistency of:

$$E\left\{\frac{\text{Cov}\left(Y_0, D|e(Z)\right)}{V(D|e(Z))}\right\} = E\left\{E\left\{Y_0|D=1, e(Z)\right\} - E\left\{Y_0|D=0, e(Z)\right\}\right)\}$$

By contrast, the naive estimator of $E\left\{\Delta\right\}$ will have an inconsistency of:

$$\frac{\text{Cov}\left(Y_0, D\right)}{V(D)} = E\left\{Y_0|D=1\right\} - E\left\{Y_0|D=0\right\}$$

It is not possible to sign the differences in these inconsistencies in general.

Instead, we will seek to sign the bias in the model of Imbens and Angrist (1994) and Angrist et al. (1996). In that model, observations may be divided according to their response to the instrument. Never-takers (N) are observations for which D would equal zero whether Z equals one or zero. Always-takers (A) are observations for which D would equal one whether Z equals one or zero. Compliers (C) are observations for which D is one if and only if Z is one, and defiers are observations for which D equals one if and only if Z is zero. In their model, it is assumed that there are no defiers (that the effect of the instrument on the assignment to treatment is monotone), and we follow.

As those authors discuss, the average treatment effect[6] may be calculated via the instrumental variables estimator described above, which they call the local average treatment effect. As above, we consider an investigator who enters an instrument Z into a propensity score but does not realize that Z is an instrument. Let us denote the expected value of Y_1 among always-takers as $E\left\{Y_1|A\right\}$ and the proportion of the population which are

[6]They discuss local average treatment effects, since they do not assume that average treatment effects are the same for always-takers, never-takers, compliers, and defiers. Since only the average treatment effect for compliers is identified in their model, the "local" in the name local average treatment effects refers to the complier group. In this section, we continue our practice of talking as if there is only one treatment effect. This may be thought of as a shorthand for LATE (for this section only). Alternatively, the reader may think of this as us imposing the assumption 2 on the model of Imbens and Angrist (1994). On this latter interpretation, the assumption of monotonicity is not required any longer, and we thank a referee for pointing this out.

always-takers as P_A, and similarly for never-takers, (N), and compliers (C). Furthermore, let $p = P(Z = 1)$. Then, the naive estimator of the average treatment effect is the sample analogue of:

$$
\begin{aligned}
&E\{Y_1|D=1\} - E\{Y_0|D=0\} = \\
&\qquad \left(\frac{P_A}{P_A + pP_C} E\{Y_1|A\} + \frac{pP_C}{P_A + pP_C} E\{Y_1|C\}\right) - \\
&\qquad \left(\frac{P_N}{P_N + (1-p)P_C} E\{Y_0|N\} + \frac{(1-p)P_C}{P_N + (1-p)P_C} E\{Y_0|C\}\right) = \\
&\qquad \frac{P_A}{P_A + pP_C} E\{Y_1|A\} - \frac{P_N}{P_N + (1-p)P_C} E\{Y_0|N\} + \\
&\qquad \frac{pP_C}{P_A + pP_C} E\{Y_1|C\} - \frac{(1-p)P_C}{P_N + (1-p)P_C} E\{Y_0|C\}
\end{aligned}
$$

By contrast, the propensity score estimator of the treatment effect is the sample analogue of:

$$
\begin{aligned}
&p\left(E\{Y_1|D=1, Z=1\} - E\{Y_0|D=0, Z=1\}\right) \\
&(1-p)\left(E\{Y_1|D=1, Z=0\} - E\{Y_0|D=0, Z=0\}\right) = \\
&p\left(\frac{P_A}{P_A + P_C} E\{Y_1|A\} + \frac{P_C}{P_A + P_C} E\{Y_1|C\} - E\{Y_0|N\}\right) \\
&+ (1-p)\left(E\{Y_1|A\} - \frac{P_N}{P_N + P_C} E\{Y_0|N\} - \frac{P_C}{P_N + P_C} E\{Y_0|C\}\right) = \\
&\frac{P_A + (1-p)P_C}{P_A + P_C} E\{Y_1|A\} - \frac{P_N + pP_C}{P_N + P_C} E\{Y_0|N\} + \\
&\frac{pP_C}{P_A + P_C} E\{Y_1|C\} - \frac{(1-p)P_C}{P_N + P_C} E\{Y_0|C\}
\end{aligned}
$$

An easy-to-work-with special case is $p = 0.5, P_A = P_N$. In this special case, we may re-write,

for the naive estimator:

$$\frac{P_A}{P_A + P_C/2} \left(E\left\{Y_1|A\right\} - E\left\{Y_0|N\right\} \right) + \frac{P_C/2}{P_A + P_C/2} \left(E\left\{Y_1|C\right\} - E\left\{Y_0|C\right\} \right)$$

and, for the propensity-score estimator:

$$\frac{P_A + P_C/2}{P_A + P_C} \left(E\left\{Y_1|A\right\} - E\left\{Y_0|N\right\} \right) + \frac{P_C/2}{P_A + P_C} \left(E\left\{Y_1|C\right\} - E\left\{Y_0|C\right\} \right).$$

In this special case, each estimator is a weighted average of a difference which reveals the treatment effect, $E\left\{Y_1|C\right\} - E\left\{Y_0|C\right\}$, and one which does not, $E\left\{Y_1|A\right\} - E\left\{Y_0|N\right\}$. As P_C approaches 0, both estimators approach the unrevealing difference, $E\left\{Y_1|A\right\} - E\left\{Y_0|N\right\}$. As P_C approaches one, the naive estimator approaches the revealing difference, $E\left\{Y_1|C\right\} - E\left\{Y_0|C\right\}$, while the propensity score estimator approaches the simple average of the revealing and unrevealing differences. Finally, for $P_C > 0$, the naive estimator always weights the revealing difference more highly than does the propensity score estimator.

In the general case, the expressions are not so convenient, but the main results follow. As P_C approaches 0, both estimators approach the unrevealing difference, $E\left\{Y_1|A\right\} - E\left\{Y_0|N\right\}$. As P_C approaches one, the naive estimator approaches the revealing difference, $E\left\{Y_1|C\right\} - E\left\{Y_0|C\right\}$, while the propensity score estimator approaches the p-weighted average of the revealing and unrevealing differences plus an additional term: $p(E\left\{Y_1|C\right\} - E\left\{Y_0|C\right\}) + (1-p)(E\left\{Y_1|A\right\} - E\left\{Y_0|N\right\}) + (1-2p)(E\left\{Y_0|N\right\} - E\left\{Y_0|C\right\})$. Finally, as P_C increases, in both estimators the weights on the compliers' expectation terms increase, but these weights increase faster in the naive estimator.

The results in the case of a discrete, monotonic Z are similar to the results in the linear case. Both estimators are inconsistent. With weak instruments (P_C near zero) the naive and matching estimators have the same inconsistency. With strong instruments (P_C near 1), the matching estimator's inconsistency is larger.

3 Monte Carlo

Since the propensity score matching estimator with the instrument in the predictor set is both more biased and more variable than is the naive estimator, it is conceivable that the confidence intervals for the matching estimator would have greater coverage rates. In this Monte Carlo simulation, we show that this need not be the case.

Let z be the instrument, ϵ_1 and ϵ_2 be error terms in the outcome and treatment equations, d be an indicator for treatment, and y be the outcome variable. We assume the following data generating process for the Monte Carlo experiment:

$$\begin{aligned} z &\sim \text{Exponential}(1) \\ (\epsilon_1, \epsilon_2) &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) \\ d &= 1(\beta z + \epsilon_1 > 0) \\ y &= \gamma d + \epsilon_2 \end{aligned}$$

The parameters of this data generating process are β, controlling the strength of the instrument, γ, the treatment effect, and ρ, controlling the strength of the unobservable confounder(s).

For all the results we present, $\gamma = 0.5$. We consider two values of β, 0.2 and 0.8, corresponding to a weak instrument and a strong instrument. In the former case, $P_C = P\{d|z = 1\} - P\{d|z = 0\} \approx 0.079$, while in the latter case, $P_C \approx 0.29$. We consider $\rho = 0, 0.1, 0.2...0.5$. When ρ is zero, strong ignorability holds. As ρ increases, the strength of the unobservable confounder increases.

For each combination of parameters, we draw 2,000 random datasets with 100 observa-

tions. For each dataset, we estimate the propensity score, $e(z)$ with a probit regression of d on z, and we estimate the treatment effect with a linear regression of y on d and $e(z)$. We also regress y on d alone for a naive estimate of the treatment effect. We use the bootstrap (with 100 replications and BCa confidence intervals) to calculate 95% confidence intervals for each combination of parameters and each Monte Carlo dataset draw.

Figure 1 shows the results of the experiment when there is a strong instrument. The top left panel shows that confidence intervals for the naive estimator are narrower than are confidence intervals for the matching estimator when the instrument is strong. The bottom left panel shows that the bias in the estimate of the treatment effect grows with the strength of unobserved confounders. It also shows that the propensity score matching estimator has a larger bias than does the naive estimator for every value of $\rho > 0$. The top right panel shows the coverage rate of the 95% confidence interval for each of the naive and matching estimators. As ρ increases, the coverage rate declines for each of the matching estimator and the naive estimator. Strikingly and despite the fact that the matching estimator has a wider confidence interval, the coverage rates for the naive estimator are above the corresponding rates for the matching estimator for every value of $\rho > 0$.

Figure 2 shows the analogous set of Monte Carlo results when there is a weak instrument. As in the strong instrument case, for every value of $\rho > 0$, the propensity score matching estimator has a wider 95% confidence interval, a larger bias, and a lower coverage rate than does the naive estimator. However, unlike in the strong instrument case, these differences are small.

4 Case studies

Including an instrument among the predictors of treatment in a propensity score analysis will increase inconsistency over the naive predictor, but how important is this effect in practice?

Figure 1: Monte Carlo Results–Strong Instrument

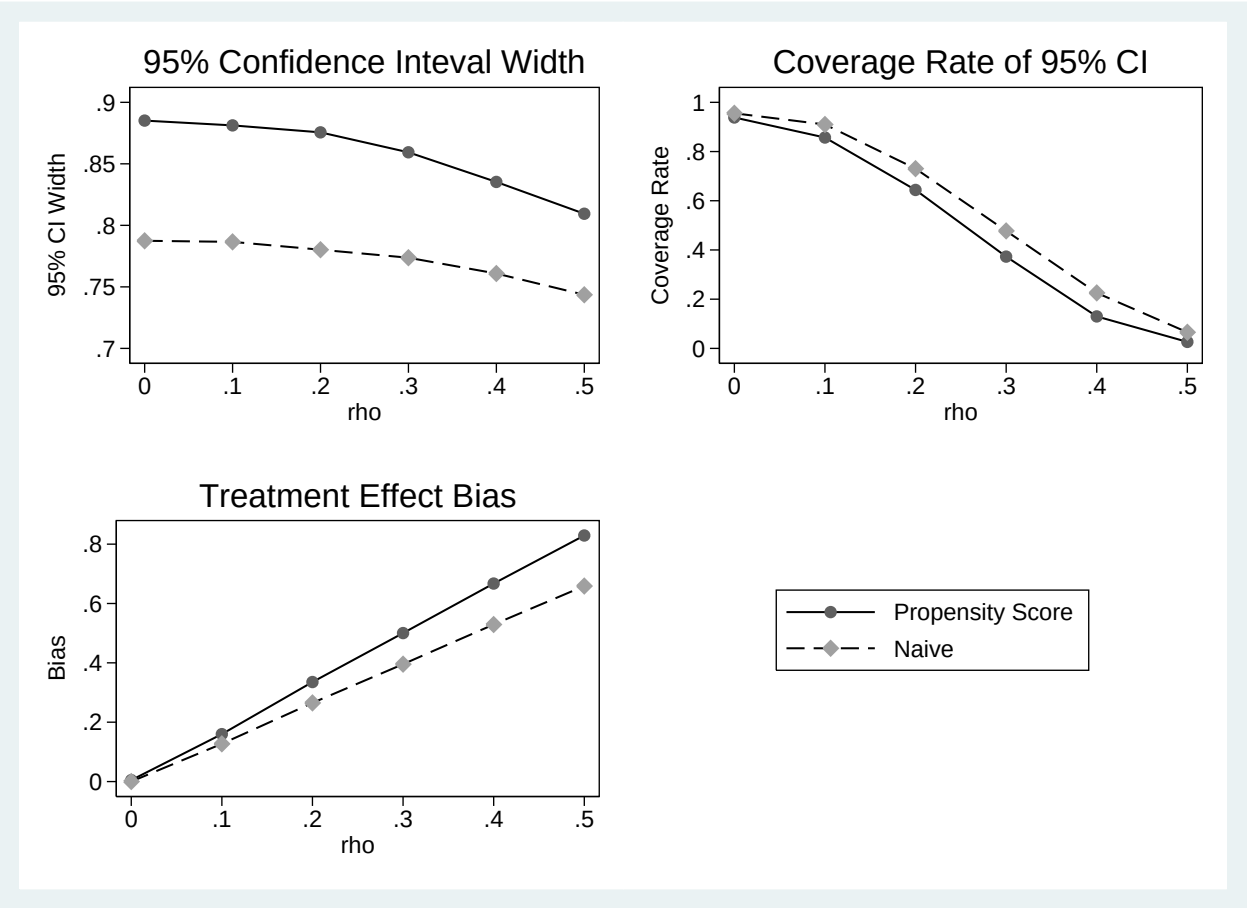
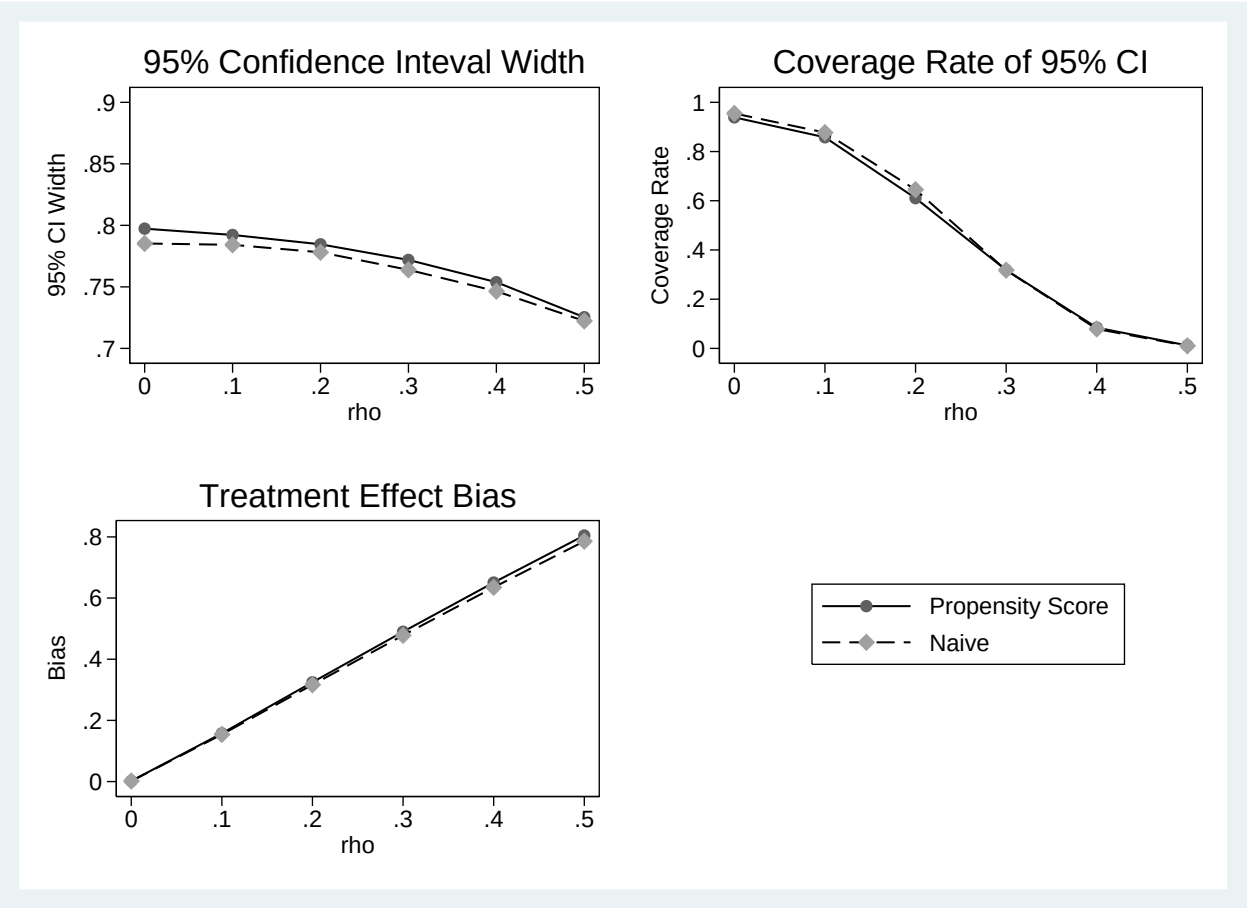


Figure 2: Monte Carlo Results–Weak Instrument



We consider two empirical examples, the first with a strong instrument and the second with a weak instrument.

4.1 Tennessee STAR experiment

The Tennessee STAR experiment was a randomized trial undertaken by the public school system in Tennessee to determine the effect of reducing class sizes for young children in kindergarten through third grade. Starting in 1985, a cohort of entering kindergarten students were randomly assigned to one of three branches: (1) a small class branch (13-17 students per classroom); (2) a regular sized class (22-25 students) with a teacher's aide; and (3) a regular sized class without a teacher's aide. The design of the study required that students assigned to a small class remain in a small class through third grade. For various practical reasons, 11% of the 929 students assigned to a small classroom in kindergarten were in a regular sized classroom by third grade. (That is, $P\{D|Z = 1\} = 0.89$). Conversely, 16.1% of the 2,052 students assigned to regular sized classrooms in kindergarten (with or without teachers aides) were in a small class by third grade. (That is, $P\{D|Z = 0\} = 0.16$). Detailed information about the experiment can be found in Finn et al. (2007).

Here, we ask whether attending a small class in third grade improved performance on standardized tests given to third graders at the end of the year. We consider only students who entered the experiment in kindergarten and stayed in the experiment through third grade. We drop students who do not have standardized test scores available in third grade from the analysis. Despite randomization in kindergarten, differential attrition from the treatment and control groups (often for unobserved reasons) means that small and regular class attendees in third grade are not balanced on their covariates. However, kindergarten randomization does provide us with an strong instrument for assignment to a small classroom, since $P_C = P\{D|Z = 1\} - P\{D|Z = 0\} = 0.73$ is large.

If an analyst came upon these data but did not understand that kindergarten assignment

was a good instrument, he would likely use it in the construction of a propensity score. It is correlated with assignment ($\rho = 0.69$, $p < 0.0001$) and is correlated with outcomes (for example, $\rho = 0.081$, $p < 0.0001$ for 3rd grade reading score).

In Table 2, we describe the results of a number of analyses aimed at finding the effect of small class size on achievement in 3rd grade. We examine two outcome measures, 3rd grade reading score and 3rd grade math score. The columns of the table describe the method used to estimate the effect. The first column uses instrumental variables, using kindergarten class size assignment as an instrument for 3rd grade class size. The second column reports a naive OLS regression of the outcome on an indicator variable for small class size. The third column is like the second, except that a propensity score constructed from the instrument is also included as a control in the regression. The fourth column is like the third, except that the propensity score there is constructed using both the instruments and other controls. The rows indicate the outcome measure, either reading or math score, and the super-rows indicate whether or not covariate controls are included linearly in the regression. The covariate controls are listed at the bottom of the table.

The table exhibits the phenomena predicted by the theory. Consider the reading score with covariates. The Naive column says that children in small classes score, on average and adjusted for covariates, 6.00 points higher than do children in large classes. When this same analysis is run using instrumental variables and adjusting for the same covariates, the estimate of the class-size effect rises to 8.73. When we move from the naive estimator of class-size effect to estimations which include the propensity score as controls in addition to covariates, the estimated effect falls to 2.97. Adding a propensity score which contains the instrumental variable increases the inconsistency relative to the naive estimator. The results without covariates show a similar pattern: adding an instrument-containing propensity score increases inconsistency relative to the naive estimator (assuming that the randomization is a valid instrument). Throughout, the IV results indicate that the true effect of small class

Table 2: Treatment Effect of Small Classroom in 3rd Grade on Test Scores

Covariates		IV	Naive	OLS w/ $e(Z)$	OLS w/ $e(X,Z)$	Mean [s.d.]
No	Reading	8.59	5.85	5.78	3.08	624
		(2.02)**	(1.40)**	(1.40)**	(1.94)	[37]
	Math	6.80	4.68	4.53	2.52	626
		(2.15)**	(1.50)**	(1.48)**	(2.05)	[40]
Yes	Reading	8.73	6.00	2.97	2.97	624
		(2.01)**	(1.34)**	(1.84)	(1.84)	[37]
	Math	6.96	4.89	2.41	2.41	626
		(2.15)**	(1.43)**	(1.95)	(1.95)	[40]

- $N = 3,019$ in the reading regressions and $N = 3,056$ in the math regressions. Sample sizes differ because not all students took all the exams.
- Standard errors in parentheses.
- * significant at 5%; ** significant at 1%.
- Demographic controls in the regressions reported in the lower half of the table include indicators for gender, race, whether the child lives in an urban, suburban, or rural area, and whether the child qualifies for a free school lunch.

size is larger than the naive estimate reveals, but, an estimate employing an instrument-containing propensity score is *lower* than the naive estimate.

4.2 Swan-Ganz catheterization

The placement of Swan-Ganz catheters is common among ICU patients – over 2 million patients in North America are catheterized each year. A Swan-Ganz catheter is a slender tube with sensors that measures hemodynamic pressures in the right side of the heart and in the pulmonary artery. Once in place, the catheter is often left in place for days, so it can continuously provide information to ICU doctors. This information is often used to make decisions about treatment, such as whether to give the patient medications that affect the functioning of the heart. It is a controversial question in medicine, however, whether a Swan-Ganz catheterization reduces patient mortality or increases it.

An influential observational study by Connors et al. (1996) finds that patients who receive

Swan-Ganz catheterization during their first day in the ICU are 1.27 times more likely to die within 180 days of their admission. Even at 7 days after ICU admission, Connors et al. (1996) find that catheterization increases mortality. This conclusion was very surprising to ICU doctors, many of whom continue to use the Swan-Ganz catheter to guide therapy in the ICU. This result led to the organization of several randomized controlled trials (e.g. Richard et al., 2003; Sandham et al., 2003) to test the effect of Swan-Ganz catheterization on survival in some clinically defined special populations. These randomized controlled trials found no effect of Swan-Ganz catheterization on mortality.

The Connors et al. (1996) data come from ICUs at five prominent hospitals – Duke University Medical Center, Durham, NC; MetroHealth Medical Center, Cleveland, OH; St. Joseph's Hospital, Marshfield, WI; and University of California Medical Center, Los Angeles, CA. The study admitted only severely ill patients admitted to an ICU. Murphy and Cluff (1990) provide a detailed description of patient recruitment procedures, including a list of exclusion criteria. Connors et al. (1996) count a patient as catheterized if the procedure was performed within 24 hours of entering the ICU.

Health Services & Outcomes Research Methodology invited a number of researchers to re-analyze the data from Connors et al. (1996) and published these reanayses in a 2001 special issue. One of these papers (Hirano and Imbens, 2001) used a propensity score method using an inclusion criterion based on goodness of fit to construct the propensity score. They found that Swan-Ganz catheterization had a large, negative effect on survival.

In the Swan-Ganz example, using the instrumental variables approach rather than the propensity score approach makes a substantive difference. The instrumental variables approach finds that catheterization improves mortality outcomes only in the short run, if at all, and increases mortality in the long run. (Redacted et al., 2005) As we note above, the propensity score approach finds that catheterization harms patients in both the short and long runs. The result using the instrumental variable approach is intuitively appealing be-

cause it suggests a possible explanation for the fact that many ICU doctors are committed to the use of the Swan-Ganz catheter. Since most ICU patients leave the ICU well before 30 days after admission have elapsed, ICU doctors may never observe the increase in mortality. The instrumental variable approach is also closer to the results of randomized trial evidence in this area (conducted in special populations with particular medical conditions), which tends to find no mortality effect of catheterization.

Here, we reanalyze the the same Connors et al. (1996) data to illustrate the consequences of including a weak instrument in a propensity score analysis. Our instrument for Swan-Ganz catheterization is the patient was admitted to the ICU on a weekday (rather than a weekend). Redacted et al. (2005) argue that, for these data, this variable meets the two crucial requirements for an instrument's validity. Unlike the STAR experiment case, the correlation between the instrument and treatment is small ($\rho = 0.057$, $p < 0.05$) but is significant. The correlation between the instrument and outcomes is also small but often significant (for example, the correlation with 60-day mortality is $\rho = 0.035$, $p < 0.05$).Thus, it seems possible that an analyst would include this variable in the predictor set of a propensity score analysis.[7] In this case, $P\{D|Z = 1\} = 0.46$ and $P\{D|Z = 0\} = 0.40$, so $P_C = 0.06$, which means that the instrument is weaker than in the previous empirical example.

Table 3 is arranged identically to Table 2. The entries in the table show the estimated effect on mortality at either 60 or 90 days from ICU admission of the use of a Swan-Ganz catheter. The columns and super-rows denote the various estimation techniques.

The results in the case of Swan-Ganz catheterization have some similarities to those in the STAR experiment. For example, the IV estimates indicate that the true mortality effect of Swan-Ganz catheterization is higher than the naive estimator would suggest, but

[7]Like Redacted et al. (2005), we confine our analysis to patients with acute respiratory failure, congestive heart failure, and massive organ system failure (with sepsis or malignancy). For other patients, the correlation between weekend admission and treatment is not statistically significant.

Table 3: Treatment Effect of Swan-Ganz Catheterization on Mortality

			IV	Naive	OLS w/ $e(Z)$	OLS w/ $e(X,Z)$	Mean [s.d.]
Covariates	No	60 days	0.600	0.094	0.094	0.074	0.387
			(0.286)*	(0.014)**	(0.014)**	(0.016)**	[0.487]
		90 days	0.629	0.093	0.093	0.073	0.419
			(0.292)*	(0.015)**	(0.015)**	(0.017)**	[0.493]
	Yes	60 days	0.642	0.076	0.074	0.074	0.387
			(0.313)*	(0.015)**	(0.015)**	(0.015)**	[0.487]
		90 days	0.674	0.075	0.073	0.073	0.419
			(0.320)*	(0.015)**	(0.015)**	(0.015)**	[0.493]

- $N = 4,572$ in all the regressions.
- Standard errors in parentheses.
- * significant at 5%; ** significant at 1%.
- Controls in the regressions reported in the lower half of the table include age, gender, race, insurance coverage, income, indicators for primary and secondary diagnoses, medical history, and a wide variety of laboratory tests. Redacted et al. (2005) (in their Tables 1-3) show summary statistics on these variables.

the propensity-score adjustment results in a reduced estimate of the effect, relative to the naive estimator. There are two interesting differences, however. Again, assuming instrument validity, the propensity-score-adjusted results are more biased than are the results from a naive comparison. Here, because the instrument is weak, the standard errors on the IV estimates are quite large. Also, again because the instrument is weak, the difference between the naive estimator and the propensity-score-adjusted estimator is small.

5 Discussion

We show theoretically that including an instrument in the predictor set for a propensity score leads to greater inconsistency than would arise from a naive estimate and that the extra inconsistency grows with the predictive power of the instrument. The methods used in much of the applied propensity score literature, methods directed at finding predictor variables highly correlated with assignment, seem prone to producing these inconsistencies.

In our empirical applications, we show that, in the case of strong instruments, mistakenly

including instruments in the predictor set of a propensity score can increase inconsistency in a substantively significant way. One might object that our strong-instrument example is unrealistic: we are imagining a researcher who ignores the fact that the outcome of randomization is a potential instrument. In our view, however, this example serves starkly to illuminate our main point: it is central to bring problem-specific knowledge to bear when using propensity score-based methods. When a researcher uses an instrumental variable in the construction of a propensity score, the estimates become more inconsistent than with a naive estimator.

This raises the question of what remedies are available to empirical researchers. There is no statistical test to determine whether a particular variable is an instrument; therefore, there is no pat statistical procedure which will either detect or solve the problem. The only solution to the problem of selection on unobservables is to identify an instrumental variable. To find an instrument, the researcher must rely on detailed knowledge of either the institutional arrangements leading to the choice of treatment or on an economic model of that choice. In a randomized controlled trial, this knowledge comes from understanding the randomization design. In an observational setting, this knowledge typically comes from behavioral assumptions made about the assignment process (in economic parlance, from exclusion restrictions). The mechanical application of propensity score matching methods in the absence of such knowledge may lead to increased bias relative to naive estimation. Such methods are not a substitute for substantive identification arguments.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the american statistical association*, 91(434):444–455.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Sturmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156.

Connors, A., Speroff, T., Dawson, N., Thomas, C., Harrell, F., Wagner, D., Desbiens, N., Goldman, L., Wu, A., Califf, R., Fulkerson, W. J., Vidaillet, H., Broste, S. Bellamy, P., Lynn, J., and Knaus, W. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. support investigators. *Journal of the American Medical Association*, 276(11):889–897.

D'Agostino, Jr., R. B. and D'Agostino, Sr., R. B. (2007). Estimating treatment effects using observational data. *Journal of the American Medical Association*, 297(3):314–316.

Finn, J. D., Boyd-Zaharias, J., Fish, R. M., and Gerber, S. B. (2007). *Project STAR and Beyond: Database User's Guide*. HEROS, Incorporated.

Heckman, J. (1997). Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. *Journal of human resources*, 32(3):441–462.

Heckman, J. and Navarro-Lozano, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of economics and statistics*, 86(1):30–57.

Heckman, J. J. and Robb, R. (1985). Alternative methods for evaluating the impact of interventions: an overview. *Journal of econometrics*, 30(1-2):239–267.

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 & outcomes research methodology*, 2(3-4):259–278.

Ichimura, H. and Taber, C. (2001). Propensity-score matching with instrumental variables. *American economic review*, 91(2):119–124.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Murphy, D. and Cluff, L. (1990). Support: Study to understand prognoses and preferences for outcomes and risks of treatments-study design. *Journal of Clinical Epidemiology*, 43(suppl):1S–123S.

Redacted, Redacted, and Redacted (2005). Redacted. Working paper, Redacted.

Richard, C., Warszawski, J., Anguel, N., Deye, N., Combes, A., Barnoud, D., Boulain, T., Lefort, Y., Fartoukh, M., Baud, F., Boyer, A., Brochard, L., Teboul, J., and Group., F. P. A. C. S. (2003). Early use of the pulmonary artery catheter and outcomes in patients with shock and acute respiratory distress syndrome: A randomized controlled trial. *JAMA*, 290(20):2713–2720.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(2):479–495.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127(8):757–763.

Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 52(1):249–264.

Sandham, J., Hull, R., Brant, R., Knox, L., Pineo, G., Doig, C., Laporta, D., Viner, S., Passerini, L., Devitt, H., Kirby, A., Jacka, M., and Group, C. C. C. C. T. (2003). A randomized, controlled trial of the use of pulmonary-artery catheters in high-risk surgical patients. *New England Journal of Medicine*, 348(1):5–14.

Weitzen, S., Lapane, K. L., Alicia Y. Toledano, A. L. H., and Mor, V. (2004). Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemioligy and drug safety*, 13(12):841–53.

Weitzen, S., Lapane, K. L., Alicia Y. Toledano, A. L. H., and Mor, V. (2005). Weakness of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. *Pharmacoepidemiology and drug safety*, 14(4):227–238.

Wooldridge, J. M. (2005). Violating ignorability of treatment by controlling for too many factors. *Econometric Theory*, 21:1026–1028.